Journal of
MATHEMATICAL
PHYSICS

Volume 14 January 1973 Number 1

A model for peaking of galactic gravitational radiation

Gerald A. Campbell and Richard A. Matzner
Department of Physics, University of Texas, Austin, Texas 78712
(Received 10 April 1972)

Geometrical optics is used to calculate the radiation pattern from a source in orbit in a strong gravitational field. No specific mechanism is postulated for the radiation itself, and only the field's effect on the radiation enters. (The model proposes a "black hole" at the galactic center.) Besides the Doppler peaking expected in these orbits, we find that the gravitational lens effect can enhance the radiation (regardless of how the radiation is produced). If the radiation arises from individual short events, the gravitational lensing leads to a scatter in the observed intensity. Formulas are presented for the probability a certain pulse will exceed the average by a given factor for a detector of finite sensitivity. Enhancement as found here, if present in the galaxy, would lower the overall galactic mass loss implied by Weber's gravitational radiation measurements.

1. INTRODUCTION

Weber[1] has recently reported experimental results indicating a large flux of gravitational radiation, apparently emanating from the galactic center. There immediately arises an energy problem for the production of such radiation, since estimates of the energy flux imply a mass loss by the center of the galaxy of $\sim 10^3 M_\odot/\mathrm{yr}$, as a conservative estimate, assuming an isotropic radiation pattern from the center of the galaxy. Such a high mass loss gives an implausibly short lifetime for the galaxy and seems to exceed by about an order of magnitude the rate of change of the galactic mass as deduced from astronomical observations.[2]

To reconcile the Weber results with the astronomical ones, the builder of galactic models must find some physical effect which reduces the over-all radiation rate from the galactic center. In view of the nearness of the position of the sun to the galactic equatorial plane (with angular alignment $\sim 10^{-3}$ or even 10^{-4}), one is led to consider the possibility of a pattern of radiation which is peaked in the disc. Two such possibilities come immediately to mind. They are (a) the forward Doppler peaking associated with radiation emitted from a moving source and (b) the focusing effect of strong gravitational fields. Both these phenomena require very strong gravitational fields: possibility (a) because strong fields are required to get relativistic velocities from particles moving in gravitational fields, and possibility (b) because, for large deflection angles ($> 2\pi$), a very strong field is needed.

As a working hypothesis, then, we consider a model of the galaxy with the galactic nucleus containing a black hole or collapsar, presumably a spinning black hole because of the ubiquity of angular momentum in galaxies. This spinning collapsar is assumed to have its axis aligned with that of the galaxy with a high degree of accuracy. Because of the tendency of contracting spinning objects to form a disc structure, we postulate that matter outside but near the collapsar also moves in orbits lying in the same plane as that of the observable disc of the galaxy. This matter may consist of stars, along with gas and dust which are sinking toward the center. We take as the mass of the collapsar, $M_c = 10^8 M_\odot$, which is an upper bound on the mass allowed, from observations of the galactic nucleus.[3]

We postulate that gravitational radiation is somehow given off from objects in orbit near the collapsar. This could be gravitational synchroton radiation as Misner[4] has suggested. (This requires very relativistic orbits, just as are required in electromagnetic radiation.[5]) Or, we may suppose that the radiation originates in some way which is local to the orbiting bodies. For instance, stimulated collapse may occur for bodies near the collapsar. Massive stars moving through the dust near the center may go supernova because of the increase of mass by accretion; inhomogeneities in the gas and dust may allow sufficiently rapid accretion onto a neutron star to cause it to collapse to a black hole; or tidal effects may trigger such an event, leading to a substantial pulse of gravitational radiation. Alternately, small ob-

jects falling into neutron stars can themselves emit pulses of radiation.

We do not suggest that these production mechanisms exhaust the possibilities. We also do not suggest that the ones mentioned here are even particularly plausible. We do not intend, in this paper, to go into the production per se, of the radiation. We simple assume it is somehow produced, and attempt to follow its fate subsequent to its production. We feel confident that attempts will be made, by others if not by us, to find acceptable models of the galactic center, involving black holes to explain the gravitational radiation flux. We note here some effects which must be present in any such model.

Although we have assumed a central black hole which would presumably be spinning, calculations based on the Kerr metric[6] which describes such a situation are still under way. In this paper we will present calculations based only on the Schwarzschild metric. In this regard we refer to the calculations by Bardeen and Cunningham[7] for the "extreme Kerr" case, which has a (angular momentum per unit mass) equal to m (the total mass of the system in geometrized units $G = c = 1$). In the general Kerr metric a is a specifiable constant, with the Schwarzschild metric having $a = 0$, while the extreme Kerr has $a = m$. Hence the work we present and the work of Bardeen and Cunningham bracket a range of behaviors. We hope to soon fill in this range of behaviors with calculations for the Kerr metric with arbitrary a. *Note added in proof:* Since this paper was submitted for publication, results similar to those presented here for Schwarzschild were obtained for maximal Kerr by J. K. Lawrence in a preprint.

2. GEOMETRICAL OPTICS

We have hypothesized emitters in high velocity orbits around the central collapsed body. Because of the orbital velocity there will be peaking of gravitons emitted forward in the direction of the motion. We will use geometrical optics to discuss this peaking, which Misner[4] describes in terms of wave optics. Drawing on experience in electrodynamics we expect that the peaking found from geometrical optics will fairly well describe the results found from the more accurate wave theory. It should be noted that the Doppler shifting of radiation emitted forward shortens its wavelength and improves the geometrical optics approximation for such radiation. We will not, of course, obtain accurate results for radiation emitted with a wavelength typical of the scale of curvature of the emitting system. Thus, if we consider neutron stars as emitters we will obtain correct results only for distances greater than several tens of kilometers from the emitter. Thus the geometrical optics description will be accurate for the radiation from a single neutron star which emits its radiation near our postualted $M_c \sim 10^8 M_\odot$ collapsar, since the radiation will have a wavelength typical of the source ($\sim M_\odot$) while the dimension of the background radius of curvature is $\sim 10^8 M_\odot$.

The geometrical optics approach gives the synchrotron-like forward peaking, but is not a completely accurate description of the wave phenomena. The wave analysis is in principle accurate but usually proceeds by decomposition into orthogonal polynomials; this means many terms are needed to demonstrate the sharp peaking of radiation we are concerned with here. In dealing with the gravitational synchrotron radiation mechanisms suggested by Misner (where the fundamental frequency is typical of $M_c \sim 10^8 M_\odot$ rather than of $M_\odot$), we cannot get correct answers for the lowest modes; however, we do obtain essentially accurate predictions for the higher modes which do show the Doppler peaking.[4,8]

One of the characteristics of radiation ignored by the geometrical optics approximation is the backscattering of radiation, since it is based on an expression which does not allow a superposition of waves traveling in opposite directions. In this problem, except for the excluded low modes as noted above, the frequencies we consider are sufficiently high so that backscattering is essentially negligible. The "effective potential" in the relevant wave equations has a small peak compared to the energy eigenvalue (essentially ω^2) associated with the wave. Hence we feel confident in applying these results to a large class of possible phenomena.

Furthermore, some effects which are rather apparent in one type of analysis may be overlooked in another. For instance, the commonplace gravitational lens effect may give large enhancements of intensities, a point which is not obvious in the wave analysis but follows straightforwardly in the geometrical optics treatment.

We will go very sketchily into the details of the geometrical optics formulation because it is well documented in the literature, and because its application, rather than its derivation, is the point of this paper.[9]

It can be shown that $k_\mu = \phi_{,\mu}$ (where ϕ is the phase of the wave quantity) is tangent to null geodesic rays (i.e., gravitons and photons travel along null geodesics) and that the polarization of the vector potential A_μ and of the gravitational potential $h_{\mu\nu}$ are parallelly propagated along these null geodesics.

We will ignore polarization in our discussion but will use the fact that null geodesic rays describe photon orbits, and will make use of the area intensity law[10]:

$$\bar{I}dA = \bar{I}_{(0)}dA_{(0)}, \tag{2.1}$$

where $\bar{I}$ is the intensity (number of photons or gravitons per square centimeter per second times $h\nu$) and dA is an elemental 2-area spanned by the radiation in the rest frame in which the photons or gravitons are counted and where it is assumed that the observations of the quantities on the two sides of the equation are made by observers who see the same spectrum for the radiation (i.e., no relative red shift). This equation generalizes the r^{-2} law for point sources. To correct for the general case where the two frames of measurement give different spectral shift, we insert the correction due to the red shift: $(\nu/\nu_0)^2$ on the right-hand side. The factor appears as the square because the energy per graviton is Doppler shifted, and the rate of detection of the gravitons is also Doppler shifted. This correction can be written in an invariant way by writing

$$(k_\mu U^\mu)^{-2} I dA = (k_\mu U^\mu{}_{(0)})^{-2} I_{(0)} dA_{(0)}, \tag{2.2}$$

where U^μ and $U^\mu{}_{(0)}$ are the 4-velocities of the two observers, $k_\mu U^\mu{}_{(0)}$ are thus the frequencies observed by the two observers; now I and $I_{(0)}$ are the intensities measured by each, and we have removed the restriction to zero redshift.

This heuristic argument assumes that dA and $dA_{(0)}$ measured by two observers in relative motion are equal.

This is in fact a consequence of the result quoted in Eq (2.2) from the exact derivation.

In our application, we take the quantities $I_{(0)}, dA_{(0)}$ to be measured in the rest frame of the emitter, and we will compute I as measured by an observer at infinity who is at rest (i.e., whose 4-velocity is parallel to the time-like Killing vector).

The derivation thus proceeds by finding null rays which connect the emitter, deep in the potential well, with the observer at infinity. We then consider a small fan of angles about this ray, defining $dA_{(0)}$. These rays can be identified by their constants of motion. A complete set of such constants are available in both the static spherically symmetric case,[11] described by the Schwarzschild metric, and in the stationary model (the Kerr metric) which describes the field of a spinning black hole.[12]

3. RAY OPTICS IN THE SCHWARZSCHILD METRIC

Although the Schwarzschild metric

$$ds^2 = -(1 - 2m/r)dt^2 + (1 - 2m/r)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2) \quad (3.1)$$

can be obtained in the limit of vanishing angular momentum from the Kerr solution, we shall present only calculations based on this simpler metric. Computations in the more difficult Kerr case are underway and will be presented elsewhere.

In view of the comments made above we can compute $dA_{(0)}$ in the Schwarzschild frame of Eq (3.1) rather than going to the moving frame of the source.

The equations governing null geodesics in the Schwarzschild metric are[11]

$$(dz/d\theta)^2 = z^3 - z^2 + \beta^2, \quad (3.2)$$

where

$$z = 2m/r, \quad (3.3)$$

β^2 is a constant

$$\beta = (2m/b), \quad (3.4)$$

and b is the impact parameter of the orbit (which can be unambiguously measured at infinity).

In writing this form we have set the angle ϕ = const; this is possible in view of the spherical symmetry and is a luxury not possible in the nonspherical Kerr solution. We also shall take the observer at $z = 0$ ($r = \infty$). Notice that only one constant (β) enters, instead of the two (energy and angular momentum) which might be expected. All photons follow the same (null) paths if they start with the same direction, regardless of their energy (in the geometrical optics approximation). Hence, only the ratio of energy to angular momentum (β) enters. If we wish to keep track of the time, as well as the orbital position, we write[11]

$$\frac{dt}{dz} = \frac{\beta z^{-2}}{1 - z}\,\frac{1}{(z^3 - z^2 + \beta^2)^{1/2}}. \quad (3.5)$$

In order to use the area-intensity law, we compute the cross section of a fan of gravitons, all passing through the source event, with a range in parameters $\Delta\beta$ and $\Delta\phi$. To simplify the discussion, we temporarily put the source at the pole $\theta = 0$.

We first compute the orthogonal distance between two orbits which lie in the same plane ϕ = const. Now

$$\Delta r\sqrt{g_{rr}} = \sqrt{g_{rr}}\,\frac{\partial r}{\partial\beta}\Big|_\theta \Delta\beta \quad (3.6)$$

is the length of the line joining two such rays along θ = const. However, this is in general not an orthogonal connecting vector, but must be multiplied by $\cos\lambda$, where

$$\tan\lambda = K^r/K^\theta \quad (3.7)$$

and

$$K^\theta = \alpha/r, \qquad K^r = \alpha F^{1/2}(1 - 2m/r)^{-1/2} \quad (3.8)$$

are the components of the momentum in an orthonormal frame, with

$$F = \beta^2/2m^2 - r^{-2} + 2mr^{-3} \quad (3.9)$$

and

$$\alpha = F^{-1/2}\,dr/d\lambda, \quad (3.10)$$

where λ is an affine parameter along the ray. We find the cross sectional length ΔD to be

$$\frac{\Delta D}{2m} = \frac{1}{z}\frac{\partial z}{\partial\beta}\Big|_\theta \Delta\beta. \quad (3.11)$$

The width of the beam in the ϕ direction is given by $r\sin\theta\Delta\phi$; hence the cross sectional area is

$$\frac{dA}{(2m)^2} = \frac{1}{z^2}\frac{\partial z}{\partial\beta}\Big|_\theta \Delta\beta\Delta\phi\,\sin\theta. \quad (3.12)$$

This is the area of the beam at a general field point; in order to use the area intensity law, we compare this area with the area spanned by the same beam, on an infinitesimal 2-sphere of radius ϵ centered on the emitter. As we mentioned above, we need not go to the rest frame of the emitter. Instead we can obtain the relevant infinitesimal area (since it is an invariant) by working in an orthonormal frame aligned with the t, θ, ϕ directions in the Schwarzschild frame. In this frame we put the event of emission at the origin, and, since we have already assumed this event is on the axis ($\theta = 0$), ϕ can be taken over as a spherical angle in this frame. However, we must introduce a new polar angle $\bar{\theta}$ in this frame.

Near the source,

$$\tan\bar{\theta} = K^\theta/K^r = [z(1 - z)^{1/2}/2mF^{1/2}]_e \quad (3.13)$$

where the subscript "e" means "at the point of emission." We thus compute $[\partial\bar{\theta}/\partial\beta]_e$ and obtain, for the infinitesimal area on the sphere:

$$\frac{dA}{(2m)^2} = \left(\frac{\epsilon}{2m}\right)^2\left[\frac{(1 - z)z^2}{2\beta^3}\,\frac{1}{2mF^{1/2}}\right]_e \Delta\phi\Delta\beta. \quad (3.14)$$

Hence with Eqs. (2.2) and (3.12) we find

$$I = \left(\frac{k_\mu U^\mu}{k_\mu U^\mu_{(e)}}\right)^2 \frac{L_0}{(2m)^2}\frac{1}{\beta}\left[\frac{z^2(1-z)}{2mF^{1/2}}\right]_e \frac{z^2}{\sin\theta}\left(\frac{\partial z}{\partial\beta}\Big|_\theta\right)^{-1} \tag{3.15}$$

Here L_0 is the luminosity of the source, equal to $4\pi I_{(0)}\epsilon^2$ measured on the infinitesimal 2-sphere.

Formulas similar to this result have been obtained for geometrical optics both in the Schwarzschild case and in the generic case by Strauss.[13]

In order to compute the intensity pattern observed at infinity, one factors out of the z^2 dependence and evaluates the derivative at $z = 0$.

The solution to Eq. (3.15), and hence the calculation of $(\partial z/\partial\beta|_\theta)$, requires an integration of an elliptic integral.[11] In general this must be carried out on a computer. We present here some computer generated plots of the radiation pattern, averaged over one period of a circular orbit, for an emitter in such a circular orbit isotropically in its rest frame.

These graphs plot observed intensity against the angle that the line to the observer makes to the normal to the orbital plane. Stable, very relativistic orbits do not exist in Schwarzschild; the smallest stable test particle orbit has radius $r = 6m$ in the usual Schwarzschild coordinates. We present in Fig. 1 a radiation pattern for the radiation intensity averaged over the emitter's orbit, when $r = 6m$, against the angle σ between the line to the observer at infinity and the normal to the orbit.

Figure 2 plots a similar calculation for a radius $r = 3.33m$, an unstable circular orbit which has orbital velocity (measured in a local Lorentz frame) $v = 0.866$. We present this result even though the orbit is unstable because we expect such an orbit to be an approximation to one of the inward spiraling orbits of the Schwarzschild solution which has the same instantaneous radius.

Figures 1 and 2 do not show the radiation intensity plotted for observers near the plane of the orbit. The peaking for small angles is quite large and is due to the gravitational lens effect. An analytic treatment suffices for these cases. Depending on how close one assumes the observer is to the plane of the orbit, the lensing effect can be quite large; as shown below, and, as can be seen by considering Eq. (3.15) near the value $\theta = \pi$, it is proportional to $1/\theta$, where θ is the alignment angle. There is not a true singularity, of course, for an emitter of finite size and of finite luminosity.

By the use of asymptotic expressions (as found in the standard texts)[14] for the elliptic integrals involved, it is possible to show that, for $\theta_{obs} - \theta_e \gg 1$,

$$I = \frac{(k_\mu U^\mu)^2}{(k_\mu U^\mu_{(e)})^2}\frac{L}{(2m)^2}\left[\frac{z^2(1-z)}{2m\beta^2F^{1/2}}\right]_e z^2 \times \frac{512}{9}\frac{(\frac{2}{3}-z_e)}{\sqrt{3}(1+\sqrt{3})^2}(1+\sqrt{\tfrac{1}{3}+z_e})^{-2}\frac{e^{-(\theta_{obs}-\theta_e)}}{\sin(\theta_{obs}-\theta_e)}, \tag{3.16}$$

where we have introduced general values θ_e and θ_{obs} for the coordinate of the emitter and observer. It is found that this limiting form holds fairly well even for $\theta_{obs} - \theta \approx \pi$. Suppose we hold θ_{obs} fixed (say $\theta_{obs} = \pi$) and consider a source with a constant luminosity surface density with angular extent θ_s centered on the point $\theta = 0$. Then if θ_s is small, all the terms in (3.16) except $\sin(\theta_{obs} - \theta_e)$ can be approximated as constant, and

$$I \propto \int_0^{\theta_s}\frac{d\theta_e\,\sin\theta_e}{\sin(\pi-\theta_e)} \cong \int_0^{\theta_s} d\theta_e = \theta_s; \tag{3.17}$$

if the $\sin(\theta_{obs} - \theta_e)$ enhancement was not present, the integral would yield $\frac{1}{2}\theta_s^2$. Hence we conclude, roughly speaking, that the focusing gives an enhancement $\propto \theta^{-1}$, where θ is the alignment angle, but cut off at $\sim 2\theta_s^{-1}$ where θ_s is the angular diameter which the emitter subtends, measured from the central hole.[15] If we envisage solar mass neutron stars as emitters, and a 10^8 solar mass collapsar, then θ_s may satisfy $\theta_s \sim 10^{-8}$, and the lens enhancement can be quite large.

If the orbit has sizeable Doppler peaking, the gravitational lens effect will mean the predominant part of the radiation comes from emission on the opposite side of the central collapsar. This follows because the radi-

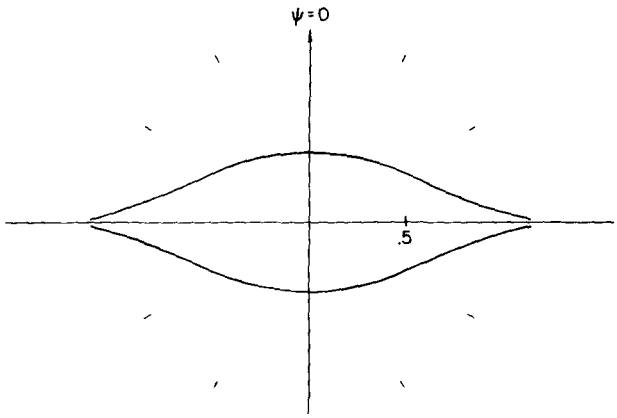

FIG. 1. Polar plot of the intensity observed at infinity, averaged over the orbit of the emitting source, for an emitter in a circular orbit $r = 6m$, the smallest stable circular orbit in the Schwarzschild solution. There is considerable radiation in a direction normal to the orbit. The Doppler peaking is not large; the peaking in the equatorial plane here is due principally to the lens effect.

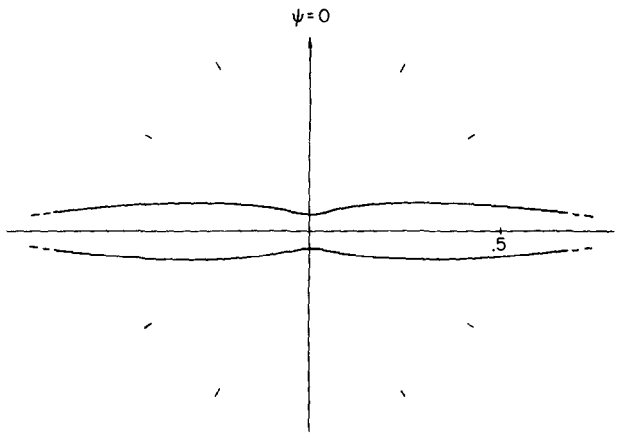

FIG. 2. Polar plot of observed intensity at infinity, averaged over the orbit, for an emitter in an unstable circular Schwarzschild orbit $r = 10/3m$. (Oribts of this radius can be stable in the Kerr solution.) The Doppler effects leads to more suppression of the radiation normal to the orbit, but the enhancement in the equatorial plane is principally due to lensing. The proper intensity of the emitting source is the same in Fig. 2 as in Fig. 1.

ation is emitted preferentially forward from tight relativistic orbits (near $r = 3m$), but it is just such orbits which are eventually deflected to the observer at infinity. The principal reason for this is that no null orbit which reaches infinity can have an apse of less than $3m$, and the bulk of the radiation which escapes to infinity is thus emitted forward (Fig. 3).

In such cases, the lensed terms dominate the expression for the intensity. If the observer does not lie in the plane of the orbit, but some small angle $\theta_{\min} > \theta_s$ away from it, the lens effect will give

$$I \cong K(\sigma^2 + \theta_{\min}^2)^{-1/2}, \tag{3.18}$$

where σ is the angle along the emitter's orbit measured from the point on the opposite side of the orbit from the observer and K is constant. In order to achieve a reduction in Weber's estimated energy flux, it is necessary to average the observed flux over the orbital motion of the emitter. We estimate

$$I_{\text{av}} \approx K\sigma_c^{-1}\ln\left(\frac{\sigma_c + \sqrt{\sigma_c^2 + \theta_{\min}^2}}{\theta_{\min}}\right), \tag{3.19}$$

where σ_c is some limiting value of σ. If the detector sensitivity allows detection of all pulses, then we can detect pulses emitted anywhere in the orbit so $\sigma_c \sim 1$ should be inserted and (3.19) is only approximately valid. If, on the other hand, the detector counts only the strongest pulses, σ_c is determined by

$$I_{\min} = K(\sigma_c^2 + \theta_{\min}^2)^{-1/2}, \tag{3.20}$$

where $I_{\min}$ is the minimum detectable pulse intensity, and Eq (3.19) becomes exact for small σ_c.

The peaking in the plane on this averaged basis is rather soft. Nonetheless, if $\theta_{\min} = 10^{-3}$, the gravitational lens effect averaged over the orbit gives an enhancement of order ~ 10 (if $\sigma_c = 1$). Such a modest reduction in the total emitted flux may be sufficient to (just) bring the energy loss down to that allowed by observations of orbits of stars in the galaxy. This average lens effect should more likely just be thought of as one mechanism reducing the over-all power needed but not providing all the enhancement needed to explain Weber's results.

When considering individual pulses (from individual collapsing neutron stars or individual infalling chunks of debris) whose duration is short compared to the orbital period, very large enhancement may take place.

Suppose $\theta_{\min}$ is fixed. Then, if we consider equal strength pulsed emissions, the probability that the intensity of a single brief pulse exceeds the average of all detectable pulses by a factor N is $P = \Delta\sigma/\sigma_c$, where

$$N\sigma_c^{-1}\ln\left(\frac{\sigma_c + \sqrt{\sigma_c^2 + \theta_{\min}^2}}{\theta_{\min}}\right) = (\Delta\sigma^2 + \theta_{\min}^2)^{-1/2} \tag{3.21}$$

defines $\Delta\sigma$. (We require $\Delta\sigma < \sigma_c$.)

If we write $\sigma_c = n\theta_{\min}$ and assume $n \gtrsim 3$ (say), we have the approximate formula correct for large n

$$P^2 = N^{-2}\ln^{-2}(2n) - n^{-2} \tag{3.22}$$

Table I is a plot of P vs N for the three values $n = 10$, 10^2, 10^3, computed from Eq. (3.21) above (assuming $\sigma_c < 1$).

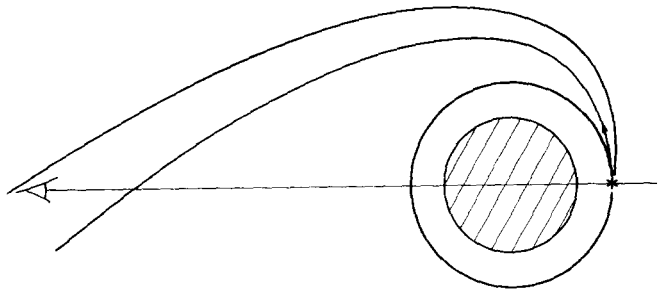

FIG. 3. The rays which are focused when the emitter is on the opposite side of the collapsar from the observer are given off almost tangentially to the orbit. This is also the direction which is preferentially peaked by the Doppler effect. Hence, near the plane of the orbit, the observed intensity is dominated by the lensed contributions.

TABLE I. The probability of observing a pulse larger than N times the average pulse, for three values of $n = \psi_c/\theta_{\min}$.

	N	$P(\%)$
$n = 10$	2.0	16.5
	2.5	11.8
	3.0	8.1
	3.5	4.7
	3.8	1.8
$n = 10^2$	2	10.6
	4	5.2
	6	3.4
	10	1.9
	14	1.1
	18	0.6
	20	0.4
$n = 10^3$	2	7.20
	5	2.88
	10	1.44
	20	0.71
	40	0.35
	60	0.22
	80	0.15
	100	0.10
	120	0.07
	140	0.02

If $n = 100$ (for instance $\theta_{\min} = 10^{-3}, \sigma_c = 10^{-1}$), we find $\sim 2\%$ probability that a pulse will exceed the average by a factor 10. Notice that in this case the maximum possible intensity (assuming a standard source) can exceed the average only by a factor 20. If the orbits are randomly oriented so that occasionally a much smaller $\theta_{\min}$ holds, then, since σ_c will not change, occasional giant pulses will result.

As is well known, $\theta_{\text{obs}} - \theta_e$ can be arbitrarily large since null rays connecting source and emitter can be found which wind around the deflector arbitrarily many times. The possibility of multiple windings means some additional enhancement because of the lensing effect for "higher order ghosts," but the intensity in these ghosts is suppressed by $e^{-2\pi} \sim 0.002$ for each successive ghost, so that they are unimportant past the first or second unless the Doppler peaking compensates.

ACKNOWLEDGMENTS

We would like to thank R. P. Kerr, H. R. Strauss, and Y. Nutku for very helpful discussions, and especially Dr. Strauss for detailed advice on the geometrical optics approximation.

*Supported in part by NSF Grant GP-20033.
†Based on part of a doctoral thesis by G. Campbell to be submitted to the University of Texas.
[1]J. Weber, Phys. Rev. Lett. **25,** 180 (1970).
[2]D. W. Sciama, G. B. Field, and M. J. Rees, Phys. Rev. Lett. **23,** 1514 (1969).
[3]Martin J. Rees and Donald Lynden Bell, Mon. Notic. Roy. Astron. Soc. **152,** 461 (1971).
[4]C. W. Misner, submitted to Phys. Rev. Lett. (1972).
[5]J. D. Jackson, *Classical electrodynamics* (Wiley, New York, 1962).
[6]R. P. Kerr, Phys. Rev. Lett. **11,** 237 (1963).
[7]J. Bardeen, and C. Cunningham, preprint (1972).
[8]C. W. Misner, private communication (1972).
[9]For a discussion of geometrical optics, see G. F. R. Ellis, "Relativistic cosmology" in *General relativity and cosmology* (1969 Varenna Summer School), edited by R. K. Sachs (Academic, New York, 1971).
[10]R. K. Sachs, Proc. R. Soc. A **264,** 309 (1961).
[11]C. Darwin, Proc. R. Soc. A **249,** 180 (1958).
[12]B. Carter, Phys. Rev. **174,** 1559 (1968).
[13]H. R. Strauss, unpublished (1971).
[14]For instance, P. F. Byrd and M. D. Friedmann, *Handbook of elliptic integrals for engineers and physicists* (Springer-Verlag, Berlin, 1954).
[15]A description of this effect for the case of small deflections was given by S. Liebes, Phys. Rev. B (1964–1965) **133,** 835 (1964).

Radiation fields in the Schwarzschild background*

James M. Bardeen†
University of Washington, Seattle, Washington

William H. Press‡
California Institute of Technology, Pasadena, California
(Received 29 March 1972; revised manuscript received 16 August 1972)

Scalar, electromagnetic, and gravitational test fields in the Schwarzschild background are examined with the help of the general retarded solution of a single master wave equation. The solution for each multipole is generated by a single arbitrary function of retarded time, the retarded multipole moment. We impose only those restrictions on the time dependence of the multipole moment which are required for physical regularity. We find physically well-behaved solutions which (i) do not satisfy the Penrose peeling theorems at past null infinity and/or (ii) do not have well-defined Newman–Penrose quantities. Even when the NP quantities exist, they are not measurable; they represent an "average" multipole moment over the infinite past, and their conservation is essentially trivial.

1. INTRODUCTION

Two general relativistic effects make it difficult to study the exact propagation of radiation fields. First, the curvature of the space–time manifold influences the propagation of the radiation. Second, the stress–energy of the radiation acts to produce curvature in the manifold. Acting in concert, these effects produce a nonlinear theory, with an extreme dearth of known, exact radiation solutions available for study.

In studying gravitational waves, it has frequently been useful to use the "linearized theory," in which the manifold is taken to be flat, and the waves are sufficiently weak that they do not destroy the flatness. Unfortunately, certain interesting phenomena vanish in the linearized case. For example, in general the propagation of radiation is not entirely along null characteristics, as Kundt and Newman[1] have shown for scalar and electromagnetic test fields in the Schwarzschild metric, as McLenaghan[2] has shown for scalar test fields in any nonflat background satisfying the vacuum Einstein equations, and as Bonnor and Rotenberg[3] have shown for asymptotically flat gravitational fields. The radiation *backscatters* off of nonuniformities in the curvature of the background space–time. For example, there is generally backscatter left behind a burst of outgoing radiation. Although the backscatter dies off in time at fixed radius, the field at any point in space does not become exactly static in a finite retarded time. Certain coefficients associated with the asymptotic field near future null infinity, the Newman–Penrose quantities (NPQ's)[4,5] are related to the backscatter from outgoing waves. In a flat background these coefficients measure properties of incoming waves and vanish identically when an outgoing-wave boundary condition is imposed. In curved space, the Einstein–Maxwell equations appear to guarantee that the NPQ's are conserved for dynamic fields; but investigations of their physical significance have been hampered by the absence of exact solutions with nontrivial NPQ's.

Backscatter and nontrivial NPQ's do not require the full nonlinear theory. They require that the background influence the radiation, but not vice versa. Thus they can be studied in detail for fields which are linearized about (i.e., weakly superimposed on) a curved background. The work of Price[6,7] on the behavior of integer–spin test fields in the collapse of a slightly nonspherical star, has furnished the key to this sort of an approach.

We have used Price's equations to analyze in detail the propagation of scalar, electromagnetic, and gravitational test fields in the Schwarzschild background. The sources of the fields are assumed to remain bounded for all time inside a radius $R > 2M$, where M is the gravitational mass (units with $c = G = 1$). We exhibit a single partial differential equation which fully describes the radiative part of the various test fields, and we solve this master equation for the general retarded solution in the region $r > R$. The solution is an expansion in powers of $(2M/r)$ which converges uniformly in this region at all retarded times.

With the general solution, we are able to examine the backscatter in some detail and to elucidate the nature and physical significance of the NPQ's. The solution also sheds considerable light on the "peeling theorems," which deal with the asymptotic radiation field at null infinity.

Our results for the NPQ's have been reported previously[8]; in this paper they are amplified from a somewhat different viewpoint. We find that the NPQ's do not always exist (i.e., the limits defining them do not always converge). When they do exist, they are a certain average of the value of the source's lowest radiatable multipole moment over the infinite past. The presence of this "average value" in the field is due to the superposition of backscatter from the outgoing radiation of all previous epochs. The conservation of the NPQ's has a simple interpretation: The contribution of the present finite epoch to the average of the infinite past is vanishingly small.

An important point is that the NPQ's, even when they exist, are not operationally measurable. An observational measurement of finite accuracy and duration, and at finite radius, can at best determine a quantity (we here call it a measurable NPQ or MNPQ) which is an average over the *recent* past (this is made precise in Sec. 6); and there is no observational way to tell whether this average agrees with the "primordial" NPQ or not.

Most previous theorems dealing with the asymptotic behavior of the fields at null infinity and, in particular, the peeling theorem of Penrose[9] based on a conformal treatment of infinity,[10] make certain mathematical regularity assumptions. For instance, Penrose assumes that the conformally transformed space–time manifold is C^4-differentiable everywhere (including future and past null infinity), with a C^3 metric. We ask, for retarded test fields in the Schwarzschild background whether all physically acceptable solutions of the field

equations are consistent with the assumptions of the Penrose theorem. The answer is no. Only solutions for which the gravitational quadrupole moment is asymptotically static in the infinite past satisfy the theorem for the gravitational field at past null infinity. Our general retarded field *is* consistent with the peeling theorem at future null infinity. We show that all solutions which are asymptotically regular in the Penrose sense at past null infinity possess NPQ's at future null infinity.

The mathematical foundation of this paper is the Newman–Penrose spin coefficient formalism, as adapted by Price[7] for test fields in the Schwarzschild metric. This is reviewed briefly in Sec. 2, and the equations for scalar, electromagnetic, and gravitational test fields are given. Section 3 solves these equations as special cases of a single "master" equation. Section 4 deals with the solutions for the lowest radiatable moment; Sec. 5 with the peeling theorems; and Sec. 6 the Newman–Penrose quantities.

2. FORMAL PRELIMINARIES

The conventional form of the Schwarzschild metric is

$$ds^2 = (1 - 2M/r)dt^2 - (1 - 2M/r)^{-1}dr^2 - r^2(d\theta + \sin^2\theta d\varphi^2). \quad (2.1)$$

Outgoing null geodesics are the surfaces of constant u, θ, φ, where u is the retarded time

$$u \equiv t - r - 2M \ln(r/2M - 1) \equiv t - r^*, \quad (2.2)$$

while for ingoing radial null geodesics,

$$v \equiv t + r + 2M \ln(r/2M - 1) \equiv t + r^* = \text{const}. \quad (2.3)$$

The radius r is both the proper circumferential radius governing the area of 2-spheres and an affine parameter along the radial null geodesics. We will always impose a boundary condition that there be no "free" incoming waves, so it is convenient to use the retarded time u and the radius r as coordinates. Then the metric is

$$ds^2 = (1 - 2M/r)du^2 + 2du\,dr - r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (2.4)$$

The Newman–Penrose spin coefficient formalism[11] is a powerful method for dealing with radiation in asymptotically flat space-times. It is based on a tetrad of complex, null 4-vectors $l^\mu, n^\mu, m^\mu, m^{*\mu}$ satisfying

$$\mathbf{l}\cdot\mathbf{n} = -\mathbf{m}\cdot\mathbf{m}^* = 1 \quad (2.5)$$

with all other dot products zero. All tensors can be reduced to (in general) complex scalars by contraction with members of this null tetrad. The "spin coefficients" are scalars constructed from covariant derivatives of the tetrad vectors. Newman–Penrose scalars have a *conformal weight* c and a *spin weight* p if under the transformation

$$\tilde{l}^\mu = \lambda l^\mu, \quad \tilde{n}^\mu = \lambda^{-1}n^\mu, \quad \tilde{m}^\mu = e^{i\eta}m^\mu, \quad (2.6)$$

the scalar T transforms as

$$\tilde{T} = \lambda^c e^{ip\eta}T. \quad (2.7)$$

In the Schwarzschild background a special choice for the null tetrad which simplifies the spin coefficients is, in u, r, θ, φ coordinates,

$$\begin{aligned} l^\mu &= [0, 1, 0, 0], \\ n^\mu &= [1, -\tfrac{1}{2}(1 - 2M/r), 0, 0], \\ m^\mu &= (1/\sqrt{2})[0, 0, 1/r, i/(r\sin\theta)]. \end{aligned} \quad (2.8)$$

Thus, l^μ is tangent to outgoing radial null geodesics and n^μ is tangent to ingoing radial null geodesics.

The physically measurable tensor associated with the electromagnetic field is the electromagnetic field tensor $F_{\mu\nu}$; that associated with the free gravitational field is the Weyl tensor $C_{\alpha\beta\gamma\delta}$ (in vacuum, identical to the Riemann tensor $R_{\alpha\beta\gamma\delta}$). The tetrad (2.8) is contracted with $F_{\mu\nu}$ to obtain the NP scalars for a test electromagnetic field

$$\begin{aligned} \Phi_1 &= F_{\mu\nu}l^\mu m^\nu, \\ \Phi_0 &= \tfrac{1}{2}F_{\mu\nu}(l^\mu n^\nu - m^\mu m^{*\nu}), \\ \Phi_{-1} &= F_{\mu\nu}m^{*\mu}n^\nu. \end{aligned} \quad (2.9)$$

Fortuitously, the subscript denotes both the spin weight and the conformal weight of the scalar. In terms of the physical electric and magnetic field components measured by an observer at rest in the Schwarzschild metric (2.1),

$$\begin{aligned} \Phi_{+1} &= 2^{-1/2}(1 - 2M/r)^{-1/2}\,[E^{(\theta)} - B^{(\varphi)} + i(E^{(\varphi)} + B^{(\theta)})] \\ \Phi_0 &= -\tfrac{1}{2}\,(E^{(r)} + iB^{(r)}) \\ \Phi_{-1} &= -2^{-3/2}(1 - 2M/r)^{1/2}[E^{(\theta)} + B^{(\varphi)} - i(E^{(\varphi)} - B^{(\theta)})]. \end{aligned} \quad (2.10)$$

The Φ_p are completely equivalent to $F_{\mu\nu}$—each contains six independent real functions.

Similarly, there are five NP scalars containing ten independent real functions which are equivalent in information content to the Weyl tensor:

$$\begin{aligned} \Psi_2 &= -C_{\alpha\beta\gamma\delta}l^\alpha m^\beta l^\gamma m^\delta, \quad \Psi_1 = -C_{\alpha\beta\gamma\delta}l^\alpha n^\beta l^\gamma m^\delta, \\ \Psi_0 &= -C_{\alpha\beta\gamma\delta}l^\alpha m^\beta m^{*\gamma}n^\delta, \quad \Psi_{-1} = -C_{\alpha\beta\gamma\delta}l^\alpha n^\beta n^\gamma m^{*\delta}, \\ \Psi_{-2} &= -C_{\alpha\beta\gamma\delta}n^\alpha m^{*\beta}n^\gamma m^{*\delta}. \end{aligned} \quad (2.11)$$

Again, the subscript gives the spin weight and the conformal weight. (The notation here follows Price[7] and differs somewhat from most authors.) For gravitational perturbations there is one additional complication: The null tetrad to be used in (2.11) is the NP special null tetrad associated with the perturbed metric, *not* the tetrad (2.8). The only Ψ_p which is nonzero in the unperturbed Schwarzschild background is $\Psi_0 = -M/r$.[3] The real parts of the Φ_p and Ψ_p are associated with even-parity fields, and the imaginary parts are associated with odd-parity fields. The letter p is used to denote the spin weight, since we reserve the letter s for the spin of the field.

It is natural to take advantage of the spherical symmetry of the background to expand the perturbations in spherical harmonics. However, the appropriate spherical harmonics for the NP scalars with nonzero spin weight are not ordinary scalar spherical harmonics, but rather the spin-weight-p spherical harmonics.[12] These harmonics are denoted by ${}_pY^l_m(\theta, \varphi)$ and involve derivatives of the ordinary spherical harmonics, which have

spin weight zero. The spin-weight index p can be increased or decreased by certain differential operators. The spin-weight-p spherical harmonics with $l < |p|$ are undefined.

The derivation of the equations governing scalar, electromagnetic, and gravitational test fields in the Schwarzschild background is described by Price.[7] Before he expands in spherical harmonics, Price "despins" Newman–Penrose scalars with nonzero spin weight. This differs from conventional practice.[4,13] Therefore, we expand a Newman–Penrose scalar of spin weight p directly in spin-weight-p spherical harmonics. For example,

$$\Phi_p(u, r, \theta, \varphi) = \sum_{l=|p|}^{\infty} {}_p\Phi^l{}_m(u, r)\,{}_pY^l{}_m(\theta, \varphi). \tag{2.12}$$

To avoid an unnecessarily complicated notation we suppress the l, m indices and write ${}_p\Phi^l{}_m(u, r)$ as $\Phi_p(u, r)$. Since the equations for the different multipoles separate, this never causes any confusion.

The differential operators in the spin-coefficient formalism which contain derivatives with respect to u and r are

$$D = l^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial r} \tag{2.13a}$$

and

$$\Delta = n^\mu \frac{\partial}{\partial x^\mu} = \frac{\partial}{\partial u} - \frac{1}{2}\left(\frac{1-2M}{r}\right)\frac{\partial}{\partial r} \tag{2.13b}$$

in the Schwarzschild background.

Of course, the spin-coefficient formalism is not needed for a scalar test field ψ, which satisfies

$$\Box\psi = (-g)^{-1/2}\frac{\partial}{\partial x^\mu}\left((-g)^{1/2}g^{\mu\nu}\frac{\partial\psi}{\partial x^\nu}\right) = 0. \tag{2.14}$$

After expanding in ordinary scalar spherical harmonics, the equation for the 2^l-pole is

$$2\frac{\partial^2\psi}{\partial r\partial u} + \frac{2}{r}\frac{\partial\psi}{\partial u}\left[\left(1-\frac{2M}{r}\right)\frac{\partial\psi}{\partial r}\right] + \frac{l(l+1)}{r^2}\psi = 0. \tag{2.15}$$

The equations governing the 2^l-pole of an electromagnetic test field are

$$D[r^2\Phi_0] = [\tfrac{1}{2}l(l+1)]^{1/2}r\Phi_1, \tag{2.16a}$$

$$D[r\Phi_{-1}] = [\tfrac{1}{2}l(l+1)]^{1/2}\Phi_0, \tag{2.16b}$$

$$\Delta[(1-2M/r)r\Phi_1] = -[\tfrac{1}{2}l(l+1)]^{1/2}(1-2M/r)\Phi_0, \tag{2.16c}$$

$$\Delta[r^2\Phi_0] = -[\tfrac{1}{2}l(l+1)]^{1/2}r\Phi_{-1}. \tag{2.16d}$$

These combine to give decoupled second-order differential equations for each of the $\Phi_p(u, r)$:

$$D\{(1-2M/r)^{-1}r^2\Delta[(1-2M/r)r\Phi_1]\} + \tfrac{1}{2}l(l+1)r\Phi_1 = 0, \tag{2.17a}$$

$$D\Delta[r^2\Phi_0] + \tfrac{1}{2}l(l+1)\Phi_0 = 0, \tag{2.17b}$$

$$\Delta[r^2D(r\Phi_{-1})] + \tfrac{1}{2}l(l+1)r\Phi_{-1} = 0. \tag{2.17c}$$

If any *one* of Eqs. (2.17) is solved, the corresponding solutions for the other two Φ_p are immediately obtained from Eqs. (2.16) as derivatives of the first Φ_p. Price works with (2.17b); in Sec. 3 we solve Eqs. (2.17a) and (2.17c).

The equations for the gravitational test field are considerably more complicated, since the perturbations involve the very geometry through which the waves propagate. The spin coefficients $\rho, \lambda, \mu, \nu, \sigma, \tau$ and the metric functions U, ω appear in the equations for the $\Psi_p(u, r)$. In the equations governing a particular 2^l-pole these subsidiary quantities, like the Ψ_p, are interpreted as the coefficients of the appropriate spin-weight spherical harmonics. The functions ρ, μ, and U have spin weight zero; τ and ω have spin weight + 1; ν, λ, and σ have spin weights − 1, − 2, and + 2, respectively. For those quantities nonzero in the back ground we distinguish the perturbations by a subscript B.

The gravitational analogs of Eqs. (2.16) are the perturbed Bianchi identities:

$$D[r^4\Psi_1] = [\tfrac{1}{2}(l-1)(l+2)]^{1/2}r^3\Psi_2; \tag{2.18a}$$

$$D[r^3\Psi_{0_B}] = [\tfrac{1}{2}l(l+1)]^{1/2}r^2\Psi_1 - 3M\rho_B; \tag{2.18b}$$

$$D[r^2\Psi_{-1}] = [\tfrac{1}{2}l(l+1)]^{1/2}r\Psi_{0_B} - 3Mr^{-2}\omega^*; \tag{2.18c}$$

$$D[r\Psi_{-2}] = [\tfrac{1}{2}(l-1)(l+2)]^{1/2}\Psi_{-1} + 3M^{-2}\lambda; \tag{2.18d}$$

$$(1-2M/r)^{-2}\Delta[(1-2M/r)^2r\Psi_2] = -[\tfrac{1}{2}(l-1)(l+2)]^{1/2}\Psi_1 - 3Mr^{-2}\sigma; \tag{2.18e}$$

$$(1-2M/r)^{-1}\Delta[(1-2M/r)r^2\Psi_1] = -[\tfrac{1}{2}l(l+1)]^{1/2}r\Psi_{0_B} + 3Mr^{-1}(\tau + r^{-1}\omega); \tag{2.18f}$$

$$\Delta[r^3\Psi_{0_B}] = -[\tfrac{1}{2}l(l+1)]^{1/2}r^2\Psi_{-1} + 3M(\mu_B - r^{-1}U_B); \tag{2.18g}$$

$$(1-2M/r)\Delta[(1-2M/r)^{-1}r^4\Psi_{-1}] = -[\tfrac{1}{2}(l-1)(l+2)]^{1/2}r^3\Psi_{-2} - 3Mr\nu. \tag{2.18h}$$

The other equations[7,11] relating the metric perturbations, the perturbations in the spin coefficients, and the Ψ_p are sufficiently complicated that it does not seem possible to combine them with Eqs. (2.18) to get a decoupled second-order differential equation for each of the Ψ_p. Price does derive such an equation for $\mathrm{Im}\Psi_{0_B}$.

However, Price turns to the Regge–Wheeler formalism,[14] as further developed by Zerilli,[15] to treat the even-parity gravitational perturbations.

Fortunately, decoupled equations do exist for Ψ_2 and Ψ_{-2}. The additional equations required are

$$D(r^2\sigma) = r^2\Psi_2 \tag{2.19}$$

and

$$(1-2M/r)\Delta[(1-2M/r)^{-1}r^2\lambda] = [\tfrac{1}{2}(l-1)(l+2)]^{1/2}r\nu - r^2\Psi_{-2}. \tag{2.20}$$

Equations (2.18a), (2.18e), and (2.19) combine to give

$$D\{(1-2M/r)^{-2}r^4\Delta[(1-2M/r)^2r\Psi_2]\} + [\tfrac{1}{2}(l-1)(l+2) + 3M/r]r^3\Psi_2 = 0, \tag{2.21a}$$

while Eqs. (2.18d), (2.18h), and (2.20) give

$$\Delta\{(1-2M/r)^{-1}r^4 D[r\Psi_{-2}]\} + [\tfrac{1}{2}(l-1)(l+2)+3M/r](1-2M/r)^{-1}r^3\Psi_{-2}=0. \tag{2.21b}$$

The outgoing radiation field near future null infinity is contained in Ψ_{-2} and the Newman–Penrose quantities are in Ψ_2; so, for our purposes a complete solution for all of the spin coefficients and the remianing Ψ_p is not necessary.

We shall see below that Eqs. (2.15), (2.17a), and (2.17c), for scalar and electromagnetic test fields, are all special cases of a single "master" equation. It is a rather remarkable fact that Eqs. (2.21a) and (2.21b), which govern the radiative behavior of *gravitational* test fields, are also special cases of the same equation. In this sense the Einstein field equations, with their particular coupling of the perturbations to the background geometry, represent the simplest spin-2 field equations in the curved Schwarzschild background. The solutions to the master equation governing all the fields depend explicitly on s (the spin of the perturbing field) only in a very minimal way.

3. THE GENERAL RETARDED SOLUTION

We consider the equation

$$2\frac{\partial^2\psi}{\partial u\partial r}+\frac{2(s+p+1)}{r}\frac{\partial\psi}{\partial u} -\frac{\partial^2\psi}{\partial r^2}-\frac{(2s+2)}{r}\frac{\partial\psi}{\partial r}+\frac{(l+s+1)(l-s)}{r^2}\psi +\left(\frac{2M}{r}\right)\left(\frac{\partial^2\psi}{\partial r^2}+\frac{(2s+1-p)}{r}\frac{\partial\psi}{\partial r}+\frac{s(s-p)}{r^2}\psi\right)=0. \tag{3.1}$$

The parameter s takes on the values 0, 1, or 2 corresponding to the spin of the test field. The parameter p takes on the values $\pm s$ corresponding to the two extreme possible spin weights. [Eq. (3.1) can *not* be used for "nonradiative" spin-weight components $-s+1 \le p \le s-1$.] With $s=0, p=0, \psi(u,r)$ is the coefficient of $Y^l{}_m(\theta,\varphi)$ in the spherical harmonic expansion of a scalar test field, and Eq. (3.1) is identical to the field equation (2.15). With $s=1, p=\pm 1, \psi$ represents the coefficient of ${}_{\pm1}Y^l{}_m$ in the spin weight ± 1 part of the electromagnetic field tensor; (3.1) then is identical to (2.17a) and (2.17c). With $s=2, p=\pm 2, \psi$ represents the coefficient of ${}_{\pm2}Y^l{}_m$ in the spin weight ± 2 part of the Weyl tensor, and (3.1) becomes identical to (2.21a) and (2.21b). In this section we obtain a general retarded solution to this master equation.

The solution is in the form of an expansion in powers of the gravitational mass M. We prove that the expansion converges for retarded fields at all $r > R > 2M$, where R is a radius bounding both the source of the background Schwarzschild metric (either a star or a black hole) and the source of the test field at all times to the past.

The general retarded solution to Eq. (3.1) must be regular at infinity in the minimal sense that $\psi \to 0$ as $r \to \infty$, and must be entirely generated by sources in the region $r < R$. It will contain one arbitrary function of the retarded time u.

First consider static solutions. Since $r^{-1}=0$ is a regular singular point of the ordinary differential equation for $\psi(r)$, the static solution regular at infinity can be written as the series

$$\psi(r) = A\,2^{(p-s)/2}\,r^{-(l+s+1)}\left[1+\sum_{k=1}^{\infty}a_k\left(\frac{2M}{r}\right)^k\right] \tag{3.2}$$

with

$$a_k=\frac{(l+k)!}{l!k!}\frac{(l+p+k)!}{(l+p)!}\frac{(2l+1)!}{(2l+1+k)!}. \tag{3.3}$$

The series converges for all $r > 2M$. The coefficient A is identified as the static multipole moment.

Now consider solutions to Eq. (3.1) which are static for all $u \le u_0$, but dynamic for $u > u_0$. These solutions are retarded, since no incoming waves are present near past null infinity. At $u = u_0$, ψ can be expanded in powers of r^{-1} at fixed u, and this analytic structure will persist for a finite retarded time after $u = u_0$ in the region $r > R$. Let

$$\psi = \sum_n f_n(u) r^{-n} \tag{3.4}$$

and substitute into Eq. (3.1). The f_n must satisfy the hierarchy of equations.

$$2(n-p-s-1)f_n' = (l+s+2-n)(n+l-s-1)f_{n-1} + (2M)(n+p-s-2)(n-s-2)f_{n-2}. \tag{3.5}$$

It is consistent with Eq. (3.5) that all f_n with $n < p+s+1$ are identically zero. Furthermore, these f_n must be zero, or some f_n with $n \le 0$ will be nonzero, and $\psi(u,r)$ will not go to zero as $r \to \infty$ at $u > u_0$. This is the peeling property at future null infinity.

Split the sum (3.4) into two parts:

$$\psi_{\rm I} = \sum_{n=p+s+1}^{l+s+1} f_n(u) r^{-n} \tag{3.6}$$

and

$$\psi_{\rm II} = \sum_{n=l+s+2}^{\infty} f_n(u) r^{-n}. \tag{3.7}$$

We shall see that the f_n in $\psi_{\rm I}$ are linear combinations of a single function of retarded time, the retarded multipole moment $A(u)$, and its first $(l-p)$ derivatives. The f_n in $\psi_{\rm II}$ cannot be represented in this way consistent with Eqs. (3.5) and the static initial conditions.

Define the multipole moment $A(u)$ from the part of the field with $p = s$. Let

$$f_{2s+1}(u) = \lim_{r\to\infty}[r^{2s+1}\psi(u,r)] = \frac{2^{l-s}(l+s)!}{(2l)!}A^{(l-s)}(u). \tag{3.8}$$

(Superscripts in parentheses denote the number of derivatives to be taken.) In the first $(l-s)$ successive integrations of Eqs. (3.5), with $p = s$, absorb the constants of integration into $A(u)$. Then $f_{l+s+1} = A$ when $A(u)$ is constant, consistent with Eq. (3.2).

The coefficient of $r^{-1}, f_1(u)$, in the $\psi(u,r)$ with $p=-s$ is related to $f_{2s+1}(u)$ in the $\psi(u,r)$ with $p=s$ by the flat-space versions of Eqs. (2.16c) and (2.16d) coupling the Φ_p $(s=1)$ and Eqs. (2.18e)–(2.18h) coupling the Ψ_p $(s=2)$. The terms in these equations from the curvature of the background are of order r^{-1}; at least as long as the infinite sum in $\psi_{\rm II}$ converges, one has

$$\frac{\partial\psi}{\partial r} = \mathcal{O}(r^{-1}\psi), \tag{3.9}$$

just as for the flat-space retarded solutions. Therefore, the analog of Eq. (3.8) when $p = -s$ is

$$f_1(u)\Big|_{p=-s} = 2^s \frac{(l-s)!}{(l+s)!} f_{2s+1}{}^{(2s)}\Big|_{p=s} = 2l \frac{(l-s)!}{(2l)!} A^{(l+s)}(u). \quad (3.10)$$

Starting with either Eq. (3.8) or Eq. (3.10), successive integrations of Eqs. (3.5) give the $f_n(u)$ in $\psi_{\rm I}(u,r)$ in the form

$$f_{p+s+1+k} = \sum_{m=0}^{[k/2]} \alpha_{k,m} (2M)^m A^{(l-p-k+m)}. \quad (3.11)$$

The upper limit to the sum $[k/2]$ is

$$[k/2] = \begin{cases} k/2, & k \text{ even} \\ (k-1)/2, & k \text{ odd} \end{cases}. \quad (3.12)$$

All the coefficients $\alpha_{k,m}$ can easily be evaluated for any particular values of s, l, and p [see Eq. (4.19) and following, for an example]; the coefficients which survive when $M = 0$ are

$$\alpha_{k,0} = 2^{l-k} 2^{-(p+s)/2} \frac{(l+p+k)!}{(2l)!} \frac{(l-p)!}{k!(l-p-k)!}. \quad (3.13)$$

When the field is static, the only nonzero f_n in $\psi_{\rm I}$ is

$$f_{l+s+1} = 2^{(p-s)/2} A. \quad (3.14)$$

Because the coefficient of f_{n-1} in Eq. (3.5) vanishes when $n = l + s + 2$, the constant of integration in f_{l+s+2} cannot be absorbed in $\int^u A(u')du'$ as, for instance, the constant of integration in f_{l+s+1} was absorbed in $A(u)$. Instead, it must be kept explicitly:

$$f_{l+s+2} = C + \frac{l(l+p)}{2(l-p+1)} \sum_{m=0}^{[(l-p-1)/2]} \alpha_{l-p-1,m} (2M)^{m+1} A^{(m)}. \quad (3.15)$$

From the static initial condition on f_{l+s+2},

$$C = \frac{1}{2} 2^{(p-s)/2} \frac{(2l+1)}{l-p+1} (2M) A(u_0). \quad (3.16)$$

The fact that the constant of integration contains $A(u_0)$ means that $f_{l+s+2}(u)$ for $u > u_0$ depends on the past history of the time dependence of $A(u)$, as well as on the instantaneous values of $A(u)$ and its derivatives. The backscatter of the outgoing radiation field $\psi_{\rm I}(u,r)$ is entirely contained in $\psi_{\rm II}(u,r)$. It was the failure to allow for the constant of integration (3.16) that led to the incorrect treatment of the backscatter in preprint versions of papers by Price[7] and Thorne[16] on the decay of radiatable multipoles during gravitational collapse.

The integration constant in f_{l+s+2} generates terms in the f_n with $n > l + s + 2$ which grow with time:

$$f_{l+s+1+k} \sim (-1)^{k-1} \frac{(k-1)!}{2^k} \frac{(l-p)!}{(l-p+k)!} \times \frac{(2l+k)!}{(2l)!} 2^{(p-s)/2} (2M) A(u_0) u^{k-1}. \quad (3.17)$$

While these terms may be partially cancelled by terms coming from successive integrals of $A(u)$, typically

$$\frac{f_{l+s+k+1}}{f_{l+s+k}} \sim -\frac{u-u_0}{2} \quad (3.18)$$

in the limit $k \gg l$ when $u - u_0 \gg 2M$, so the expansion (3.7) of $\psi_{\rm II}$ in powers of r^{-1} will diverge once

$$u - u_0 > 2r. \quad (3.19)$$

Thus, the power series expansion of the form (3.4) is not a satisfactory solution of Eq. (3.1).[3]

To obtain a solution which converges uniformly at *all* future times, we keep $\Psi_{\rm I}$ in the form (3.6), but represent $\psi_{\rm II}(u,r)$ by

$$\psi_{\rm II}(u,r) = r^{-(l+s+1)} \sum_{k=1}^{\infty} a_k \left(\frac{2M}{r}\right)^k g_k(u,r). \quad (3.20)$$

This new expansion is an expansion in powers of the gravitational mass M, instead of powers of r^{-1}. The coefficients a_k are the coefficients (3.3) in the static solution. For the purposes of the new expansion $\psi_{\rm I}(u,r)$ is considered zeroth order in M, even though M appears in the $f_n(u)$, $n \le l + s + 1$, through Eq. (3.11).

Substitute the expression (3.20) into Eq. (3.1), along with $\psi_{\rm I}(u,r)$ in the form (3.6), and require that the coefficent of each explicit power of M vanish. The result is a hierarchy of partial differential equations for the $g_k(u,r)$: When $k > 1$,

$$2\frac{\partial^2 g_k}{\partial u \partial r} - \frac{2(k+l-p)}{r}\frac{\partial g_k}{\partial u} - \frac{\partial^2 g_k}{\partial r^2} + \frac{2(k+l)}{r}\frac{\partial g_k}{\partial r} - \frac{k(k+2l+1)}{r^2} g_k = -\frac{k(k+2l+1)}{(k+l)(k+l+p)}\left(\frac{\partial^2 g_{k-1}}{\partial r^2} - \frac{(2k+2l+p-1)}{r}\frac{\partial g_{k-1}}{\partial r} + \frac{(k+l)(k+l+p)}{r^2} g_{k-1}\right); \quad (3.21)$$

and when $k = 1$,

$$2\frac{\partial^2 g_1}{\partial u \partial r} - \frac{2(l-p+1)}{r}\frac{\partial g_1}{\partial u} - \frac{\partial^2 g_1}{\partial r^2} + \frac{2(l+1)}{r}\frac{\partial g_1}{\partial r} - \frac{(2l+2)}{r^2} g_1 = -\frac{(2l+2)}{r^2} f_{l+s+1}(u) - \frac{2l(l+p)}{l+p+1}\frac{f_{l+s}(u)}{r}. \quad (3.22)$$

The right-hand side of Eq. (3.22) comes from using Eqs. (3.5) on the f_n in $\psi_{\rm I}$.

Equations (3.21) are scale invariant under the transformation $u \to Ku$, $r \to Kr$. Equation (3.22) is not generally scale invariant; but it is if the multipole moment $A(u)$ is constant, which implies that f_{l+s+1} is constant and f_{l+s} is zero. In this special case the entire hierarchy of equations for the g_k is invariant under the scale transformation. The scale invariance suggests that a solution to the hierarchy exists which depends on only one independent variable, a scale invariant combination of u and r. Since the equations are also invariant under a translation in u when $A(u)$ is constant, the most general form for the similarity variable is

$$y = (u - u_1)/2r. \quad (3.23)$$

The similarity solutions will be superimposed to give the general retarded solution.

The ansatz $g_k(u, r) = g_k(y)$ reduces the partial differential equations (3.21) and (3.22) to the ordinary differential equations

$$y(1+y)\frac{d^2g_k}{dy^2} + [k+l-p+1+2(k+l+1)y]\frac{dg_k}{dy} + k(k+2l+1)g_k = \frac{k(k+2l+1)}{(k+l)(k+l+p)}\left(y^2\frac{d^2g_{k-1}}{dy^2} + (2k+2l+p+1)y\frac{dg_{k-1}}{dy} + (k+l)(k+l+p)g_{k-1}\right). \tag{3.24}$$

When $k = 1$ in Eq. (3.24) the g_0 appearing on the right-hand side is understood to be f_{l+s+1}, a constant.

The solution of Eqs. (3.24) can be reduced to quadratures by standard methods. A particular solution to the hierarchy is the solution, for which

$$g_k(y) = f_{l+s+1} = 2^{(p-s)/2}A, \tag{3.25}$$

for all $k \geq 1$. Any dynamic solution to the hierarchy is a particular solution plus a homogeneous solution. To join a dynamic solution to a static solution at $u = u_0$ it is necessary to take $u_1 = u_0$, or

$$y = (u - u_0)/2r, \tag{3.26}$$

since only at $y = 0$ is y independent of r at fixed u. Therefore, the homogeneous solutions for constructing initially static dynamic solutions must be regular at $y = 0$.

Only one of the two independent homogeneous solutions to the kth equation (3.24) is regular at $y = 0$. Normalized to be one at $y = 0$, it is

$$h_k(y) = (1+y)^{-(k+l+p)}\sum_{m=0}^{l+p}\frac{(l+p)!}{m!(l+p-m)!} \times \frac{(l-p+k)!}{l-p+k+m)!}\frac{(l-p+m)!}{(l-p)!}y^m. \tag{3.27}$$

A homogeneous solution to the hierarchy is composed of inhomogeneous solutions to Eq. (3.24) for all $k > n$, generated by the homogeneous solution (3.27) for $k = n$. Let the functions $H_{n,k}(y)$ be the $g_k(y)$ generated by $h_n(y)$:

$$g_k(y) = H_{n,k}(y). \tag{3.28}$$

For $k < n$;

$$H_{n,k} = 0. \tag{3.29}$$

The nonzero $H_{n,k}$ are all normalized to be one at $y = 0$. Thus

$$H_{n,n}(y) = h_n(y) \tag{3.30}$$

and for $k > n$

$$H_{n,k}(y) = h_k(y)\left(1 + \int_0^y dy_1 y_1^{-(k+l-p+1)}(1+y_1)^{-(k+l+p+1)} \times h_k(y_1)^{-2}\int_0^{y_1} dy_2 y_2^{k+l-p}(1+y_2)^{k+l-p}h_k(y_2)S_k(y_2)\right). \tag{3.31}$$

The function S_k is the right-hand side of the kth equation (3.24), with $g_{k-1} = H_{n,k-1}$.

Some important properties of the nonzero $H_{n,k}$ are

$$H_{n,k} = 1 - O(y^{k-n+1}) \tag{3.32}$$

in the limit $y \ll 1$, while when $y \gg 1$

$$H_{n,k} \simeq \frac{k!}{(k-n)!}\frac{(k+2l+1)!}{(k+2l+1+n)!}\frac{(k+l-n)!}{(k+l)!} \times \frac{(k+l+p-n)!}{(k+l+p)!}\frac{(l+n)!}{l!n!}\frac{(l-p+n)!}{(l-p)!}\frac{(l+p+n)!}{(l+p)!} \times \frac{(2l)!}{(2l+n)!}\frac{(2l+1)!}{(2l+n+1)!}y^{-n}. \tag{3.33}$$

The nonzero $H_{n,k}$ decreases monotonically from one at $y = 0$ to zero in the limit $y \to \infty$.

The leading term in a homogeneous solution $H_{n,n}(y)$ can be interpreted as an ingoing wave in flat space. That is,

$$\psi = h_n(y)\, r^{-(l+s+1+n)} \tag{3.34}$$

solves Eq. (3.1) with $M = 0$ and can be put in the form

$$\psi = \sum_{q=0}^{l+p} 2^{-(s+p)/2}(-2)^q\frac{(2l-q)!}{(2l)!} \times \frac{(l+p)!}{q!(l+p-q)!}B^{(q)}(v)r^{-(l+s+1-q)}. \tag{3.35}$$

The advanced multipole moment $B(v)$, as a function of the flat-space advanced time

$$v = u + 2r, \tag{3.36}$$

is

$$B(v) = 2^{(s+p)/2}\frac{(2l)!}{(2l+n)!}\frac{(l-p+n)!}{(l-p)!}2^n(v-u_0)^{-n}. \tag{3.37}$$

A similarity solution solves the following problem: The field is static for all $u < u_0$; at $u = u_0$ the retarded multipole moment changes by a step function to a new constant value for all $u > u_0$. With the help of the above homogeneous solutions it is possible to fit the initial conditions on the $g_k(y)$ at u infinitesimally greater than u_0. The instantaneous changes in the g_k due to the change in $A(u)$ are found from Eqs. (3.5).

An arbitrary continuous variation of $A(u)$ can be approximated arbitrarily closely by a superposition of step-function changes. Since the test-field equations are linear, the general retarded solution to Eq. (3.1) can be represented as a continuous superposition of similarity solutions. The constant u_0 in the similarity variable y becomes a dummy integration variable. By letting the range of integration extend to $u_0 = -\infty$, we include cases in which the field was never static at any time in the past.

Our general retarded solution for $g_k(u, r)$ is

$$g_k(u,r) = f_{l+s+1}(u) - \int_{-\infty}^{u} du_0\left[\left(\frac{df_{l+s+1}}{du_0} - \frac{l(l+p)}{(l-p+1)(l+p+1)}f_{l+s}(u_0)\right)H_{1,k}(y) + \frac{l(l+p)}{(l-p+1)(l+p+1)}f_{l+s}(u_0)H_{2,k}(y)\right]. \tag{3.38}$$

That this does indeed solve Eqs. (3.21) and (3.22) can easily be checked by substitution. The integral in Eq. (3.38) is the *backscatter*; when $k = 1$ it is a superposition of purely ingoing waves generated by previous changes in the multipole moment.

While Eq. (3.38) is best for the physical interpretation of $g_k(u, r)$, a different form of the solution is best for proving convergence of the integrals and of the series (3.20). Integrate by parts in Eq. (3.38) and define

$$F_{l+s}(u) \equiv \frac{l(l+p)}{(l-p+1)(l+p+1)} \int^u f_{l+s}(u_0)du_0$$

$$\equiv \frac{l(l+p)}{(l-p+1)(l+p+1)} \times \sum_{m=0}^{[(l-p-1)/2]} \alpha_{l-p-1,m}(2M)^m A^{(m)}(u). \quad (3.39)$$

When r is finite, so that $u_0 \to -\infty$ implies $y \to \infty$, the result is

$$g_k(u, r) = F_{l+s}(u)\delta_{k1} - \frac{1}{2r}\int_{-\infty}^u du_0\{[f_{l+s+1}(u_0) - F_{l+s}(u_0)]H_{1,k}'(y) + F_{l+s}(u_0)H_{2,k}'(y)\}. \quad (3.40)$$

The primes denote derivatives with respect to y, and δ_{kl} is the Kronecker delta.

In going from Eq. (3.38) to Eq. (3.40) we have implicitly assumed that $f_{l+s+1}(u)$ and $F_{l+s}(u)$ are bounded in the limit $u \to -\infty$. We now impose the slightly stronger condition that $f_{l+s+1}(u_0)$ and $F_{l+s}(u_0)$ be bounded for all $u_0 < u$, if the field is being evaluated at the retarded time u. Both $f_{l+s+1}(u_0)$ and $F_{l+s}(u_0)$ contain at most $(l-p)$ derivatives of $A(u_0)$, so the condition follows if $\psi_I(u_0, r)$ was bounded at all times to the past.

Since the $H_{n,k}$ decrease monotonically from one at $y = 0$ to zero at $y = \infty$, the $H_{n,k}'(y)$ in Eq. (3.40) are negative or zero over the whole range of integration. If $f_{l+s+1}(u_0)$ and $F_{l+s}(u_0)$ satisfy

$$|f_{l+s+1}(u_0)| \le K_1, \quad (3.41)$$

$$|F_{l+s}(u_0)| \le K_2 \quad (3.42)$$

for all $-\infty < u_0 \le u$, then

$$-K_1 \le -\frac{1}{2r}\int_{-\infty}^u du_0 f_{l+s+1}(u_0)H_{1,k}'(y) \le -\frac{K_1}{2r}\int_{-\infty}^u du_0 H_{1,k}'(y) = K_1, \quad (3.43)$$

$$-K_2 \le -\frac{1}{2r}\int_{-\infty}^u du_0 F_{l+s}(u_0)H_{n,k}'(y) \le K_2, \quad (3.44)$$

so

$$|g_k(u, r)| \le K_1 + 2K_2. \quad (3.45)$$

Both K_1 and K_2 are the same order as the bound on $|A(u_0)|$, since the time scale over which $A(u)$ changes is typically greater than or equal to $(2M)$.

The integrals in Eq. (3.38) will not necessarily converge to any definite value at $r = \infty$, where $y = 0$ for all u_0. Since $r = \infty$ is not in the physical space-time, there is not physical requirement that the $g_k(u, \infty)$ have well-defined values.

Derivatives of the g_k with respect to u and r do not affect the convergence of the integrals, since

$$\frac{\partial}{\partial u}H_{n,k}(y) = \frac{1}{2r}H_{n,k}'(y) \sim \frac{1}{r}y^{-(n+1)} \quad (3.46)$$

and

$$\frac{\partial}{\partial r}H_{n,k}(y) = -\frac{y}{r}H_{n,k}'(y) \sim \frac{1}{r}y^{-n}, \quad (3.47)$$

$y \gg 1$. Equation (3.9) is valid for the general solution, not only initially static solutions.

From Eq. (3.45), the absolute values of the $g_k(u, r)$ are bounded uniformly in k. We conclude that the infinite sum in Eq. (3.20) for $\psi_{II}(u, r)$ is absoutely convergent at all $r > R$, for any $R > 2M$, and that

$$\psi_{II}(u, r) = \mathcal{O}(r^{-(l+s+2)}). \quad (3.48)$$

The only restrictions on the time dependence of the retarded multipole moment $A(u)$ are boundedness conditions on $A(u)$ and its first $[(l-p)/2]$ derivatives. These are physically necessary conditions if the field is to have finite energy density at all times to the past. Our general retarded solution constructed from Eqs. (3.6) and (3.20), with the $f_n(u)$ given by Eq. (3.11) and the $g_k(u, r)$ given by Eq. (3.38) or (3.40), almost certainly contains all physically nonsingular retarded solutions to Eq. (3.1).

Some results of this section are not new. The solutions of Couch *et al.*[17,18] for the backscatter of electromagnetic and gravitational radiation first order in M are essentially the same as Eq. (3.40), with $k = 1$.

4. THE LOWEST RADIATABLE MULTIPOLES

The physically most important multipoles are the electromagnetic dipole and the gravitational quadrupole. These typically dominate in electromagnetic and gravitational radiation processes, respectively. They are the lowest multipoles which can radiate, i.e., contribute r^{-1} terms in the respective field tensors at future null infinity. Furthermore, these multipoles contain the apparently conserved Newman–Penrose quantities. In this section we write out explicitly the general retarded solutions for $\Phi_{\pm 1}$ (electromagnetic) and $\Psi_{\pm 2}$ (gravitational) through order $(2M/r)$ in all cases and through order $(2M/r)^2$ for Ψ_{-2}. Higher-order terms do not contribute to the NPQ's.

A. The electromagnetic dipole field

In view of Eq. (3.8) the retarded electromagnetic dipole moment $E(u)$ is defined by

$$E(u) \equiv \lim_{r\to\infty}[r^3\Phi_1(u, r)] \quad (4.1)$$

in the dipole part of the field. We have shown in Sec. 3 that this limit always exists for retarded fields.

In the spin-weight-one part of the dipole field the function $h_1(y)$ is

$$h_1(y) = (1 + y + \tfrac{1}{3}y^2)/(1+y)^3, \quad (4.2)$$

so that

$$g_1(u, r) = E(u) - \int_{-\infty}^u du_0 \frac{dE}{du_0}\frac{1 + y + \frac{1}{3}y^2}{(1+y)^3}, \quad (4.3a)$$

with $y = (u - u_0)/2r$, or

$$g_1(u, r) = \int_0^\infty dy\, E(u_0)\frac{2 + \frac{4}{3}y + \frac{1}{3}y^2}{(1+y)^4} \quad (4.3b)$$

with

$$u_0 = u - 2ry. \tag{4.4}$$

The result for $\Phi_1(u,r)$ is

$$\Phi_1(u,r) = r^{-3}E(u) + \tfrac{3}{2}(2M)r^{-4}\int_0^\infty dy E(u_0)\,\frac{2 + \frac{4}{3}y + \frac{1}{3}y^2}{(1+y)^4} + \mathcal{O}[(2M/r)^2]. \tag{4.5}$$

Alternatively, we could have begun with the spin-weight-minus-one part of the dipole field. Here Eq. (3.10) gives

$$f_1(u) = E^{(2)}(u). \tag{4.6}$$

Applying Eq. (3.5) twice,

$$f_2(u) = E^{(1)}(u) \tag{4.7}$$

and

$$f_3(u) = f_{l+s+1}(u) = \tfrac{1}{2}E(u). \tag{4.8}$$

The function $h_1(y)$ is simply

$$h_1(y) = (1+y)^{-1}, \tag{4.9}$$

so

$$g_1(u,r) = \frac{1}{2}E(u) - \int_{-\infty}^{u} du_0\,\frac{1}{2}\frac{dE}{du_0}\frac{1}{1+y} \tag{4.10}$$

$$= \frac{1}{2}\int_0^\infty dy\,E(u_0)(1+y)^{-2}. \tag{4.11}$$

In Eq. (4.11), as in Eq. (4.3b), u_0 is given by Eq. (4.4). Putting everything together, we obtain

$$\Phi_1(u,r) = r^{-1}E^{(2)}(u) + r^{-2}E^{(1)}(u) + \tfrac{1}{2}r^{-3}E(u) + \tfrac{1}{4}(2M)r^{-4}\int_0^\infty dy E(u_0)(1+y)^{-2} + \mathcal{O}[(2M/r)^2]. \tag{4.12}$$

For a given time dependence of the dipole moment $E(u)$, the solution (4.12) for Φ_{-1} must be consistent with the solution (4.5) for Φ_1. This is easily checked by applying Eqs. (2.16a) and (2.16b) to the solution (4.12). First,

$$\Phi_0 = \frac{\partial}{\partial r}[r\Phi_{-1}] = -r^2E^{(1)}(u) - r^{-3}E(u) - (2M)r^{-4}\int_0^\infty dy E(u_0)\,\frac{1+\frac{1}{2}y}{(1+y)^3} - \mathcal{O}[(2M/r)^2]. \tag{4.13}$$

Then

$$\Phi_1 = r^{-1}\frac{\partial}{\partial r}[r^2\Phi_0] \tag{4.14}$$

gives Eq. (4.5).

B. The gravitational quadrupole field

The gravitational quadrupole moment $G(u)$ is also very simply defined by Eq. (3.8):

$$G(u) \equiv \lim_{r\to\infty}[r^5\Psi_2(u,r)]. \tag{4.15}$$

The limit is again guaranteed to exist for retarded fields.

The analogs of Eqs. (4.2), (4.3a), (4.3b), and (4.5) for the spin-weight-two part of the gravitational quadrupole field are

$$h_1(y) = \frac{1 + 2y + 2y^2 + y^3 + \frac{1}{5}y^4}{(1+y)^5} \tag{4.16}$$

$$g_1(u,r) = G(u) - \int_{-\infty}^{0} du_0\,\frac{dG}{du_0}\,\frac{1 + 2y + 2y^2 + y^3 + \frac{1}{5}y^4}{(1+y)^5} \tag{4.17a}$$

$$= \int_0^\infty dy\,G(u_0)\,\frac{3 + 4y + 3y^2 + \frac{6}{5}y^3 + \frac{1}{5}y^4}{(1+y)^6} \tag{4.17b}$$

$$\Psi_2(u,r) = r^{-5}G(u) + \tfrac{5}{2}(2M)r^{-6}\int_0^\infty dy G(u_0)\,\frac{3 + 4y + 3y^2 + \frac{6}{5}y^3 + \frac{1}{5}y^4}{(1+y)^6} + \mathcal{O}[(2M/r)^2]. \tag{4.18}$$

The spin-weight-minus-two part of the gravitational quadrupole field shows how the effects of the background curvature can enter $\psi_I(u,r)$. [See discussion following Eq. (3.10)] Start with

$$f_1(u) = \tfrac{1}{6}G^{(4)}(u). \tag{4.19}$$

The successive integrations of Eq. (3.5) give

$$f_2(u) = \tfrac{1}{3}G^{(3)}(u), \tag{4.20}$$

$$f_3(u) = \tfrac{1}{2}G^{(2)}(u) + \tfrac{1}{8}(2M)G^{(3)}(u), \tag{4.21}$$

$$f_4(u) = \tfrac{1}{2}G^{(1)}(u) + \tfrac{1}{8}(2M)G^{(2)}(u), \tag{4.22}$$

$$f_5(u) = f_{l+s+1}(u) = \tfrac{1}{4}G(u) - \tfrac{1}{64}(2M)^2G^{(2)}(u). \tag{4.23}$$

Similar curvature terms appear in Ψ_2 when $l \geq 4$ and in Φ_1 when $l \geq 3$. Note that the $f_n(u)$, $n < 5$, cannot be expressed as a finite sum over derivatives of $f_5(u)$. For this reason, the quadrupole moment should not be defined as the coefficient of r^{-5} in $\Psi_{-2}(u,r)$; rather, Eq. (4.15)—which leads to Eq. (4.19)—is the better definition.

In $\Psi_{II}(u,r)$ for Ψ_{-2} the function $h_1(y)$ is again simple,

$$h_1(y) = (1+y)^{-1}. \tag{4.24}$$

This makes it feasible to go on and solve for $H_{1,2}(y)$, which appears in the solution for $g_2(u,r)$. The result of applying Eq. (3.31) is

$$H_{1,2}(y) = (1+y)^{-2}\{1 + \tfrac{7}{8}[y + ln(1+y) + \tfrac{11}{30} - \tfrac{1}{2}y^{-1} + \tfrac{3}{4}y^{-2} - \tfrac{4}{3}y^{-3} + \tfrac{7}{2}y^{-4} + 5y^{-5} - (5+6y)y^{-6}\ln(1+y)]\}. \tag{4.25}$$

We finally have for Ψ_{-2}:

$$\Psi_{-2}(u,r) = \tfrac{1}{6}r^{-1}G^{(4)} + \tfrac{1}{3}r^{-2}G^{(3)} + \tfrac{1}{2}r^{-3}[G^{(2)} + \tfrac{1}{4}(2M)G^{(3)}] + \tfrac{1}{2}r^{-4}[G^{(1)} + \tfrac{1}{4}(2M)G^{(2)}] + \tfrac{1}{4}r^{-5}[G - \tfrac{1}{16}(2M)^2G^{(2)}] + \tfrac{1}{6}r^{-5}(2M/r)\int_0^\infty dy[G(u_0) - \tfrac{1}{16}(2M)^2G^{(2)}(u_0)](1+y)^{-2} + \tfrac{1}{14}r^{-5}(2M/r)^2\int_0^\infty dy[G(u_0) - \tfrac{1}{16}(2M)^2G^{(2)}(u_0)][-H_{1,2}'(y)] + \mathcal{O}[(2M/r)^3]. \tag{4.26}$$

The coefficient of f_{l+s} vanishes in Eqs. (3.38)–(3.40) when $l = -p = s$, which means that the Newman–Penrose

constants appear in a simple way in Φ_{-1} and Ψ_{-2} (see Sec. 6).

5. PEELING PROPERTIES

There are three distinct types of infinities in an asymptotically flat space-time, corresponding to three possible choices of time coordinate. If $r \to \infty$ with the static time coordinate t in the Schwarzschild metric held constant, the limit is called spacelike infinity. The limit $r \to \infty$ at constant retarded time u is future null infinity, while the limit $r \to \infty$ at constant advanced time v [see Eq. (2.3)] is past null infinity. Penrose[9,10] has pioneered the study of the conformal structure of infinity. He transforms coordinates to bring $r = \infty$ in an asymptotically flat space-time to finite coordinate values and then removes the induced singularity in the metric by a conformal transformation. The original open, noncompact manifold $\tilde{M}$ is converted to a manifold M which contains future and past null infinity as regular null hypersurfaces ($\mathcal{I}^+$ and $\mathcal{I}^-$, respectively). Spacelike infinity is represented by a point I_0, which is generally a singular point of M. A spin-s zero-rest-mass field in the physical open, noncompact manifold $\tilde{M}$ can be described by a totally symmetric spinor $\tilde{\phi}_{A\cdots K}$, with $2s$ indices. The corresponding conformally transformed spinor

$$\phi_{A\cdots K} = \Omega^{-(s+1)}\tilde{\phi}_{A\cdots K} \tag{5.1}$$

satisfies the spin-s zero-rest-mass field equation in M. If $\phi_{A\cdots K}$ is continuous at $\mathcal{I}^-$ and $\mathcal{I}^+$, the field is called asymptotically regular. Penrose[9] shows that an asymptotically regular field has the following peeling behavior:

$$(r^{s+p+1}\psi_p) \tag{5.2}$$

has a limit at future null infinity, and

$$(r^{s-p+1}\psi_p) \tag{5.3}$$

has a limit of past null infinity, where ψ_p is the spin-weight-p part of the field tensor. For an electromagnetic field $\psi_p = \Phi_p$ and for a gravitational field $\psi_p = \Psi_p$.

Penrose then prove asymptotic regularity of the gravitational field on the assumption that the geometry of the spacetime is sufficiently smooth at null infinity, specifically that the manifold M is C^4-differentiable with a C^3 metric and that the conformal factor Ω is C^3 on M. (This step is close to a tautology, since the geometry is the gravitational field.) With the same regularity of the geometry an electromagnetic field is not required to be asymptotically regular—expressions (5.2) and (5.3) need only be bounded at future and past null infinity, respectively.

On the other hand, Couch and Torrence[19] have shown that a very much weaker type of peeling behavior, in which $r^{s+1}\psi_p$ need not be bounded for $p \geq 0$ at future null infinity and for $p \leq 0$ at past null infinity, is consistent with asymptotic flatness.

How much regularity at future or past null infinity can be expected purely on the basis of a certain set of assumptions about the physical nature of the source? We begin an exploration of this question using our general retarded test field solutions.

The only restriction we impose in deriving the solutions are the absence of incoming radiation at past null infinity, the boundedness of the retarded multipole moment at all times to the past, and the validity of the test field approximation.

The first two restrictions correspond to the physical condition that the sources of the test fields be bounded within a compact region for all times to the past. If the source is contained within a radius $R > 2M$, a rough estimate of the maximum possible gravitational multipole moment is MR^l. Charge separation for electromagnetic (or possibly scalar) sources increases the limit to the order of R^{l+1} (charges measured in Gaussian units), at which point the test field approximation breaks down.

The validity of the test field approximation requires that the nonlinear contributions of the test fields in the exact Einstein equations do not significantly modify the background Schwarzschild metric. The first $(l + s)$ derivatives of $A(u)$ must be bounded, so that the $\Psi_p(u, r)$ will be bounded. Also, the energy radiated over all times to the past must be small compared with the gravitational mass M. An explicit positive-definite[20] expression for the energy radiated from a particular 2^l-pole is

$$\frac{1}{2\pi}\left(2^l\frac{(l-s)!}{(2l)!}\right)^2\int_{-\infty}^{u}|A^{(l+1)}(u_0)^2|\,du_0 \ll M(u=-\infty). \tag{5.4}$$

We claim that no further constraints on $A(u)$ are physically necessary.

At future null infinity our general retarded solutions for scalar, electromagnetic, and gravitational fields are all asymptotically regular in the Penrose sense. Moreover, the coefficients of $r^{-(s+p+1)}$ in the ψ_p are related to each other and (by definition) to $A(u)$ in the same way as in a flat background:

$$\lim_{r\to\infty} r^{s+p+1}\psi_p(u,r) = (-1)^{s-p}\left(2^{s-p}\frac{(l+p)!}{(l-p)!}\frac{(l-s)!}{(l+s)!}\right)^{1/2} \times \frac{2^{l-s}(l+s)!}{(2l)!}A^{(l-p)}(u). \tag{5.5}$$

In the limit $r \to \infty$ Eqs. (2.16c) and (2.16d) and Eqs. (2.18e)–(2.18h) acting on the general retarded solution reduce to the flat-space equations. Note that Eq. (3.9) has been established by our proof of Eqs. (3.46) and (3.47).

To find the peeling behavior at past null infinity, let

$$u = v - 2r \tag{5.6}$$

and take the limit $r \to \infty$ with v constant. While v is not the exact advanced time in the Schwarzschild background, it becomes exact in the limit. The backscatter part of the field $\Psi_{II}(u, r)$ is of order $(2M/r)$ compared with $\psi_I(u, r)$ and does not contribute to $\Psi_p(u, r)$ in the limit. Also, we can neglect terms of order $(2M/r)$ in the coefficient of a given derivative of $A(u)$ in $\psi_I(u, r)$. Without assuming anything about relative magnitudes of the derivatives at past null infinity, we obtain

$$r^{s-p+1}\psi_p \simeq 2^{(p-s)/2}\lim_{r\to\infty}\sum_{k=0}^{l-p}\frac{(2l-k)!}{(2l)!}\frac{(l-p)!}{k!(l-p-k)!}2^k \times r^{-l-p+k}A^{(k)}(v-2r). \tag{5.7}$$

The content of Eq. (5.7) is that the general retarded solu-

tion in the Schwarzschild background approaches a retarded flat-space solution at past null infinity as well as at future null infinity.

Since the $A^{(k)}$ are bounded, $r^{s-p+1}\Psi_p$ approaches a limit of zero at past null infinity for $p > 0$. However, $r^{s+1}\psi_0$ need only be bounded. The condition (5.4) on the radiated energy allows $r^{s-p+1}\psi_p$ to be unbounded at past null infinity for $p < 0$. For instance,

$$A(u) \sim \sin[b(u/u_1)^{1/3}] \tag{5.8}$$

gives a finite energy radiated for all $l \geq s \geq 0$, but

$$r^{2s+1}\psi_{-s} \sim (r^{2/3})^{(2s-l)} \to \infty$$

at past null infinity for $0 < s \leq l < 2s$.

A mathematical condition that the field be asymptotically regular at past null infinity is equivalent to a restriction on the time dependence of the multipole moment in the distant past which is much stronger than Eq. (5.4). For example, if $l = s$ the restriction is that

$$\lim_{u\to-\infty} u^k A^{(k)}(u) \tag{5.9}$$

exist for $0 \leq k \leq 2s$. The multipole moment must be asymptotically static in the infinite past—to any given accuracy it must be static for an infinite time. Such a restriction is not required by any physical regularity condition. We conclude that the geometrical regularity conditions assumed by Penrose are not physically necessary.

A physically more appropriate approach to peeling theorems is a direct argument from the general retarded solution to the field equations. Sachs[21] and Goldberg and Kerr[22] have proved such theorems for linearized gravitational fields and for electromagnetic fields in flat space. Our results support the proposition that backscatter cannot be strong enough in an asymptotically flat space-time to destroy the asymptotic regularity of the field at future null infinity.

6. THE NEWMAN-PENROSE QUANTITIES

The standard prescription for calculating the NPQ's associated with a spin-s field[4,5,23] is as follows. Consider the Newman–Penrose field scalar with spin weight $p = s$. Extract the lowest radiatable multipole $l = s$, by performing an angular integration. Denote the resulting function of u and r by $\psi_s(u, r)$. For example, in the gravitational case

$$\psi_s(u, r) = 2\pi \int \sin\theta \, d\theta \, d\varphi \; \Psi_2(u, r, \theta, \varphi) \; {}_2Y^2{}_m(\theta, \varphi). \tag{6.1}$$

The $\psi_s(u, r)$ are really $2s + 1$ complex functions, corresponding to the $2s + 1$ possible values of the axial eigenvalue m. Now let

$$P(u, r) = r^{2s+1}\psi_s(u, r) \tag{6.2}$$

and

$$Q(u, r) = \frac{\partial}{\partial(1/r)} [r^{2s+1}\psi_s(u, r)]. \tag{6.3}$$

The lowest radiatable multipole moment is

$$A_s(u) = \lim_{r\to\infty} P(u, r), \tag{6.4}$$

and the NPQ is

$$\text{NPQ} = \lim_{r\to\infty} Q(u, r). \tag{6.5}$$

Both limits are at future null infinity. There are $2(2s + 1)$ NPQ's associated with the spin-s field, corresponding to the real and imaginary parts of the $(2s + 1)$ functions ψ_s.

Newman and Penrose[4,5] *assume* that

$$\psi_s(u, r) = \frac{A_s(u)}{r^{2s+1}} + \frac{\text{NPQ}}{r^{2s+2}} + \mathcal{O}(r^{-(2s+3)}), \tag{6.6}$$

so that both limits (6.4) and (6.5) exist, and then show that the gravitational NPQ's are conserved (independent of u) if the field satisfies the vacuum Einstein equations near null infinity. Exton, Newman, and Penrose[23] prove that the electromagnetic and gravitational NPQ's are conserved by the vacuum Einstein–Maxwell equations in asymptotically flat space-times.

Subsequent papers[24] have examined various mathematical properties of the NPQ's, but have not shed much light on their physical meaning. We try to fill this gap by asking the following equestions in the context of test fields in the Schwarzschild background: (i) Are the NPQ's measurable in any physically meaningful sense? (ii) Under what conditions does the limit (6.5), and therefor the NPQ, exist as a formal mathematical property of the test field? (iii) If the limit does exist, what is the physical interpretation of the value of the NPQ?

The answers, in brief, are: (i) The NPQ's are not measurable and, therefore, have no direct physical significance. (ii) The NPQ's do not exist for all physically nonsingular retarded solutions to the field equations. (iii) When the NPQ does exist, its value is proportional to a certain average of the lowest radiatable multipole moment in the infinite past.

We begin with the general retarded solution for $\psi_s(u, r)$ as obtained in Sec. 3:

$$\psi_s(u, r) = A_s(u) r^{-(2s+1)} + r^{-(2s+1)} \sum_{k=1}^{\infty} a_k g_k(u, r) \left(\frac{2M}{r}\right)^k. \tag{6.7}$$

Since the $g_k(u, r)$ are uniformly bounded if $A_s(u)$ is uniformly bounded to the past

$$P(u, r) = A_s(u) + \mathcal{O}(2M/r). \tag{6.8}$$

If the general retarded solution (6.7) is substituted into Eq. (6.3) for $Q(u, r)$, the result is

$$Q(u, r) = (2s + 1)M\left[\left(g_1 - r\frac{\partial g_1}{\partial r}\right) + \mathcal{O}(2M/r)\right]. \tag{6.9}$$

For general s [see Eqs. (5.4a), (5.4b) and (4.17a), (4.17b)],

$$g_1(u, r) = A_s(u) - \int_{-\infty}^{u} du_0 \frac{dA_s}{du_0} h_1(y), \tag{6.10}$$

with

$$h_1(y) = \frac{1}{(2s + 1)y}\left(1 - \frac{1}{(1 + y)^{2s+1}}\right) \tag{6.11}$$

and

$$y = (u - u_0)/2r. \tag{6.12}$$

Note that $f_{l+s}(u_0)$ appears in Eq. (3.38) for $g_1(u,r)$, vanishes identically when $l=p=s$ as a consequence of the peeling conditions at future null infinity. Equations (6.9)–(6.12) combine to give at $r \gg 2M$.

$$Q(u,r) = (2s+1)M\left(A_s(u) - \int_{-\infty}^{u} du_0 \frac{dA_s}{du_0} \frac{1}{(1+y)^{2s+2}}\right) \quad (6.13a)$$

or

$$Q(u,r) = (2s+1)(2s+2)M \times \int_0^{\infty} dy A_s(u-2ry)(1+y)^{-(2s+3)}. \quad (6.13b)$$

The integrals in Eqs. (6.10) and (6.13a) come from a superposition of incoming waves (the backscatter) generated by previous changes in the multipole moment.

Both $P(u,r)$ and $Q(u,r)$ are directly measurable by a network of observers covering a finite region of space-time surrounding the source region at $r \gg 2M$. The observers measure the field tensor as a function of position. They must project the tensor at each point on an appropriate null tetrad and then perform an angular integration to obtain $\psi_s(u,r)$ over the span of radius r and over the span of retarded time u covered by the observers. The choice of null tetrads is not unique in general; but the spherical symmetry allows a unique choice for the background Schwarzschild metric. To first order in the gravitational field perturbations only the spin-weight $p=0,\pm 1$ parts of the perturbed Weyl tensor are affected by the uncertainty in the tetrad induced by the first-order deviations from spherical symmetry. In all cases, then, $\psi_s(u,r)$ is unambiguous to first order in the test field. Given measurements of $\psi_s(u,r)$ for $l=s$ to some finite accuracy over a range of r and u the observers can extract the values of $P(u,r)$ and $Q(u,r)$ to a corresponding accuracy.

Equation (6.8) says that at values of $r \gg 2M$ the observers will find that $P(u,r)$ is independent of r along an outgoing radial null geodesic, and assures that its variation with u can safely be interpreted as the variation of the multipole moment $A_s(u)$, as defined at future null infinity, with u. In this sense, the lowest radiatable multipole moment is measurable.

The Newman–Penrose constant is the limit (6.5) of $Q(u,r)$ at future null infinity. However, Eq. (6.13) gives no assurance that the limit exists, let alone that the value of the limit can be extracted from measurements of $Q(u,r)$ over a finite region of space–time.

Consider as an example $A_s(u)=A_1$ for all $u<u_1$ and $A_s(u)=A_2$ for all $u>u_1$. Equation (6.13) gives

$$Q(u,r) = \begin{cases} (2s+1)MA_1, & u<u_1 \quad (6.14a) \\ (2s+1)M\{A_2 - (A_2-A_1)[1+(u-u_1)/2r]^{-(2s+2)}\}, & u>u_1. \quad (6.14b) \end{cases}$$

The NPQ at all u is the intial static value of Q given by Eq. (6.14a). However, at any fixed, finite value of r, Q goes smoothly toward a new asymptotically static value appropriate to the new value of the multipole moment in the limit $u-u_1 \gg 2r$.

Measurements of finite accuracy will not detect any deviation from the new static value of Q if the change in the multipole moment occurred at a time u_1 sufficiently far in the past, such that $u_1 \ll u-2r$. It is not physically reasonable to require that measurements be made infinitely far in the past ($u \to -\infty$) or at infinitely large radii ($r \to \infty$ at finite u), or that they be infinitely accurate. An apparently constant Q need not be constant all the way out to future null infinity, so the value of Q at future null infinity, the NPQ, is not measurable in a physically realistic sense.

As Eq. (6.13b) makes explicit, the value of Q at given u,r is proportional to a weighted time average of the multipole moment over the entire past history. The weighting function $(2s+2)(1+y)^{-(2s+3)}$ cuts off at $y \sim 1$ or $u_0 \sim u-2r$, so the average is effectively over a time $\Delta u = u-u_0 \sim 2r$ previous to the retarded time at which Q is being evaluated. In the limit $r \to \infty$, the interval Δu expands to cover the entire past history uniformly. The NPQ is a uniform average of $A_s(u_0)$ over the entire past, if the average exists. We shall see below that such an average may not exist. Since any finite range of u_0 makes a negligible contribution to the average over the entire past, the value of the NPQ, if it exists, cannot be extracted from measurements of the field at finite u and r, which are only sensitive to $A_s(u_0)$ over a finite time to the past. In physical terms, the presence of an "average value" in the field is due to the local superposition of backscatter from the outgoing radiation of all previous retarded times.

We define a measurable Newman–Penrose quantity (MNPQ) to be the value of Q in a region of space–time where Q, to the finite accuracy of the measurements, is a constant independent of u and r. For an MNPQ to exist, the average value of $A_s(u_0)$ must have been constant over times $\Delta u \gg 2r$ to the past. Either $A_s(u_0)$ itself was constant or substantial net deviations of $A_s(u_0)$ from the average value only lasted for a time $\delta u \ll 2r$.

The above definition of an MNPQ differs from our previous[8] identification of the MNPQ as the coefficient of $r^{-(2s+2)}$ in an asymptotic expansion of the field in powers of r^{-1}. The old MNPQ was not defined very precisely mathematically, since an asymptotic expansion of the field in powers of r^{-1} is not always possible. Once established, the old MNPQ does persist until a time $u-u_1 = 2r$ after the field becomes dynamic at $u=u_1$. The old MNPQ fails at the $\frac{1}{3}$-speed of-light cone $u-u_1=2r$, because on this cone the maximum value of y which appears in the integral (6.13a), with $A_s(u_0)$ static for $u_0<u_1$, is $y=1$. At $y=1$ an expansion of $(1+y)^{-(2s+2)}$ in powers of r^{-1} diverges.

The $\frac{1}{3}$-speed-of-light cone has no special meaning for the new MNPQ's. These are associated with the quantity $Q(u,r)$, which is always well defined and varies continuously when the multipole moment changes. For a given measurement accuracy ϵ the new MNPQ persists until $(u-u_1)/2r = \mathcal{O}(\epsilon)$ after the multipole moment changes [see Eq. (6.14b), for example].

Goldberg[25] has still another definition of MNPQ's which relates them to an artificially constructed "conserved flux." Goldberg's MNPQ's also change continuously when the multipole moment changes. Their values at finite u and r are no more closely related to the values of the NPQ's, if they exist, than our MNPQ's.

There is a limit on how rapidly $Q(u,r)$ can vary. Note that since $A_s(u)$ is bounded

$$\frac{\partial Q}{\partial u} = (2s+1)(2s+2)M\int_0^\infty dy A'_s(u-2ry)(1+y)^{-(2s+3)}$$
$$= \mathcal{O}(r^{-1}Q). \tag{6.15}$$

We now consider some examples of when the limit (6.5) defining the NPQ at future null infinity does and does not exist. First, a sufficient condition for the NPQ to exist is that the limit

$$A_s(-\infty) = \lim_{u\to-\infty} A_s(u) \tag{6.16}$$

exist. Then the average value of $A_s(u)$ in the limit $r\to\infty$ in Eq. (6.13b) is just $A_s(-\infty)$, and

$$\text{NPQ} = (2s+1)MA_s(-\infty). \tag{6.17}$$

A multipole moment which has the limit (6.16) is, by definition, asymptotically static in the infinite past.

A retarded solution to the test field equations which satisfies the strong Penrose peeling condition at past null infinity is asymptotically static in the infinite past and, therefore, possesses NPQ's. However, we have seen in Sec. 5 that there are no physical restrictions which require the field to satisfy the Penrose peeling condition at past null infinity or to be asymptotically static in the infinite past.

As an example of a solution which is not asymptotically static in the infinite past, but still possesses NPQ's, consider

$$A_s(u) = \sin bu \tag{6.18}$$

for a scalar field ($s = 0$). The integrals in Eqs. (6.13a) or (6.13b) for $Q(u,r)$ involve sine and cosine integrals. Let

$$\text{Si}(x) = \tfrac{1}{2}\pi - f(x)\cos(x) - g(x)\sin(x) \tag{6.19}$$

and

$$\text{Ci}(x) = \gamma + \ln(x) + f(x)\sin(x) - g(x)\cos(x). \tag{6.20}$$

The functions $f(x)$ and $g(x)$ have asymptotic expansions

$$f(x) \simeq \frac{1}{x}\left(1 - \frac{2!}{x^2} + \frac{4!}{x^4} - \cdots\right) \tag{6.21}$$

and

$$g(x) \simeq \frac{1}{x^2}\left(1 - \frac{3!}{x^2} + \frac{5!}{x^4} - \cdots\right) \tag{6.22}$$

when $x \gg 1$. The result for Q is

$$Q(u,r) = [1 - (2rb)^2 g(2rb)]\sin(u)$$
$$- (2rb)[1 - (2rb)f(2rb)]\cos(u). \tag{6.23}$$

In the limit $r\to\infty$

$$Q(u,r) \simeq -\frac{\cos(u)}{br} + \mathcal{O}\left(\frac{1}{(br)^2}\right) \to 0, \tag{6.24}$$

so the NPQ exists and equals zero. In other words, the average of the multipole moment can approach a limit as $u_0\to\infty$, even though the multipole moment itself does not.

If the multipole moment varies on time scales which are the order of $u - u_0$ in the limit $u_0\to-\infty$, the NPQ does not exist. For instance, if

$$A_s(u) = \sin(c\sinh^{-1}bu) \tag{6.25}$$

and $c \ll 1$, $A_s(u_0)$ is approximately constant over most of the range $u > u_0 > u - 2r$ when $2rb \gg 1$. Thus

$$Q(u,r) \simeq (2s+1)M\{\sin[b\sinh^{-1}(-2rb)] + \mathcal{O}(c)\} \tag{6.26}$$

and oscillates indefinitely in the limit $2rb\to\infty$.

We conclude that the NPQ exists if and only if the average $A_s(u_0)$, over a time $\Delta u = u - u_0$, approaches a limit as $u_0\to-\infty$. Either A_s is asymptotically static in the infinite past, or the time variation of A_s is *entirely* on time scales infinitesimally short compared with $u - u_0$ in the limit $u_0\to-\infty$. Both of these conditions are rather special, and the NPQ's will not exist for the generic retarded test field solution.

The conservation of the NPQ's, when they exist, has no predictive powers; it is a reflection of the fact that an average of $A_s(u_0)$ over an infinite time is not affected by time variations over any finite time span. Therefore, the existence of the NPQ at any finite retarded time automatically implies the existence of the NPQ with the same value at any other finite retarded time. The value of the field at finite u and r depends, to any finite accuracy, only on the multipole moment over a finite range of retarded time to the past and is, thus, in principle completely independent of the value of the NPQ.

We are left with NPQ's which, when they exist, have only a formal mathematical significance. For a spin-s zero-rest-mass test field in any static, spherically symmetric, asymptotically flat background (in any metric theory of gravity), this mathematical significance has a simple origin. As long as the metric coefficients for the static background are analytic in $1/r$ in some neighborhood of $1/r = 0$, it is possible to expand the general retarded solution for the spin-weight $p = s$, $l = s$ part of the field as

$$\psi_s = A_s(u)r^{-(2s+1)} + r^{-(2s+1)}\sum_{n=1}^{\infty} a_k\left(\frac{2M}{r}\right)^k g_k(u,r). \tag{6.27}$$

Here M is just a parameter indicating the scale of radius on which the deviations from flat space become large; it need not have an interpretation as a gravitational mass. The a_k are coefficients chosen so that the static solution for ψ_s has $g_k = A_s = \text{const}$. The relativistic equations may couple $g_k(u,r)$ to g_{k-2}, g_{k-3}, etc. as well as the g_{k-1}. However, $g_1(u,r)$ can only couple to $f_{2s+1}(u)$, as before, since the $f_n(u)$ with $n < 2s+1$ are identically zero. The function $H_{1,1}(y)$ in this context, also, is the homogeneous similarity solution to the flat-space spin-s field equations. In general, then,

$$Q(u,r) = 2(2s+2)Ma_1$$
$$\times\left[\int_0^\infty dy A_s(u_0 = u - 2ry)(1+y)^{-(2s+3)} + O\left(\frac{2M}{r}\right)\right]. \tag{6.28}$$

The existence and value of the NPQ is related to an average over the lowest radiatable multipole moment in the infinite past in essentially the same way as before. The only possible difference is the value of the coefficient a_1 in the relativistic static solution. Thus, the conservation of the NPQ's (when they exist) depends only on asymptotic flatness; it is independent of any special properties of the curvature correction to the field equations.

An apparent special property of the Einstein–Maxwell equations for test fields in the Schwarzschild background is the appearance of the NPQ's in the spin weight $p = -s$ part of the lowest radiatable multipole. In Eq. (3.38), for example, f_{l+s} is not identically zero when $l = s = -p$; but it has a vanishing coefficient and does not contribute to $g_k(u, r)$. However, even this is a result only of asymptotic flatness, plus consistency of the equations for the different spin-weight parts of the field. The leading backscatter at future null infinity can always be interpreted as an incoming wave in flat space, so the coefficients of the incoming wave in the different spin-weight parts of the field must be related by the flat-space equations. If the backscatter compensates for the changes in multipole moment in $g_1(u, r)$ of the spin weight $p = s$ part of the field, as it does when $l = s$, it must do so in $g_1(u, r)$ of the other spin-weight parts of the field, as well.

The generalization of our results to asymptotically flat solutions of the full nonlinear Einstein and Einstein–Maxwell field equations is not quite as straightforward. For instance, the NPQ's of the gravitational field have a different form when a dynamic electromagnetic field is present.[23] It does seem safe to conclude that if the lowest radiatable multipole moments of the electromagnetic and gravitational fields do not have the asymptotic behavior in the infinite past necessary for the existence of the test field NPQ's, the NPQ's of the respective fields will not exist in the full nonlinear theory either. The conservation of the NPQ's, when they exist, is probably as trivial a consequence of asymptotic flatness as it is for test fields.

It may be possible to obtain general retarded solutions to the exact field equations at large r similar to our test-field equations and check the validity of these conjectures directly. Care must be taken not to assume more regularity at future or past null infinity than is physically justified.

Our approach to the physical interpretation of the NPQ's has concentrated on their existence and measurability. Glass and Goldberg[24] have interpreted the conservation of the NPQ's in terms of invariant transformations and an artificially constructed differential conservation law. They assume that the NPQ's exist and then show that their conservation is related to a superposition principle for ingoing and outgoing waves valid asymptotically in the lowest radiatable multipole in asymptotically flat space times. We have not found any physical content to the "conserved flux" they define.

7. SUMMARY AND CONCLUSION

Using the general retarded solution of our master equation for the radiative parts of test fields in the Schwarzschild background, we have examined the nature of the fields' Newman–Penrose quantities and peeling properties.

The explicit retarded test-field solution shows that the NPQ's are a certain average of the lowest radiatable multipole moment over the infinite past and do not exist unless the average exists. Even when they do exist, the NPQ's are not measurable and, therefore, have no direct physical significance.

Of course, the asymptotically flat boundary condition is an abstraction which ignores the existence of other matter in the universe. In practice, for a star in the galaxy, one can ignore the other stars out to a radius of about one light year at most. At such a radius the multipole moments are very well established, since if $M = 1M_\odot$, $2M/r$ is the order of 10^{-12}. However, any net change in the lowest radiatable multipole moment on a year's time scale or longer makes it impossible to talk about a conserved NPQ.

The peeling properties of test fields in the Schwarzschild background are identical to the peeling properties in a flat background. The mathematical regularity assumptions of the Penrose peeling theorem are justified at future null infinity, but not at past null infinity.

The general retarded solution can also be used to study the detailed development and decay of the backscatter and wavetails for all radiatable multipoles. The wavetail is outgoing radiation at future null infinity at retarded times after the source has become static. This material, however, will be presented in a subsequent paper.

ACKNOWLEDGMENT

We thank Kip Thorne and Richard Price for introducing us to this problem and for many helpful discussions. This work was begun while one of us (J.M.B.) was visiting Caltech as a Senior Research Associate in the spring of 1971.

*Supported in part by the National Science Foundation (GP-15267) at the University of Washington and (GP-27304, GP-28027) at the California Institute of Technology.

†Present address: Yale University, New Haven, Connecticut.

‡Fannie and John Hertz Foundation Fellow.

[1]W. Kundt and E. T. Newman, J. Math. Phys. 9, 2193 (1968).
[2]R. G. McLenaghan, Proc. Camb. Philos. Soc. **65**, 139 (1969).
[3]W. B. Bonner and M. A. Rotenberg, Proc. R. Soc. A **289**, 247 (1965).
[4]E. T. Newman and R. Penrose, Phys. Rev. Lett. **15**, 231 (1965).
[5]E. T. Newman and R. Penrose, Proc. R. Soc. A **305**, 175 (1968).
[6]R. H. Price, Phys. Rev. D 5, 2419 (1972).
[7]R. H. Price, Phys. Rev. D 5, 2439 (1972).
[8]W. H. Press and J. M. Bardeen, Phys. Rev. Lett. **27**, 1303 (1971).
[9]R. Penrose, Proc. R. Soc. A **284**, 159 (1965).
[10]R. Penrose, in *Relativity, groups, and topology*, edited by C. DeWitt and B. DeWitt (Gordon and Breach, New York, 1964).
[11]E. T. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).
[12]E. T. Newman and R. Penrose, J. Math. Phys. **7**, 863 (1966).
[13]A. I. Janis and E. T. Newman, J. Math. Phys. **6**, 902 (1965).
[14]T. Regge and J. A. Wheeler, Phys. Rev. **108**, 1063 (1957).
[15]F. J. Zerilli, Phys. Rev. D **2**, 2141 (1970).
[16]K. S. Thorne, in *Magic without magic: John Archibald Wheeler*, edited by J. Klauder (Freeman, San Francisco, 1972).
[17]W. E. Couch and R. J. Torrence, J. Math. Phys. **9**, 484 (1968).
[18]W. E. Couch and W. H. Halliday, J. Math. Phys. **12**, 2170 (1971).
[19]W. E. Couch and R. J. Torrence, J. Math. Phys. **13**, 69 (1972).
[20]H. Bondi, M. van der Burg, and A. Metzner, Proc. R. Soc. A **269**, 21 (1962).
[21]R. K. Sachs, Proc. R. Soc. A **264**, 309 (1961); Proc. R. Soc. A **270**, 103 (1962); Phys. Rev. **128**, 2851 (1962).
[22]J. N. Goldberg and R. P. Kerr, J. Math. Phys. **5**, 172 (1964).
[23]A. R. Exton, E. T. Newman, and R. Penrose, J. Math. Phys. **10**, 1566 (1969).
[24]For references see E. N. Glass and J. N. Goldberg, J. Math. Phys. **11**, 3400 (1970).
[25]J. N. Goldberg, Phys. Rev. Lett. **28**, 1400 (1972).

Physical applications of multiplicative stochastic processes

Ronald Forrest Fox
Georgia Institute of Technology, Atlanta, Georgia 30332
(Received 24 July 1972)

The theory of multiplicative stochastic processes has been shown to lead to a density matrix description of nonequilibrium quantum mechanical phenomena. In the present paper a detailed treatment of the approach to the uniform, microcanonical, and canonical equilibrium density matrices is presented. The canonical equilibrium density matrix is approached by the density matrix which represents a subsystem in contact with a constant temperature heat reservoir.

INTRODUCTION

In a recent paper the theory of multiplicative stochastic processes was explained, and it was shown how such a theory potentially leads to a description of nonequilibrium phenomena.[1] In the present paper the density matrix formulation of nonequilibrium quantum mechanical phenomena will be presented with a detailed account of the approach to the uniform, microcanonical, and canonical equilibrium density matrices. The circumstances in which the canonical equilibrium density is approached are of particular interest since they correspond to a subsystem which is in contact with a constant temperature heat reservoir.

RECAPITULATION

The Schrödinger equation for nonrelativistic quantum mechanics may be written in matrix form as[2]

$$i\frac{d}{dt}C_\alpha(t) = \sum_{\alpha'} M_{\alpha\alpha'}C_{\alpha'}(t), \tag{1}$$

where $M_{\alpha\alpha'} = M^*_{\alpha'\alpha}$, which is the condition of Hermiticity, and where $\sum_\alpha C^*_\alpha(t)C_\alpha(t) = 1$ for all t, which is the condition of conservation of total probability. The Hermiticity of $M_{\alpha\alpha'}$ in (1) is a necessary and sufficient condition for the conservation of total probability. Suppose that a fluctuating contribution to the Hamiltonian is considered. Then (1) becomes

$$i\frac{d}{dt}C_\alpha(t) = \sum_{\alpha'} M_{\alpha\alpha'}C_{\alpha'}(t) + \sum_{\alpha'}\tilde{M}_{\alpha\alpha'}(t)C_{\alpha'}(t), \tag{2}$$

where $\tilde{M}_{\alpha\alpha'}(t) = \tilde{M}^*_{\alpha'\alpha}(t)$, and the following properties hold for the averaged moments of $\tilde{M}_{\alpha\alpha'}(t)$[1]:

$$\langle\tilde{M}_{\alpha\alpha'}(t)\rangle = 0, \tag{3}$$

$$\langle\tilde{M}_{\alpha\alpha'}(t)\tilde{M}_{\beta\beta'}(s)\rangle = 2Q_{\alpha\alpha'\beta\beta'}\delta(t-s), \tag{4}$$

$$\langle\tilde{M}_{\mu_1\nu_1}(t_1)\cdots\tilde{M}_{\mu_{2n-1}\nu_{2n-1}}(t_{2n-1})\rangle = 0 \quad \text{for } n = 1, 2, \ldots, \tag{5}$$

$$\begin{aligned}\langle\tilde{M}_{\mu_1\nu_1}(t_1)\cdots\tilde{M}_{\mu_{2n}\nu_{2n}}(t_{2n})\rangle &= \frac{1}{2^n n!}\sum_{p\in S_{2n}}\prod_{j=1}^{n}\langle\tilde{M}_{\mu_{p(2j-1)}\nu_{p(2j-1)}}(t_{p(2j-1)}) \\ &\quad\times\tilde{M}_{\mu_{p(2j)}\nu_{p(2j)}}(t_{p(2j)})\rangle \\ &= \frac{1}{2^n n!}\sum_{p\in S_{2n}}\prod_{j=1}^{n}2Q_{\mu_{p(2j-1)}\nu_{p(2j-1)}\mu_{p(2j)}\nu_{p(2j)}} \\ &\quad\times\delta(t_{p(2j-1)} - t_{p(2j)}),\end{aligned} \tag{6}$$

where S_{2n} is the symmetric group of order $(2n)!$ The properties given by (3), (4), (5), and (6) are those appropriate for a purely random, Gaussian, stochastic, matrix process. Such stochastic processes are the only type of stochastic process to be considered in this paper.

A density matrix representation for the Schrödinger equation is obtained in terms of the density matrix $\rho_{\alpha\beta}(t)$, which is defined by[3]

$$\rho_{\alpha\beta}(t) \equiv C^*_\alpha(t)C_\beta(t). \tag{7}$$

With the definitions

$$L_{\alpha\beta\alpha'\beta'} \equiv \delta_{\alpha\alpha'}M_{\beta\beta'} - \delta_{\beta\beta'}M^*_{\alpha\alpha'} \tag{8}$$

$$\tilde{L}_{\alpha\beta\alpha'\beta'} \equiv \delta_{\alpha\alpha'}\tilde{M}_{\beta\beta'}(t) - \delta_{\beta\beta'}\tilde{M}^*_{\alpha\alpha'}(t), \tag{9}$$

Eq. (2) may be used to directly verify

$$i\frac{d}{dt}\rho_{\alpha\beta}(t) = \sum_{\alpha'}\sum_{\beta'}[L_{\alpha\beta\alpha'\beta'} + \tilde{L}_{\alpha\beta\alpha'\beta'}(t)]\rho_{\alpha'\beta'}(t). \tag{10}$$

This is the fluctuating density matrix equation. Using (3)–(6) to average over the stochastic contribution to (10), an equation for the averaged density matrix $\langle\rho_{\alpha\beta}(t)\rangle$ may be obtained, although only after considerable computation[1]

$$\frac{d}{dt}\langle\rho_{\alpha\beta}(t)\rangle = -i\sum_{\alpha'}\sum_{\beta'}L_{\alpha\beta\alpha'\beta'}\langle\rho_{\alpha'\beta'}(t)\rangle - \sum_{\alpha'}\sum_{\beta'}R_{\alpha\beta\alpha'\beta'}\langle\rho_{\alpha'\beta'}(t)\rangle. \tag{11}$$

The "matrix" $R_{\alpha\beta\alpha'\beta'}$ which appears in (11) is defined by[1]

$$R_{\alpha\beta\alpha'\beta'} \equiv \delta_{\alpha\alpha'}\sum_\theta Q_{\beta\theta\theta\beta'} + \delta_{\beta\beta'}\sum_\theta Q_{\theta\alpha\alpha'\theta} - Q_{\beta\beta'\alpha'\alpha} - Q_{\alpha'\alpha\beta\beta'}, \tag{12}$$

where $Q_{\alpha\beta\mu\nu}$ is the "matrix" which appears in (4) and (6).

From (12) it also follows that for arbitrary complex matrices $X^1_{\alpha\beta}$,

$$\sum_\alpha\sum_\beta\sum_{\alpha'}\sum_{\beta'}X^*_{\alpha\beta}R_{\alpha\beta\alpha'\beta'}X_{\alpha'\beta'} \geq 0 \tag{13}$$

and for arbitrary μ and ν[1],

$$\sum_\alpha R_{\mu\nu\alpha\alpha} = 0. \tag{14}$$

APPROACH TO THE UNIFORM EQUILIBRIUM DENSITY MATRIX

Equations (13) and (14) lead to a proof that (11) describes the approach of $\langle\rho_{\alpha\beta}(t)\rangle$ to an equilibrium density matrix which is uniform[1]:

$$\langle \rho_{\alpha\beta}(t) \rangle \xrightarrow[t\to\infty]{} \rho_0 \delta_{\alpha\beta}. \tag{15}$$

If there are N eigenstates involved, then conservation of total probability implies $\rho_0 = 1/N$. The equilibrium density matrix given by (15) has been discussed by Tolman.[4] If the different eigenstates also have different energy eigenvalues, then the uniform equilibrium density matrix nevertheless gives equal weight to each eigenstate independently of its energy eigenvalue. Physically, the uniform equilibrium density matrix has found almost no applications, and only possesses theoretical interest. The reason that the uniform equilibrium density matrix asymptotically occurs is that in (2) no restrictions with respect to eigenstate–eigenstate couplings have been imposed upon $\tilde{M}_{\alpha\alpha'}(t)$. Unrestricted, $\tilde{M}_{\alpha\alpha'}(t)$ may couple any two eigenstates, even if their energy eigenvalues are greatly different. It is this feature of $\tilde{M}_{\alpha\alpha'}(t)$ which leads to the uniform equilibrium density matrix.

APPROACH TO THE MICROCANONICAL EQUILIBRIUM DENSITY MATRIX

Consider Eq. (2). Because both $M_{\alpha\alpha'}$ and $\tilde{M}_{\alpha\alpha'}(t)$ are Hermitian matrices, there exists a unitary transformation which diagonalizes $M_{\alpha\alpha'}$ while transforming $\tilde{M}_{\alpha\alpha'}(t)$ into another Hermitian matrix $\tilde{M}^T_{\alpha\alpha'}(t)$. This transformation may be schematized by

$$M_{\alpha\alpha'} \to d_\alpha \delta_{\alpha\alpha'}, \tag{16}$$

$$\tilde{M}_{\alpha\alpha'}(t) \to \tilde{M}^T_{\alpha\alpha'}(t). \tag{17}$$

In the following, the superscript T will be dropped. Therefore, by unitary transformation, Eq. (2) may always be written in the form

$$i\frac{d}{dt}C_\alpha(t) = d_\alpha C_\alpha(t) + \sum_{\alpha'} \tilde{M}_{\alpha\alpha'}(t) C_{\alpha'}(t), \tag{18}$$

where the d_α are real numbers and correspond to the energy eigenvalues.[2] Equation (18) is as general as (2).

In order to obtain the approach to the microcanonical equilibrium density matrix instead of the uniform equilibrium density matrix, it will be required that the $\tilde{M}_{\alpha\alpha'}(t)$ in (18) be restricted by the condition that it does not generate couplings between eigenstates α and β, for which $d_\alpha \neq d_\beta$. In this way the fluctuating contribution to the Hamiltonian only couples eigenstates corresponding to the same degenerate energy eigenvalue. Formally, this situation is achieved by the replacement of $\tilde{M}_{\alpha\alpha'}(t)$ with $\tilde{M}_{\alpha\alpha'}(t)\delta(d_\alpha - d_{\alpha'})$, where $\delta(d_\alpha - d_{\alpha'})$ is the Kronecker delta symbol in the two arguments d_α and $d_{\alpha'}$. Equation (18) then becomes

$$i\frac{d}{dt}C_\alpha(t) = d_\alpha C_\alpha(t) + \sum_{\alpha'} \tilde{M}_{\alpha\alpha'}(t)\delta(d_\alpha - d_{\alpha'})C_{\alpha'}(t), \tag{19}$$

which clearly separates into an equation for each distinct degenerate energy eigenvalue manifold of eigenstates.

Therefore, in the following, consideration will be confined to one particular, but otherwise arbitrary, degenerate manifold of eigenstates for which the energy eigenvalue will be denoted by d.[2] Therefore, (19) becomes

$$i\frac{d}{dt}C_\alpha(t) = dC_\alpha(t) + \sum_{\alpha'} \tilde{M}_{\alpha\alpha'}(t) C_{\alpha'}(t), \tag{20}$$

where it is understood that all the eigenstates coupled by $\tilde{M}_{\alpha\alpha'}(t)$ have energy eigenvalue d. With this in mind, properties (3)–(6) characterize the various averaged moments of $\tilde{M}_{\alpha\alpha'}(t)$.

The averaged density matrix equation corresponding with (20) is given by

$$\frac{d}{dt}\langle \rho_{\alpha\beta}(t) \rangle = -\sum_{\alpha'}\sum_{\beta'} R_{\alpha\beta\alpha'\beta'} \langle \rho_{\alpha'\beta'}(t) \rangle. \tag{21}$$

The analog to the first term in the right-hand side of (11) is zero in (21) because

$$L_{\alpha\beta\alpha'\beta'} \equiv \delta_{\alpha\alpha'} M_{\beta\beta'} - \delta_{\beta\beta'} M^*_{\alpha\alpha'} \to d\delta_{\alpha\alpha'}\delta_{\beta\beta'} - d\delta_{\alpha\alpha'}\delta_{\beta\beta'} = 0. \tag{22}$$

The equilibrium state follows from conditions (13) and (14) and is given by

$$\langle \rho_{\alpha\beta}(t) \rangle \xrightarrow[t\to\infty]{} \rho_0 \delta_{\alpha\beta}, \tag{23}$$

where $\rho_0 = 1/N$ if the degeneracy is equal to N. Therefore, each eigenstate in the degenerate manifold becomes equally probably in equilibrium.

It is of particular interest to consider how the total energy behaves during the approach to the microcanonical equilibrium density matrix. The total energy is given by[3]

$$E_{\text{total}}(t) \equiv \sum_\alpha \sum_\beta [d\delta_{\alpha\beta} + \tilde{M}_{\alpha\beta}(t)]\rho_{\alpha\beta}(t) = d + \sum_\alpha \sum_\beta \tilde{M}_{\alpha\beta}(t)\rho_{\alpha\beta}(t). \tag{24}$$

The first term in the second expression follows from conservation of total probability. In the Appendix it is proved that

$$\left\langle \sum_\alpha \sum_\beta \tilde{M}_{\alpha\beta}(t)\rho_{\alpha\beta}(t) \right\rangle = 0. \tag{25}$$

Therefore, the total energy, on the average, is d for all times t, while $\sum_\alpha \sum_\beta \tilde{M}_{\alpha\beta}(t)\rho_{\alpha\beta}(t)$ represents the fluctuations of the total energy around the average value d. These energy considerations complete the treatment of the approach to the microcanonical equilibrium density matrix.

APPROACH TO THE CANONICAL EQUILIBRIUM DENSITY MATRIX

Consider a subsystem in contact with a constant temperature heat reservoir. The complete system will have an equilibrium state characterized by a microcanonical density matrix. However, the equilibrium state of the subsystem will be characterized by a canonical equilibrium density matrix if the heat reservoir is very large. Of interest here is the situation in which the heat reservoir remains in its equilibrium state throughout time while the subsystem relaxes into equilibrium with the reservoir from an initial nonequilibrium state. The problem will be to determine the dynamical equations for the relaxation of the subsystem into its canonical equilibrium density matrix.

Denote the Hamiltonians for the subsystem and the heat reservoir by $\mathbf{H}_S$ and $\mathbf{H}_R$, respectively. It is assumed that the state of reservoir is given, on the average, by its equilibrium state throughout time. Therefore, the in-

teraction between the reservoir and the subsystem is represented by a stationary, purely random, Gaussian Hamiltonian $\tilde{\mathbf{H}}(t)$.

Latin indices will be used to denote eigenstates of the subsystem

$$\mathbf{H}_S|i\rangle = E_i|i\rangle. \tag{26}$$

Greek indices will be used to denote reservoir eigenstates

$$\mathbf{H}_R|\alpha\rangle = E_\alpha|\alpha\rangle. \tag{27}$$

Generally, $\tilde{\mathbf{H}}(t)$ will have nonzero matrix elements in the direct product manifold of the two eigenstate manifolds of $\mathbf{H}_S$ and $\mathbf{H}_R$. Denoting the identity matrices for the subsystem eigenstate manifold and for the reservoir eigenstate manifold by $\mathbf{1}_S$ and $\mathbf{1}_R$, respectively, the total Hamiltonian for the complete system may be written as

$$\mathbf{H}_{\text{total}} = \mathbf{H}_S \otimes \mathbf{1}_R + \mathbf{1}_S \otimes \mathbf{H}_R + \tilde{\mathbf{H}}(t), \tag{28}$$

The Schrödinger wavefunction $\psi(t)$ may be expanded in terms of direct product basis states.

$$\psi(t) = \sum_i \sum_\alpha C_{i\alpha}(t)|i\rangle|\alpha\rangle. \tag{29}$$

With the Hamiltonian given by (28), (29) leads to[2]

$$i\frac{d}{dt}C_{i\alpha}(t) = (E_i + E_\alpha)C_{i\alpha}(t) + \tilde{H}_{i\alpha j\beta}(t)C_{j\beta}(t), \tag{30}$$

where $\tilde{H}_{i\alpha j\beta}(t)$ is defined by

$$\tilde{H}_{i\alpha j\beta}(t) \equiv \langle\alpha|\langle i|\tilde{H}(t)|j\rangle|\beta\rangle. \tag{31}$$

In (30) and throughout the remainder of this section the repeated index summation convention is used.

In order to insure that this description leads to a microcanonical equilibrium density matrix for the complete system, it is necessary to restrict $\tilde{H}_{i\alpha j\beta}(t)$ to be zero unless $E_i + E_\alpha = E_j + E_\beta$. This restriction is schematized in Fig. 1. With this restriction, (3) is a special case of (20) if the substitutions

$$d \to (E_i + E_\alpha) \equiv E_{\text{total}} \quad \text{and} \quad \tilde{M}_{\alpha\alpha'}(t) \to \tilde{H}_{i\alpha j\beta}(t) \tag{32}$$

are made, and if (30) is restricted to a single degenerate manifold of eigenstates for the complete system with total energy E_{total}. Because of a result analogous with (25), the energy eigenstates of the complete Hamiltonian are, on the average, direct products of the energy eigenstates of the subsystem and reservoir Hamiltonians.

The density matrix is defined by[3]

$$\rho_{i\alpha j\beta}(t) \equiv C^*_{i\alpha}(t)C_{j\beta}(t). \tag{33}$$

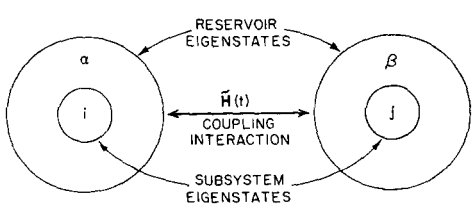

FIG. 1. $\tilde{H}_{i\alpha j\beta} = 0$ unless $E_i + E_\alpha = E_j + E_\beta$.

The stochastically averaged density matrix satisfies the equation[2]

$$\frac{d}{dt}\langle\rho_{i\alpha j\beta}(t)\rangle = -iL_{i\alpha j\beta i'\alpha' j'\beta'}\langle\rho_{i'\alpha' j'\beta'}(t)\rangle - R_{i\alpha j\beta i'\alpha' j'\beta'}\langle\rho_{i'\alpha' j'\beta'}(t)\rangle, \tag{34}$$

where

$$L_{i\alpha j\beta i'\alpha' j'\beta'} = \delta_{ii'}\delta_{\alpha\alpha'}(E_j + E_\beta)\delta_{jj'}\delta_{\beta\beta'} - \delta_{jj'}\delta_{\beta\beta'} \times (E_i + E_\alpha)\delta_{ii'}\delta_{\alpha\alpha'}. \tag{35}$$

The expression in (35) is actually equal to zero since the direct product eigenstates are restricted to the manifold satisfying

$$E_i + E_\alpha = E_{\text{total}} = E_j + E_\beta, \tag{36}$$

as is implied by (32). This corresponds to the absence of an L term in (21). However, an L term is explicitly indicated in (34) since it will actually manifest itself later. The $R_{i\alpha j\beta i'\alpha' j'\beta'}$ in (34) is defined by

$$R_{i\alpha j\beta i'\alpha' j'\beta'} \equiv \delta_{ii'}\delta_{\alpha\alpha'}Q_{j\beta\theta\theta'\theta\theta' j'\beta'} + \delta_{jj'}\delta_{\beta\beta'}Q_{\theta\theta' i\alpha i'\alpha'\theta\theta'} - Q_{j\beta j'\beta' i'\alpha' i\alpha} - Q_{i'\alpha' i\alpha j\beta j'\beta'}, \tag{37}$$

where

$$\langle\tilde{H}_{i\alpha j\beta}(t)\tilde{H}_{i'\alpha' j'\beta'}(s)\rangle = 2Q_{i\alpha j\beta i'\alpha' j'\beta'}\delta(t-s). \tag{38}$$

The condition that the reservoir state remain the equilibrium state throughout time is imposed by assuming that the averaged density matrix $\langle\rho_{i\alpha j\beta}(t)\rangle$ factors into a direct product of the subsystem density matrix, and the reservoir density matrix in which the reservoir density matrix is given, for all times, by its equilibrium density matrix. The reservoir equilibrium density matrix must be the canonical density matrix because of: (36), the condition that the complete system has a microcanonical equilibrium density matrix, and the condition that the subsystem has a canonical equilibrium density matrix. Formally, the factorization is given by

$$\langle\rho_{i\alpha j\beta}(t)\rangle \to \langle\rho_{ij}(t)\rangle\langle\rho_{\alpha\beta}\rangle, \tag{39}$$

where

$$\langle\rho_{\alpha\beta}\rangle = (1/Q_R)\exp[-(E_\alpha/K_BT)]\delta_{\alpha\beta}, \tag{40}$$

where $Q_R = \sum_\alpha \exp[-(E_\alpha/K_BT)]$, K_B is Boltzmann's constant, and T is the temperature of the reservoir.

Putting (39) and (40) into (34), followed by taking the trace over reservoir eigenstates, produces the following equation for the averaged subsystem density matrix:

$$\frac{d}{dt}\langle\rho_{ij}(t)\rangle = -i(E_j - E_i)\langle\rho_{ij}(t)\rangle - R_{i\alpha j\alpha i'\alpha' j'\alpha'}\frac{1}{Q_R}\exp\left(-\frac{E_{\alpha'}}{K_BT}\right)\langle\rho_{i'j'}(t)\rangle. \tag{41}$$

Note that the L term of (34) does contribute to (41).

By defining $T_{iji'j'}$ by

$$T_{iji'j'} \equiv R_{i\alpha j\alpha i'\alpha' j'\alpha'}(1/Q_R)\exp[-(E_{\alpha'}/K_BT)], \tag{42}$$

Eq. (41) becomes

$$\frac{d}{dt}\langle \rho_{ij}(t)\rangle = -i(E_j - E_i)\langle \rho_{ij}(t)\rangle - T_{iji'j'}\langle \rho_{i'j'}(t)\rangle. \tag{43}$$

It is of interest to consider the relationship between $T_{iji'j'}$ and $T_{i'j'ij}$. The use of (37), (38), and Fig. 1 is required.

Figure 1 and (38) imply that

$$Q_{i\alpha j\beta i'\alpha' j'\beta'} = 0 \quad \text{unless} \quad E_i + E_\alpha = E_j + E_\beta$$

$$\text{and} \quad E_{i'} + E_{\alpha'} = E_{j'} + E_{\beta'}. \tag{44}$$

This may be written with Kronecker deltas as

$$Q_{i\alpha j\beta i'\alpha' j'\beta'}\,\delta(E_i + E_\alpha - E_j - E_\beta)\delta(E_{i'} + E_{\alpha'} - E_{j'} - E_{\beta'}). \tag{45}$$

Using expressions similar to (45) in (37) leads to

$$\begin{aligned} R_{i\alpha j\beta i'\alpha' j'\beta'} = [&\delta_{ii'}\delta_{\alpha\alpha'}Q_{j\beta\theta\theta'\theta\theta' j'\beta'}\,\delta(E_j + E_\beta - E_\theta - E_{\theta'}) \\ &\times \delta(E_\theta + E_{\theta'} - E_{j'} - E_{\beta'}) \\ &+ \delta_{jj'}\delta_{\beta\beta'}Q_{\theta\theta' i\alpha i'\alpha'\theta\theta'}\,\delta(E_\theta + E_{\theta'} - E_i - E_\alpha) \\ &\times \delta(E_{i'} + E_{\alpha'} - E_\theta - E_{\theta'}) - 2Q_{j\beta j'\beta' i'\alpha' i\alpha}] \\ &\times \delta(E_j + E_\beta - E_{j'} - E_{\beta'})\delta(E_i + E_\alpha - E_{i'} - E_{\alpha'}). \end{aligned} \tag{46}$$

Using (46), along with appropriate index changes, in (42) gives

$$\begin{aligned} T_{iji'j'} &= [\delta_{ii'}\delta_{\alpha\alpha'}Q_{j\alpha\theta\theta'\theta\theta' j'\alpha'}\,\delta(E_j + E_\alpha - E_\theta - E_{\theta'}) \\ &\quad\times \delta(E_\theta - E_{\theta'} - E_{j'} - E_{\alpha'}) + \delta_{jj'}\delta_{\alpha\alpha'}Q_{\theta\theta' i\alpha i'\alpha'\theta\theta'} \\ &\quad\times \delta(E_\theta + E_{\theta'} - E_i - E_\alpha)\delta(E_{i'} + E_{\alpha'} - E_\theta - E_{\theta'}) \\ &\quad - 2Q_{j\alpha j'\alpha' i'\alpha' i\alpha}]\delta(E_j + E_\alpha - E_{j'} - E_{\alpha'}) \\ &\quad\times \delta(E_i + E_\alpha - E_{i'} - E_{\alpha'})(1/Q_R)\exp[-(E_{\alpha'}/K_B T)] \\ &= [\delta_{ii'}\delta_{\alpha\alpha'}Q^*_{j'\alpha'\theta\theta'\theta\theta' j\alpha}\,\delta(E_j + E_\alpha - E_\theta - E_{\theta'}) \\ &\quad\times \delta(E_\theta + E_{\theta'} - E_{j'} - E_{\alpha'}) + \delta_{jj'}\delta_{\alpha\alpha'}Q^*_{\theta\theta' i'\alpha' i\alpha\theta\theta'} \\ &\quad\times \delta(E_\theta + E_{\theta'} - E_i - E_\alpha)\delta(E_{i'} + E_{\alpha'} - E_\theta - E_{\theta'}) \\ &\quad - 2Q^*_{j'\alpha' j\alpha i\alpha i'\alpha'}]\delta(E_j + E_\alpha - E_{j'} - E_{\alpha'}) \\ &\quad\times \delta(E_i + E_\alpha - E_{i'} - E_{\alpha'}) \\ &\quad\times (1/Q_R)\exp[-(E_{\alpha'}/K_B T)] \\ &= [\delta_{ii'}\delta_{\alpha\alpha'}Q^*_{j'\alpha\theta\theta'\theta\theta' j\alpha}\,\delta(E_{j'} + E_\alpha - E_\theta - E_{\theta'}) \\ &\quad\times \delta(E_\theta + E_{\theta'} - E_j - E_{\alpha'}) + \delta_{jj'}\delta_{\alpha\alpha'}Q^*_{\theta\theta' i'\alpha i\alpha'\theta\theta'} \\ &\quad\times \delta(E_\theta + E_{\theta'} - E_{i'} - E_\alpha)\delta(E_i + E_{\alpha'} - E_\theta - E_{\theta'}) \\ &\quad - 2Q^*_{j'\alpha j\alpha' i\alpha' i'\alpha}]\delta(E_{j'} + E_\alpha - E_j - E_{\alpha'}) \\ &\quad\times \delta(E_i + E_{\alpha'} - E_{i'} - E_\alpha) \\ &\quad\times (1/Q_R)\exp[-(E_\alpha/K_B T)]. \end{aligned} \tag{47}$$

The second equality in (47) follows from the Hermiticity of $\tilde{H}(t)$ in (38), while the third equality follows from renaming the indices α and α' according to the interchange $\alpha \leftrightarrow \alpha'$. In the last expression for $T_{iji'j'}$ in (47), the factors $\delta(E_{j'} + E_\alpha - E_j - E_{\alpha'})\delta(E_i + E_{\alpha'} - E_{i'} - E_\alpha)$ require that $E_\alpha = E_{\alpha'} + E_j - E_{j'} = E_{\alpha'} + E_i - E_{i'}$, which may be combined to give

$$E_\alpha = E_{\alpha'} + \tfrac{1}{2}(E_j + E_i - E_{j'} - E_{i'}). \tag{48}$$

This means that

$$\begin{aligned} &\delta(E_{j'} + E_\alpha - E_j - E_{\alpha'})\delta(E_i + E_{\alpha'} - E_{i'} - E_\alpha) \\ &\quad\times (1/Q_R)\exp[-(E_\alpha/K_B T)] \\ &= \delta(E_{j'} + E_\alpha - E_j - E_{\alpha'})\delta(E_i + E_{\alpha'} - E_{i'} - E_\alpha)(1/Q_R) \\ &\quad\times \exp[-(E_{\alpha'}/K_B T)]\exp[-(E_j + E_i \\ &\quad - E_{j'} - E_{i'})/2K_B T]. \end{aligned} \tag{49}$$

Putting (49) into the last expression for $T_{iji'j'}$ in (47) gives

$$\begin{aligned} T_{iji'j'} = [&\delta_{ii'}\delta_{\alpha\alpha'}Q^*_{j'\alpha\theta\theta'\theta\theta' j\alpha'}\,\delta(E_{j'} + E_\alpha - E_\theta - E_{\theta'}) \\ &\times \delta(E_\theta + E_{\theta'} - E_j - E_{\alpha'}) + \delta_{jj'}\delta_{\alpha\alpha'} \\ &\times Q^*_{\theta\theta' i'\alpha i\alpha'\theta\theta'}\,\delta(E_\theta + E_{\theta'} - E_{i'} - E_\alpha) \\ &\times \delta(E_i + E_{\alpha'} + E_\theta - E_{\theta'}) - 2Q^*_{j'\alpha j\alpha' i\alpha' i'\alpha}] \\ &\times \delta(E_{j'} + E_\alpha - E_j - E_{\alpha'})\delta(E_i + E_{\alpha'} - E_{i'} - E_\alpha) \\ &\times \frac{1}{Q_R}\exp\left(-\frac{E_{\alpha'}}{K_B T}\right) \\ &\times \exp\left(-\frac{(E_j + E_i - E_{j'} - E_{i'})}{2K_B T}\right). \end{aligned} \tag{50}$$

However, using (46) in (42) in order to calculate $T_{i'j'ij}$, directly verifies that (50) is simply

$$T_{iji'j'} = T^*_{i'j'ij}\exp[-(E_j + E_i - E_{j'} - E_{i'})/2K_B T]. \tag{51}$$

Together, Eqs. (43) and (51) provide the dynamical description of the temporal approach to equilibrium of the averaged density matrix for the subsystem. Equation (51) is a generalized detailed balancing condition.

It can now be proved that the canonical equilibrium density matrix is obtained asymptotically:

$$\langle \rho_{ij}(t)\rangle \xrightarrow[t\to\infty]{} (1/Q_S)\exp[-(E_j/K_B T)]\delta_{ij}, \tag{52}$$

where $Q_S = \sum_j \exp[-(E_j/K_B T)]$. The proof of (52) uses the first equality given in (47) to show that

$$T_{iji'j'}(1/Q_S)\exp[-(E_{j'}/K_B T)]\delta_{i'j'} = 0. \tag{53}$$

Proof of (53):

From (47) it follows that

$$\begin{aligned} &T_{iji'j'}(1/Q_S)\exp[-(E_{j'}/K_B T)]\delta_{i'j'} \\ &\quad= [\delta_{ii'}\delta_{\alpha\alpha'}Q_{j\alpha\theta\theta'\theta\theta' j'\alpha'}\,\delta(E_j + E_\alpha - E_\theta - E_{\theta'}) \\ &\quad\quad\times \delta(E_\theta + E_{\theta'} - E_{j'} - E_{\alpha'}) + \delta_{jj'}\delta_{\alpha\alpha'} \end{aligned}$$

$$\times Q_{\theta\theta' i\alpha i'\alpha'\theta\theta'}\,\delta(E_\theta + E_{\theta'} - E_i - E_\alpha)$$
$$\times \delta(E_{i'} + E_{\alpha'} - E_\theta - E_{\theta'})$$
$$- 2Q_{j\alpha j'\alpha' i'\alpha' i\alpha}]\delta(E_j + E_\alpha - E_{j'} - E_{\alpha'})$$
$$\times \delta(E_{i'} + E_{\alpha'} - E_i - E_\alpha)$$
$$\times \frac{1}{Q_R}\exp\left(-\frac{E_{\alpha'}}{K_B T}\right)\frac{1}{Q_S}\exp\left(-\frac{E_{j'}}{K_B T}\right)\delta_{i'j'}$$
$$= [\delta_{\alpha\alpha'} Q_{j\alpha\theta\theta'\theta\theta' i\alpha'}\,\delta(E_j + E_\alpha - E_\theta - E_{\theta'})$$
$$\times \delta(E_\theta + E_{\theta'} - E_i - E_{\alpha'}) + \delta_{\alpha\alpha'} Q_{\theta\theta' i\alpha j\alpha'\theta\theta'}$$
$$\times \delta(E_\theta + E_{\theta'} - E_i - E_\alpha)\delta(E_j + E_{\alpha'} - E_\theta - E_{\theta'})$$
$$- 2Q_{j\alpha\theta\alpha'\theta\alpha' i\alpha}]\delta(E_j + E_\alpha - E_\theta - E_{\alpha'})\delta(E_\theta + E_{\alpha'}$$
$$- E_i - E_\alpha)\,(1/Q_R Q_S)\exp[-(E_{\text{total}}/K_B T)]$$
$$= (Q_{j\alpha\theta\theta'\theta\theta' i\alpha} + Q_{\theta\theta' i\alpha j\alpha\theta\theta'} - 2Q_{j\alpha\theta\theta'\theta\theta' i\alpha})$$
$$\times \delta(E_j + E_\alpha - E_\theta - E_{\theta'})\delta(E_\theta + E_{\theta'} - E_i - E_\alpha)$$
$$\times (1/Q_R Q_S)\exp[-(E_{\text{total}}/K_B T)] = 0. \tag{54}$$

The second equality in (54) used (36), while the fourth equality used (38). This completes the proof of (53).

SUMMARY

It has been shown how the theory of multiplicative stochastic processes leads to a description of the approach to equilibrium of the density matrix for a quantum mechanical system. In the case of a system which maintains a constant average total energy, the approach to the microcanonical equilibrium density matrix is described. In the case of a subsystem in contact with a constant temperature heat reservoir, the approach to the canonical equilibrium density matrix is described. This last case provides a unified treatment of some aspects of the problems of magnetic resonance relaxation, spectral line shape when the line shape is Lorentzian, and molecular reaction relaxation phenomena.[5,6] Detailed accounts of these and other problems from the perspective presented in this paper remains to be presented.

APPENDIX

Proof of (25): Using Eq. (20), $C_\beta(t)$ can be written as

$$C_\beta(t) = e^{-idt}\sum_{\beta'}\sum_{k=0}(-i)^k\int_0^t\int_0^{S_k}\int_0^{S_{k-1}}\cdots\int_0^{S_3}\int_0^{S_2}\sum_{\mu_{k-1}}\sum_{\mu_{k-2}}$$
$$\times\cdots\sum_{\mu_2}\sum_{\mu_1}\tilde{M}_{\beta\mu_{k-1}}(S_k)\tilde{M}_{\mu_{k-1}\mu_{k-2}}(S_{k-1})\cdots\tilde{M}_{\mu_2\mu_1}(S_2)$$
$$\times\tilde{M}_{\mu_1\beta'}(S_1)dS_1\cdots dS_k C_{\beta'}(0). \tag{55}$$

Similarly, $C_\alpha^*(t)$ can be written as

$$C_\alpha^*(t) = e^{idt}\sum_{\alpha'}\sum_{l=0}^{\infty}(i)^l\int_0^t\int_0^{S_l}\int_0^{S_{l-1}}\cdots\int_0^{S_3}\int_0^{S_2}\sum_{\nu_1}\sum_{\nu_2}$$
$$\times\cdots\sum_{\nu_{l-2}}\sum_{\nu_{l-1}}\tilde{M}_{\alpha'\nu_1}(S_1)\tilde{M}_{\nu_1\nu_2}(S_2)\cdots\tilde{M}_{\nu_{l-2}\nu_{l-1}}(S_{l-1})$$
$$\times\tilde{M}_{\nu_{l-1}\alpha}(S_l)dS_1\cdots dS_l C_{\alpha'}^*(0). \tag{56}$$

In (56) the Hermiticity of $\tilde{M}(t)$ has been used. In both (55) and (56) the multiple integrals are time ordered with $t \geq S_k \geq S_{k-1} \geq \cdots \geq S_2 \geq S_1 \geq 0$. Using (7), the quantity of interest in (25) may be written as

$$\left\langle\sum_\alpha\sum_\beta\tilde{M}_{\alpha\beta}(t)\rho_{\alpha\beta}(t)\right\rangle$$
$$= \left\langle\sum_\alpha\sum_\beta C_\alpha^*(t)\tilde{M}_{\alpha\beta}(t)C_\beta(t)\right\rangle$$
$$= \sum_\alpha\sum_\beta\sum_{\alpha'}\sum_{\beta'}\sum_{l=0}^{\infty}\sum_{k=0}^{\infty}(i)^l(-i)^k\int_0^t\int_0^{S_l'}\cdots\int_0^{S_2'}\int_0^t\int_0^{S_k}$$
$$\times\cdots\int_0^{S_2}\sum_{\nu_1}\cdots\sum_{\nu_{l-1}}\sum_{\mu_{k-1}}\cdots\sum_{\mu_1}\langle\tilde{M}_{\alpha'\nu_1}(S_1')\cdots\tilde{M}_{\nu_{l-1}\alpha}(S_l')$$
$$\times\tilde{M}_{\alpha\beta}(t)\tilde{M}_{\beta\mu_{k-1}}(S_k)\cdots\tilde{M}_{\mu_1\beta'}(S_1)\rangle\,C_{\alpha'}^*(0)C_{\beta'}(0)dS_1$$
$$\times\cdots dS_k dS_1'\cdots dS_l'$$
$$= \sum_{l=0}^{\infty}\sum_{k=0}^{\infty}(i)^l(-i)^k\sum_{\alpha'}\sum_{\beta'}\int_0^t\int_0^{S_l'}\cdots\int_0^{S_2'}\int_0^t\int_0^{S_k}\cdots\int_0^{S_2}$$
$$\times\sum_\alpha\sum_\beta\sum_{\nu_1}\cdots\sum_{\nu_{l-1}}\sum_{\mu_{k-1}}\cdots\sum_{\mu_1}\langle\tilde{M}_{\alpha'\nu_1}(S_1')$$
$$\times\cdots\tilde{M}_{\nu_{l-1}\alpha}(S_l')\,\tilde{M}_{\alpha\beta}(t)\tilde{M}_{\beta\mu_{k-1}}(S_k)\cdots\tilde{M}_{\mu_1\beta'}(S_1)\rangle$$
$$\times C_{\alpha'}^*(0)C_{\beta'}(0)dS_1\cdots dS_k\,dS_1'\cdots dS_l'. \tag{57}$$

Two cases need to be considered in evaluating the last expression (57). These two cases are (a) $k + l$ is even and (b) $k + l$ is odd.

Case (a): If $k + l$ is even, then there are $k + l + 1$ $\tilde{\mathbf{M}}(t)$'s in the product, and the average will be zero because of condition (5).

Case (b): If $k + l$ is odd, then there are $k + l + 1$ $\tilde{\mathbf{M}}(t)$'s in the product, and the average will not be zero because of condition (6). However, it will be shown below that these nonzero terms occur in pairs of opposite sign so that the sum of all such nonzero terms is zero.

Let $p + q$ be odd and consider the two terms: $k = p$ and $l = q$, and $k = q$ and $l = p$. In the first case the last expression in (57) contains the factor $(i)^q(-i)^p$, whereas in the second case it contains the factor $(i)^p(-i)^q$. However, $p + q$ is odd implies

$$(i)^q(-i)^p = (-1)^p(i)^{p+q} = -(-1)^q(i)^{q+p} = -(i)^p(-i)^q. \tag{58}$$

Therefore, if the remaining integral factors of these two terms are equal, then a pair of nonzero terms with opposite signs have been identified.

The remaining integral factor in the $k = p$ and $l = q$ case is

$$\sum_{\alpha'}\sum_{\beta'}\int_0^t\int_0^{S_q'}\cdots\int_0^{S_2'}\int_0^t\int_0^{S_p}\cdots\int_0^{S_2}\sum_\alpha\sum_\beta\sum_{\nu_1}\cdots\sum_{\nu_{q-1}}\sum_{\mu_{p-1}}\cdots$$
$$\times\sum_{\mu_1}\langle\tilde{M}_{\alpha'\nu}(S_1')\cdots\tilde{M}_{\nu_{q-1}\alpha}(S_q')\tilde{M}_{\alpha\beta}(t)\tilde{M}_{\beta\mu_{p-1}}(S_p)$$
$$\times\cdots\tilde{M}_{\mu_1\beta'}(S_1)\rangle\,C_{\alpha'}^*(0)C_{\beta'}(0)dS_1\cdots dS_p\,dS_1'\cdots dS_q' \tag{59}$$

In the $k = q$ and $l = p$ case the remaining integral factor is

$$\sum_{\alpha'}\sum_{\beta'}\int_0^t\int_0^{S_p'}\cdots\int_0^{S_2'}\int_0^t\int_0^{S_q}\cdots\int_0^{S_2}\sum_{\alpha}\sum_{\beta}\sum_{\nu_1}\cdots\sum_{\nu_{p-1}}\sum_{\mu_{q-1}}\cdots$$

$$\times\sum_{\mu_1}\langle\tilde{M}_{\alpha'\nu_1}(S_1')\cdots\tilde{M}_{\nu_{p-1}\alpha}(S_p')\tilde{M}_{\alpha\beta}(t)\tilde{M}_{\beta\mu_{q-1}}(S_q)$$

$$\times\cdots\tilde{M}_{\mu_1\beta'}(S_1)\rangle\, C_{\alpha'}^*(0)C_{\beta'}(0)dS_1\cdots dS_q\,dS_1'\cdots dS_p'. \tag{60}$$

Therefore, the proof of (25) is reduced to proving that (59) and (60) are equal. That (59) and (60) are equal follows from the property that the trace of a product of matrices equals the trace of the transpose of the product of matrices. This is explicitly illustrated above by considering $C_{\alpha'}^*(0)C_{\beta'}(0)$ as a matrix, using the transpose property of the trace, renaming indices, and relabeling the variables of integration. Thereby, the expression in (60) may be transformed into the expression in (59). This completes the proof of (25).

---

[1]R. F. Fox, J. Math. Phys. **13**, 1196 (1972).

[2]Throughout this paper physical units are assumed in terms of which Planck's constant h has the value 1 and, therefore, does not explicitly appear in any equation.

[3]This definition is to be contrasted with the definition given by R. C. Toleman, *The principles of statistical mechanics* (Oxford U.P., New York, 1962), Chap. IX. The expressions for density matrix averages of quantum mechanical operators are, therefore, also different.

[4]R. C. Tolman, Ref. 3, Chap. IX.

[5]R. Kubo and N. Hashitsume, Prog. Theor. Phys. Suppl. (46), 210 (1970).

[6]M. Blume, Phys. Rev. **174** (2), 351 (1968); M. Clauser and M. Blume, Phys. Rev. B **3** (3), 583 (1971).

Locality conditions on form factors*

Bo Andersson†
Joseph Henry Laboratories of Physics, Princeton, New Jersey 08540
(Received 9 June 1972)

We derive a set of integral equations which are necessary and sufficient conditions on the form factors of local field theory, i.e. on the matrix elements of local operators. The basic idea is that out of the set of all (distribution-valued) functions defined on the boundary of an analyticity domain, in general only a subset are boundary values of functions which are analytic within the domain. The form factors are boundary values of a vertex function which, due to the general assumptions of locality, reasonable energy, and mass spectrum and Poincaré covariance, is analytic at least in the domain constructed by Källén and Wightman. The characteristic boundary of the domain ("the distinguished boundary") is the set of physical values of the arguments of the form factors, and the integral equations in that way only involve such values. The main advantages in formulating the locality conditions in this way are that (1) only the physical quantities of the field theory, i.e., the matrix elements between the field operators, enter into the equations and (2) the frustrating complications which are met in the construction of the domains of analyticity for n-point functions with $n > 3$ might hopefully be avoided because the distinguished boundary can be constructed even if the whole domain is not known. The integral equations have naturally no unique solutions, because, e.g., all perturbation-theoretical form factors must, of course, fulfill them. The equations may, however, function as a convenient starting point for approximations and "model building" for form factors outside the presently used perturbation theories. The integral equations are straightforward generalizations of the notion of "weak local commutativity" for the two-point function. This condition means that the two spectral functions connected to two locally commuting operators should be equal. The conditions on the form factors (which are the generalizations of the two-point spectral functions) are that the difference should vanish when integrated over certain physical sets of mass space.

1. INTRODUCTION

In a series of earlier papers[1,2] we have investigated the properties of the vertex-function particularly in momentum space. The basic input has been the analyticity properties proved by Källén and Wightman[3] (this paper is hereafter referred to as KW) from some general assumptions which are expected to be fulfilled in all (interesting) field theories. In a nonepsilontic way the assumptions are that the field theory exhibits[4,5]

(1) *Lorentz covariance* and *translation invariance*,

(2) *"reasonable" mass and energy spectrum*

(3) *locality* in the sense of *local commutativity*.

The following results from KW are of interest for this paper:

There is a function $G = G(Z_1, Z_2, Z_3)$ depending upon the three complex variables Z_j, $j = 1, 2, 3$. These variables can be interpreted as the Lorentz squares ("mass squares") of three complex energy–momentum vectors $\tilde{p}_j$ fulfilling energy–momentum conservation:

$$\sum_{j=1}^{3} \tilde{p}_j = 0, \quad \tilde{p}_j = p_j + ik_j, \quad Z_j = -\tilde{p}_j^2, \quad j = 1, 2, 3. \quad (1)$$

(metric $p^2 = p^2 - p_0^2$)

There is further a domain D_{KW} in the three-dimensional complex space of the variables Z_j defined above such that

(i) the function G is analytic at least inside the domain D_{KW},

(ii) the matrix elements of the field operators in the theory are different boundary values of the function G on the real axes of the variables Z_j (which constitute parts of the boundaries of the domain D_{KW}).

In order to make use of the above-mentioned analyticity properties of the vertex function, we will make the additional assumption that G exhibits certain boundedness properties along the boundary of the domain D_{KW}. We will in particular assume moderate integrability properties for the boundary values of G (fulfilled, e.g., for temperate distributions) and at most polynomial growth in asymptotic directions (i.e., when one or more of the variables Z_j approaches infinity) inside the domain D_{KW}.

We will actually be satisfied to investigate the properties of the "absorptive parts" of the vertex function G. Due to the symmetry between the variables Z_j it is sufficient to consider the quantity $\Delta_3 G$ defined by

$$\Delta_3 G(Z_1, Z_2, b_3) = \lim_{\epsilon \to 0} [G(Z_1, Z_2, b_3 + i\epsilon) - G(Z_1, Z_2, b_3 - i\epsilon)]. \quad (2)$$

The quantity b_3 in Eq. (2) will always be a positive real number. Then the absorptive part defined in that way is the discontinuity of the vertex function across the real positive axis of the third variable. This is one of the possible "axiomatic" singularity surfaces of the vertex function G according to the investigation in KW ("the Z_3-cut"). This discontinuity can be interpreted as the contribution from the states with squared mass b_3 and with the quantum numbers of one of the fields (here the C-field; cf. the formulas in Sec. 2B). The absorptive part $\Delta_3 G$ will in general exhibit distribution properties considered as function of b_3 while it is analytic as function of the remaining variables Z_1 and Z_2.

The corresponding quantities $\Delta_1 G$ and $\Delta_2 G$ with obvious definitions exhibit the same properties as $\Delta_3 G$ after appropriate exchange of indices. It is possible to derive representation formulas ("dispersion relations") (cf. Ref. 1 hereafter referred to as DI and DII) by means of which the absorptive part $\Delta_3 G$ is determined from its boundary values. In some cases these boundary values are the above-mentioned matrix elements which we will refer to in the following as "form factors" (a precise definition is given in Sec. 2B). In the other cases it is

possible (according to the results in Sec. 3) to express the boundary values in the representation formula as functionals of the form factors.

Additional information on the properties of the vertex function will then via the representation formulas imply restrictions on the form factors. If, e.g., the vertex functions is assumed to vanish at a certain rate at infinity it is possible to write "super convergence relations" or "sum rules" for the form factors (cf. Ref. 2 hereafter referred to as SI, SII, SIII).

There are, however, *basic restrictions on the form factors from the mere fact that they are boundary values of an analytic function.* To see that in more detail, consider the set of functions $\mathcal{K}(D)$, which are analytic in a certain domain D. From $\mathcal{K}(D)$ it is possible to construct the set of distributions $\delta\mathcal{K}$ which are boundary values on the boundary of D, δD, of functions in $\mathcal{K}(D)$. Then the set of $\delta\mathcal{K}$ is a subset of the set $\mathcal{L}(\delta D)$ of *all* distributions defined on δD. *In general* $\delta\mathcal{K}$ *does not coincide with* $\mathcal{L}(\delta D)$.

The kind of restrictions that are necessary and sufficient in order that a distribution in $\mathcal{L}(\delta D)$ also is in $\delta\mathcal{K}$ depends upon the domain D and in particular upon "the appearance" of the boundary δD. To clarify the problem, we will consider a few examples with well-known solutions.

In the first example the domain is the unit circle in one complex dimension. Any function $k(Z)$, analytic within the unit circle, can be represented by Cauchy's theorem as

$$k(Z) = \frac{1}{2\pi}\int_0^{2\pi} d\theta\ f(\theta)(1 - e^{-i\theta}Z)^{-1}. \tag{3}$$

The weight-function in (3) can be chosen as the (in general, distribution valued) boundary value $k(e^{i\theta})$. On the other hand an arbitrary weight function f (only restricted so that the integral exists) defines via Eq. (3) a function k analytic inside the unit circle. The problem is to characterize those weight functions $f(\theta)$ which are boundary values of the corresponding analytic function, i.e., which (in a distribution theoretical sense) fulfills

$$f(\theta) = \lim_{\eta\to 1} k(\eta e^{i\theta}). \tag{4}$$

That the problem is nontrivial can be seen from the weight function $f(\theta) = e^{-i\theta}$. This weight function gives the result $k = 0$ in Eq. (3), and therefore Eq. (4) cannot be fulfilled.

The solution to the problem is actually related to this fact. If the weight function $f(\theta)$ is represented as a Fourier series (which is always possible in distribution sense)

$$f(\theta) = \sum_{n=-\infty}^{\infty} f_n e^{-in\theta}; \tag{5}$$

then the condition that $f(\theta)$ is a limit according to Eq. (4) can be formulated as the following set of integral equations:

$$0 = f_n = \frac{1}{2\pi}\int_0^{2\pi} d\theta\ f(\theta)e^{-in\theta}; \qquad n < 0. \tag{6}$$

As a second and maybe more well-known example (and actually a special case of the one above), we will consider functions analytic in the whole complex plane with a cut along the positive real axis. Any function γ with these analyticity properties and vanishing around infinity can be written as

$$\gamma(Z) = \frac{1}{2\pi i}\int_0^{\infty} \frac{\phi(a)}{a - z}\, da. \tag{7}$$

The function ϕ *can be identified* with "the imaginary part" of γ in the following sense:

$$\begin{aligned}\lim_{\epsilon\to 0}[\gamma(b + i\epsilon) - \gamma(b - i\epsilon)] &= (2\pi i)^{-1}\int \phi(a)da\ 2\pi i\delta(a - b) = \phi(b), \quad b > 0,\\ &= 0, \quad b < 0.\end{aligned} \tag{8}$$

Then "the real part" of the function γ is determined by "the imaginary part" ϕ by means of a principal-value integral:

$$\lim_{\epsilon\to 0}[\gamma(b + i\epsilon) + \gamma(b - i\epsilon)] = (\pi i)^{-1}\int da\phi(a)/(a - b)p. \tag{8'}$$

This relation ("dispersion relation") is actually equivalent to the conditions in Eq. (6) if the "cut plane domain" above is mapped into the unit circle, e.g., by the mapping $w = (i - \sqrt{z})(i + \sqrt{Z})^{-1}$.

In the case of several variables there are some further complications.[6] Thus an arbitrary domain in the $2n$ (real) dimensional space of n complex numbers cannot be a domain of analyticity. In general it is possible to continue every analytic function regular in a given domain into a larger domain. This larger domain is called the "envelope of holomorphy" of the given domain. The reason for this phenomena is that the Cauchy–Riemann equations, which are the conventional criterion of analyticity, do in the case of several complex variables "overdetermine" the function. This can also be seen in the fact that all the functions analytic in an analyticity domain are actually determined by the values on only a low-dimensional part of the boundary of the domain. As an example consider the function K which is analytic in the topological product of n unit circles:

$$K(Z_1, \ldots, Z_n) = \frac{1}{(2\pi)^n}\int\prod_{j=1}^{n} \frac{d\theta_j}{1 - e^{-i\theta}jZ_j} f(\theta_1, \ldots, \theta_n) \tag{9}$$

The actual integration region is in Eq. (9) only n (real) dimensional while the dimension of the boundary in general is $(2n - 1)$.

The integration domain, which consequently determines K completely, is known as "the distinguished boundary," and intuitively it corresponds to "the utmost corners" of a domain of analyticity.

The problem in several variables which corresponds to the one discussed above for one variable is then to characterize those distributions which are defined on the distinguished boundary of a given domain of analyticity and are boundary values of functions which are analytic inside the domain.

For the particular domain above the problem is a straightforward generalization of the condition in Eq. (6). Thus, the weight function f is the limit of

$$f(\theta_1, \ldots \theta_n) = \lim_{\eta_1\to 1} \ldots \lim_{\eta_n\to 1} K(\eta_1 e^{i\theta_1}, \ldots, \eta_n e^{i\theta_n}), \tag{10}$$

iff all coefficients vanish

$$f_{j_1}\cdots j_n = 0, \quad \text{and } j_n < 0,$$

$$f_{j_1}\cdots j_n = \frac{1}{(2\pi)^n}\int_0^{2\pi} d\theta_1 \cdots \int_0^{2\pi} d\theta_n\, f(\theta_1,\ldots,\theta_n) \times e^{-i(j_1\theta_1+\cdots+j_n\theta_n)} \quad (11)$$

in the multiple Fourier series for f:

$$f(\theta_1,\ldots,\theta_n) = \sum_{j_1=-\infty}^{\infty}\cdots\sum_{j_n=-\infty}^{\infty} f_{j_1\ldots j_n} e^{i(j_1\theta_1+\cdots+j_n\theta_n)}. \quad (12)$$

As a final example we consider the topological product of "cut planes" like the one in the earlier example. In that case a function analytic in the domain and vanishing in all directions around infinity is given by

$$\Gamma(Z_1,\ldots Z_n) = \frac{1}{(2\pi i)^n}\int_0^{\infty}\prod_{j=1}\frac{da_j}{a_j - Z_j}\,\phi(a_1,\ldots,a_n). \quad (13)$$

In that case the distinguished boundary is the topological product of positive real axes. As each one of them is "exposed" in the same sense as in connection with Eq. (8) we may deduce that any distribution ϕ (with the right support properties and such that the integral makes sense) is "allowed" and is related to Γ similarly as in Eq. (8). Unfortunately, the domain of analyticity found for the vertex function in KW is more complicated than the domains discussed so far. Nevertheless, certain features of these examples also appear in connection with the vertex function.

By an appropriate limit procedure [similar to the one in Eq. (8)] it is possible to express the form factors as limits of the representation formula for the vertex function. This then defines the form factors as functionals of the weight functions in the representation formula. On the other hand, these weight functions, being combinations of boundary values of the vertex function, can actually be expressed as functionals of the form factors by means of another limit procedure and another set of representation formulas (cf. Sec. 2A and 3).

In this paper we derive a set of conditions for the compatibility of the procedures above. It turns out that just as in Eqs. (6) and (11) the conditions can be expressed as integral equations, in this case for the form factors. In contrast to the cases above, however, there is only a finite number of such equations. It should be mentioned that the integral domains of the equations (the distinguished boundary of the domain) only contain physical values of the form factors. Inasmuch as the (freely performed) changes of orders of integrations and orders of limits and integrations are allowed, the conditions which are derived are both necessary and sufficient conditions.

The possibility of reformulating the general assumptions of field theory into integral equations of this kind is useful because:

(1) Only the physical quantities of the field theory, i.e., the matrix elements of the operators, enter into the equations,

(2) One of the major obstacles in order to carry through a program similar to KW for n-point functions with $n > 3$ is that the analyticity domains are rather complicated.[7,8] The distinguished boundaries of the domains are, however, possible to find and to discuss even if the whole domain might be frustratingly complicated to construct.

The integral equations do, of course, have a large set of solutions. All perturbation-theoretical vertex functions should, as a matter of fact, be included in this set because all of them fulfill the general assumptions in KW.

On the other hand, the equations can serve as a convenient starting point for dynamical approximations and model building. By the equations the, in general, nontrivial problem of incorporating the constraints of locality into a model of some form factors is solved. The integral equations are also straightforward generalizations of the results for the corresponding two-point function problem. In that case local commutativity implies "weak local commutativity," i.e., that two "spectral functions" should be equal. Here we will find that the difference between two form-factors (which generalize the spectral functions of the two-point function) shall vanish when integrated over certain physical sets in mass space.

In Sec. 2 we have gathered some earlier results, in order to make this paper reasonably self-contained. Thus in Sec. 2A some properties of the representation formulas from DI and DII are presented.[9] In Sec. 2B we define the connection between the vertex-functions and the matrix-elements of the field-operators. To that end, we make use of the well-known properties of the causal functions, i.e., the retarded and advanced functions.[7,4,10] In Sec. 3A we briefly discuss the appearance of the distinguished boundary of the analyticity domain of the vertex function. In particular, we show (with an explicit construction in Appendix B) that all points on the distinguished boundary are surrounded by a neighborhood contained in the domain of analyticity.[11] We may consequently deduce that the boundary points can be approached from inside along many different directions and that the corresponding limiting procedures nevertheless yield the same boundary value of the function.

In Sec. 3B we use this freedom to express the weight functions of the representation formulas in terms of the form factors. It is shown that the procedure is unambiguous unless the field operators of the theory have nonvanishing matrix elements between the vacuum state and states with vanishing mass.

In Sec. 4 we compute by an appropriate limit procedure (according to the results in Sec. 2B) the form factors as functionals of the weight functions.

In Sec. 5 the compatibility between the results of Sec. 3 and Sec. 4 is investigated and the integral equations are derived.

In Sec. 6 some extensions and further conclusions are presented.

Some of the details of the limiting procedures in Sec. 4 and 5 are gathered in Appendices C and D.

2. SURVEY OF EARLIER RESULTS

A. Representation formulas for the absorptive part of the vertex function

The analyticity properties which are proved in KW and are mentioned in the Introduction can be explicitly seen from the following representation-formula which is derived in DII:

$$\Delta_3 G(Z_1, Z_2, b_3) = \frac{1}{(2\pi i)^2} \int\int db_1 db_2 (K^{\mathrm{I}}(Z_1, Z_2; b_3; b_1, b_2) \times \sigma_{\mathrm{I}}(\Delta_3 G)(b_1, b_2, b_3) + K^{\mathrm{II}}(Z_1, Z_2; b_3; b_1, b_3)\sigma_{\mathrm{II}}(\Delta_3 G)(b_1, b_2, b_3). \quad (14)$$

The kernel functions K^{I} and K^{II} in Eq. (14) are given by the formulas

$$K^{\mathrm{I}}(Z_1, Z_2; b_3; b_1, b_2) = \int_0^\infty dr\ \delta(rb_3 + (r - b_1)(r - b_2)) \times \left[\frac{\theta(b_1 - r)\theta(r - b_2)}{b_1 - Z_1} \left(1 + \tfrac{1}{2}(b_1 + Z_1 - 2r) \times \frac{2r + b_3 - Z_1 - Z_2}{(r - Z_1)(r - Z_2) + rb_3} \right) + \frac{\theta(b_2 - r)\theta(r - br)}{b_2 - Z_2} \times \left(1 + \tfrac{1}{2}(b_2 + Z_2 - 2r) \frac{2r + b_3 - Z_1 - Z_2}{(r - Z_1)(r - Z_2) + rb_3} \right) \right], \quad (15)$$

$$K^{\mathrm{II}}(Z_1, Z_2; b_3; b_1, b_2) = \int_0^1 \int_0^1 d\alpha d\beta\ \delta(1 - \alpha - \beta)\delta(\alpha b_2 + \beta b_1 - \alpha\beta b_3) \times \left(\frac{[\theta(b_1)\theta(-b_2) + \tfrac{1}{2}\theta(b_1)\theta(b_2)]}{b_1 - Z_1} \frac{Z_1 - \alpha^2 b_3}{\alpha\beta b_3 - \alpha Z_2 - \beta Z_1} + \frac{[\theta(-b_1)\theta(b_2) + \tfrac{1}{2}\theta(b_1)\theta(b_2)]}{b_2 - Z_2} \frac{Z_2 - \beta^2 b_3}{\alpha\beta b_3 - \alpha Z_2 - \beta Z_1} \right). \quad (16)$$

The weight functions in the representation formula are combinations of boundary values on the real axes of $\Delta_3 G$ defined by

$$\begin{aligned} \sigma_{\mathrm{I}}(\Delta_3 G) &= \Delta_3 G(b_1 + i\epsilon', b_2 + i\epsilon'', b_3) + \Delta_3 G(b_1 - i\epsilon', b_2 - i\epsilon'', b_3), \\ \sigma_{\mathrm{II}}(\Delta_3 G) &= \Delta_3 G(b_1 + i\epsilon', b_2 - i\epsilon'', b_3) + \Delta_3 G(b_1 - i\epsilon', b_2 + i\epsilon'', b_3). \end{aligned} \quad (17)$$

We note especially that both the kernel functions K^{I} and K^{II} are explicitly *symmetric* by the exchange of the indices 1 and 2. It is further evident that the integrals used in the definitions in Eqs. (15) and (16) are only formal, because of the occurrence of the δ functions Concerning *the singularity structure* of the kernel functions, it should be noted that there are singularities along the real positive axes of both the Z variables ("the Z_1 and Z_2 cuts," respectively) with a similar interpretation as in connection with the Z_3 cut above. There are, however, also singularities along two very different surfaces, known as "anomalous cuts", i.e.,

$$\text{AC I:}\quad (r - Z_1)(r - Z_2) + rb_3 = 0, \quad r > 0, \quad (18)$$

$$\text{AC II:}\quad \alpha\beta b_3 - \alpha Z_2 - \beta Z_1 = 0, \quad \alpha > 0, \beta > 0, \alpha + \beta = 1. \quad (19)$$

Both these singularities occur in KW and also in perturbation-theoretical examples and are connected to the singularities of the so-called triangle graph. As a matter of fact, both kernel functions can be considered as absorptive parts in the sense of Eq. (2) for particular perturbation-theory functions.

As an example consider the perturbation-theory function F corresponding to the graph of Fig. 1. By explicit calculation one finds[12] [in the notation of Eq. (1) and Fig. 1] for the derivative $\partial F/\partial r$

$$\frac{\partial F}{\partial r}(Z_1, Z_2, Z_3; b_1, r) = (\text{const}) \frac{1}{b_1 - Z_1} \frac{1}{(r - Z_1)(r - Z_2) + rZ_3} \times [\log(-rZ_3) - \log(r - Z_1) - \log(r - Z_2)]. \quad (20)$$

By comparison with Eq. (15) it is easily seen that the first term in the brackets $[\cdots]$ equals the absorptive part of $\partial F/\partial r$ for a particular value of the "internal mass" variable $r = r(b_1, b_2, b_3)$ according to the δ functions (except for a polynomial). There are similar connections between the second term in Eq. (15) and the graph obtained by the interchange of the indices 1 and 2 in Fig. 1. The terms in the expression for the kernel function K^{II} are in the same way connected to the graphs obtained from Fig. 1 and the above-mentioned one by exchanging the internal mass variables 0 and r (putting $r = \alpha b_3$ or $r = \beta b_3$, respectively).

We further note that the "weight functions" $\sigma(\Delta_3 G)$ multiplying the kernel functions K^{I} and K^{II} are particular combinations of boundary values of $\Delta_3 G$ on the real axes of the variables. In particular one needs the boundary values of $\Delta_3 G$ for such values of the variables b_1, b_2, and b_3 that the δ functions in the integrals of Eq. (15) or (16) can be fulfilled. The *support properties* of the kernel functions in this way imply different restrictions on the integration variables, the "mass squares" b_j.

For the subsequent discussion in this paper, it is useful to introduce the following six real domains:

$$\left.\begin{aligned} T_j(b_1, b_2, b_3):&\quad b_j < 0, b_k > 0, b_l > 0 \\ D_j(b_1, b_2, b_3):&\quad b_j > (\sqrt{b_k} + \sqrt{b_l})^2, b_k > 0, b_l > 0 \end{aligned}\right\}, \quad j \neq k \neq l \neq j. \quad (21)$$

It is easily seen that the domains are nonoverlapping and further that inside these domains there are real vectors p_j in Lorentz space such that [cf. Eq. (1)]

$$\sum_{j=1}^{3} p_j = 0, \quad -p_j^2 = b_j, \quad j = 1, 2, 3. \quad (22)$$

Inside, e.g., the domain D_1 the vectors p_2 and p_3 are timelike and belong to the same light cone, etc.

In terms of the domains T and D the support properties

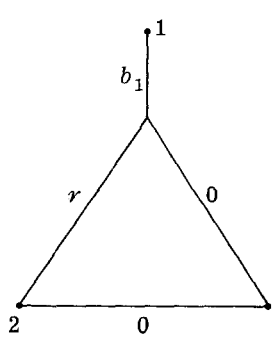

FIG. 1. The perturbation theory graph related to the function in Eq. (20).

of the kernel functions are the following:

The first term in Eq. (15) for the kernel function K^{I} is nonvanishing only inside T_2 and D_1, while the second term (found by interchanging the indices 1 and 2) is nonvanishing inside T_1 and D_2.

The first term in the kernel function K^{II} is nonvanishing only inside T_2 and D_3 while the second term is nonvanishing inside T_1 and D_3.

We note in particular that the first and second term in K^{II} has overlapping support in the domain D_3. The contribution to $\Delta_3 G$ has for symmetry reasons been attributed with one half to each term.

In the derivation of the representation formula it is explicitly assumed that the boundary values $\sigma(\Delta_3 G)$ are independent of the integration parameters r, α, and β. This is indicated in Eq. (14) by the fact that these integrals have been "extracted." It will subsequently turn out that this assumption is very important. The basis behind the assumption is investigated in Sec. 3 and Appendix B. It is shown that there is a unique representation of the boundary values $\sigma(\Delta_3 G)$ in terms of limits of the causal functions, and we may consequently deduce that the assumption is fulfilled. For particular values of the arguments (Z_1, Z_2, b_3) there is another and sometimes more convenient way of representing the absorptive part $\Delta_3 G$. Thus, the following formula, which is proved in DI, is valid if (Z_1, Z_2, b_3) is on ACI:

$$\Delta_3 G_r(Z_1, Z_2, b_3) = 2\pi i \iint da_1 da_2\ \delta((r-a_1)(r-a_2) + rb_3) \times \left(G_B(a_1, a_2, b_3)\theta(a_1 - r)\frac{1}{2}\frac{a_1 + Z_1 - 2r}{a_1 - Z_1} + G_A(a_1, a_2, b_3)\theta(a_2 - r)\frac{1}{2}\frac{a_2 + Z_2 - 2r}{a_2 - Z_2}\right). \quad (23)$$

In the same way the following formula is valid for (Z_1, Z_2, b_3) on ACII:

$$\Delta_3 G_{\alpha\beta}(Z_1, Z_2, b_3) = 2\pi i \iint da_1 da_2\ \delta(\alpha a_2 + \beta a_1 - \alpha\beta b_3) \times \left(G_B(a_1, a_2, b_3)\theta(a_1)\frac{\alpha}{a_1 - Z_1} + G_A(a_1, a_2, b_3)\theta(a_2)\frac{\beta}{a_2 - Z_2}\right). \quad (24)$$

The quantities r [in Eq. (23)] and α, β [in Eq. (24)] are the real positive numbers which parametrize the anomalous cut surfaces in Eqs. (18) and (19), respectively. They should then be considered as being defined by Eqs. (18) and (19) as functions of (Z_1, Z_2, b_3). The conditions that the equations have solutions with the "right" reality and positivity properties are then conditions on the coordinates in order that Eqs. (23) and (24) be valid [i.e., that (Z_1, Z_2, b_3) be on ACI or ACII, respectively]. We note in passing that outside the positive real axes, i.e., unless both Z_1 and Z_2 are on the (normal) cut surfaces, there is at most *one* solution for r, α, and β with the required properties. This is so despite the apparent quadratic nature of the equations. The quantities G_A and G_B occurring as weight functions in Eqs. (23) and (24) are the form factors and are defined precisely by Eqs. (36) and (37) below. It is obvious that the quantities $\Delta_3 G_r$ and $\Delta_3 G_{\alpha\beta}$ as defined by Eqs. (23) and (24) depend explicitly upon r, α, and β. The formulas *cannot* be used to continue $\Delta_3 G$ outside the surfaces ACI and ACII. We have indicated the limitations by the extra indices r and α, β.

The results are reached by a limiting procedure, when a Cauchy contour inside the domain of analyticity is made to approach the anomalous cut at the same time as the third variable of the vertex function approaches the (normal) cut (i.e., the Z_3 cut) (cf. DI)

We finally note that there is a certain *assumed asymptotic behavior* of the absorptive part $\Delta_3 G$ behind the derivation of Eqs. (14), (23), (24). Thus the function $\Delta_3 G$ is assumed to vanish in all direction around infinity inside the analyticity domain. The modifications, which are necessary in case this is not fulfilled (but $\Delta_3 G$ according to the assumptions of Sec. 2A is at most polynominally growing), have been gathered in Appendix A ("subtracted dispersion relations").

The assumption that $\Delta_3 G$ vanishes when one of the arguments approaches infinity "along the anomalous cut" ACI is actually already incorporated in Eq. (23). Thus the following superconvergence relation or sum rule may be derived from Eq. (23) in the limit $Z_2 \to \infty, Z_1 \to r$ (or $Z_1 \to \infty, Z_2 \to r$) (cf. SII and Ref. 7 for a direct proof of this relation):

$$\iint da_1 da_2\ \delta((r-a_1)(r-a_2) + rb_3)[G_B(a_1, a_2, b_3)\theta(a_1 - r) - G_A(a_1, a_2, b_3)\theta(a_2 - r)] = 0. \quad (25)$$

According to the results in SII the same relation is also valid when G_B and G_A are replaced by the weight functions $\sigma_{\mathrm{I}}(\Delta_3 G)$ above.

B. The vertex function and the operator matrix elements

The vertex function G has so far been considered as an abstract entity equipped with the analyticity properties proved in KW. We wish in this section briefly to touch upon the connections between G and the matrix elements of an underlying field theory. We will in particular assume the existence of three scalar fields A, B, C which are local and interact in such a way that the Wightman axioms are fulfilled. For the physical interpretation it is useful to assume asymptotic properties admitting LSZ[10] or reduction formalism[7] though no explicit use is made of this formalism. We will, however, just as in KW, make extensive use of the causal (retarded and advanced) functions and their well-known representation formulas.

We will use the notations of SII and define, e.g., the retarded function R_A by the Fourier integral:

$$R_A(\tilde{p}_2, \tilde{p}_3) = -\int dx_2 dx_3\ \exp[i\tilde{p}_2(x_2 - x_1) + i\tilde{p}_3(x_3 - x_1)] \times r_A(x_1 - x_2; x_1 - x_3). \quad (26)$$

The weight function r_A in the Fourier transform is known as the vacuum expectation value of the retarded commutator.[7,10] We will only need the following properties (for explicit formulas, cf., e.g., Appendix B, SII):

(i) The retarded commutator is a sum of operator products of A, B, and C multiplied in different orders and with step functions in time.

(ii) Due to the assumed locality of the field theory, r_A is essentially[13] a Lorentz scalar and vanishes unless the field points of B and C are retarded ("before") the field point of the field A.

Due to the support properties of r_A, Eq. (26) exhibits the retarded function R_A as an analytic function of the vectors p_2 and p_3 when the imaginary parts of the vectors belong to the forward light cones (V^+). Due to the

Lorentz covariance of the theory the function R_A only depends upon the Lorentz scalar variables, i.e., the variables of Eq. (1). We now *define* the relation between the field operators and the vertex function G by

$$R_A = G. \tag{27}$$

Then it is possible to represent G inside the analyticity domain of R_A by Eq. (26).

It is also possible to define[7,10] retarded functions R_B and R_C in which the fields B and C are distinguished just like the field A in Eq. (26). Further we may define the advanced functions A_A, A_B, A_C essentially be reversing the sign of the arguments in the step functions for the respective Fourier weight functions. All these functions exhibit analyticity properties and coincide in a well-known way in a common region.[3,7,4] By the uniqueness of analytic continuation[6] they are then *all* equal to the vertex function G and can be used as representation formulas for it inside their regions of analyticity.

We will now make use of the different retarded and advanced functions to express the absorptive part $\Delta_3 G(Z_1, Z_2, b_3)$ in terms of the matrix elements of the operators. To that end we note that the energy–momentum vector p_3, corresponding to the mass square b_3 according to Eq. (1) and above, will because of the positivity of b_3 be a timelike vector. We will for definiteness choose it to belong to the forward light-cone remembering that due to the inherent CPT symmetry in a field theory fulfilling the Wightman axioms[4] this choice implies no loss of generality. The different limiting situations in Eq. (2) can then be achieved by choosing the limiting imaginary part of the vector p_3 to approach the origin inside a definite lightcone, i.e.,

$$b_3 \pm i\epsilon = -(p_3 + ik_3)^2, \qquad k_3 \in V^{\pm} \to 0. \tag{28}$$

Then, for the case when the imaginary part of the vector p_2 belongs to V^+ [note that, due to Eq. (1), $\mathrm{Im} p_1 = k_1 = -k_2 - k_3 \to -k_2 \in V^-$], the absorptive part $\Delta_3 G$ is given by

$$\Delta_3 G = R_A(\tilde{p}_2, p_3) - A_B(\tilde{p}_1, p_3). \tag{29}$$

Insertion of the representation formulas for R_A and A_B like the one in Eq. (26) and use of the explicit expressions for r_A and a_B then results in the following integral (cf. SII):

$$\int d^2x \exp\left(i\sum_{j=1}^{3} \tilde{p}_j x_j\right)\langle 0|\theta(12) \times [C(x_3), [A(x_1), B(x_2)]]|0\rangle. \tag{30}$$

We have used the shorthand notation $\theta(12) = \theta(x_1 - x_2)$. The step function with a vector argument means that the vector belongs to V^+. We note that the argument of the field C *does not* occur inside the step functions, and we can consequently perform the integral over x_3 by straightforward means. Insertion of a complete set of states $|n\rangle$ with energy–momentum vectors p_n and use of translation invariance[3,2,4] results in

$$\Delta_3 G|_{k_2 \in V^+} = (2\pi)^4 \int dx \exp(i\tilde{p}_1 x_1 + i\tilde{p}_2 x_2) \times \sum_{|n\rangle} \delta(p_3 - p_n)\langle 0|\theta(12)[B(x_2)A(x_1)]|n\rangle\langle n|C|0\rangle. \tag{31}$$

We have in Eq. (31) neglected one term containing a δ function with the vector argument $(p_3 + p_n)$. Due to the spectrum assumptions and our assumption above that $p_3 \in V^+$ this vector can never vanish and the term will consequently give no contribution.

In that way the absorptive part $\Delta_3 G$ is given by the matrix element of a retarded commutator between the vacuum state $|0\rangle$ and a state $|n\rangle$ with definite energy–momentum vector and the quantum numbers of the field $C(\langle n|C|0\rangle \neq 0)$.

There is a similar formula valid when the imaginary part of p_2 belongs to V^-:

$$\Delta_3 G|_{k_2 \in V^-} = (2\pi)^4 \int dx \exp(i\tilde{p}_1 x_1 + i\tilde{p}_2 x_2) \times \sum_{|n\rangle} \delta(p_3 - p_n)\langle 0|\theta(21)[A(x_1), B(x_2)]|n\rangle\langle n|C|0\rangle. \tag{32}$$

We note in particular that the limit

$$\mathfrak{L}(\Delta_3 G) = \lim_{k_2 \to 0}[\Delta_3 G|_{k_2 \in V^+} - \Delta_3 G|_{k_2 \in V^-}] \tag{33}$$

can be expressed as

$$\mathfrak{L}(\Delta_3 G) = (2\pi)^4 \int dx_2 \exp(ip_1 x_1 + ip_2 x_2) \times \sum_{|n\rangle} \delta(p_3 - p_n)\langle 0|[B(x_2), A(x_1)]|n\rangle\langle n|C|0\rangle. \tag{34}$$

The quantity $\mathfrak{L}(\Delta_3 G)$ is thus the Fourier transform of a matrix element which does *not* contain any step functions in time.

Introducing another complete set of states $|m\rangle$ with energy–momentum vectors p_m, we can perform the integral in Eq. (34) in the same way as in connection with Eq. (31). The result is

$$\mathfrak{L}(\Delta_3 G)(p_1, p_2, p_3) = (2\pi)^2 \theta(p_3)[\theta(-p_2)G_A(-p_1^2 - p_2^2 - p_3^2 - \theta(-p_1)G_B(-p_1^2 - p_2^2 - p_3^2)]. \tag{35}$$

The quantities G_A and G_B, introduced in Eq. (35), are given by

$$G_A(-(p_2 + p_3)^2, -p_2^2, -p_3^2) = (2\pi)^6 \sum_{|n\rangle, |m\rangle} [\theta(p_3)\theta(-p_2)\delta(p_3 - p_n)\delta(p_2 + p_m) \times \langle 0|B|m\rangle\langle m|A|n\rangle\langle n|C|0\rangle + \theta(-p_3)\theta(p_2)\delta(p_3 + p_n)\delta(p_2 - p_m) \times \langle 0|C|n\rangle\langle n|A|m\rangle\langle m|B|0\rangle, \tag{36}$$

$$G_B(-p_1^2, -(p_1 + p_3)^2, -p_3^2) = (2\pi)^6 \sum_{|n\rangle, |m\rangle} [\theta(p_3)\theta(-p_1)\delta(p_3 - p_n)\delta(p_1 + p_m) \times \langle 0|A|m\rangle\langle m|B|n\rangle\langle n|C|0\rangle + \theta(-p_3)\theta(p_1)\delta(p_3 + p_n)\delta(p_1 - p_m) \times \langle 0|C|n\rangle\langle n|B|m\rangle\langle m|A|0\rangle]. \tag{37}$$

The definitions of G_A and G_B in Eqs. (36) and (37) have been chosen in an explicitly CPT symmetric way. These quantities will in the following be referred to as "form factors" of the fields A and B, respectively.

It can be seen from the definition that the form factor G_A is essentially the matrix element of the operator A between states with the quantum numbers of the operators B and C and with well-defined energy–momentum

vectors. The notation is chosen so that the vectors p_1, p_2, and p_3 (with mass squares b_1, b_2, and b_3, respectively) fulfilling energy–momentum conservation according to Eq. (22) always refer to states with the quantum numbers of the fields A, B, and C, respectively.

The support-properties of the form factors mirrors the mass spectrum of the theory. We further note that the form factor $G_B(b_1, b_2, b_3)$ is nonvanishing only inside the regions $D_1(b_1, b_2, b_3)$, $D_3(b_1, b_2, b_3)$, and $T_2(b_1, b_2, b_3)$ in the notations of Sec. 2A. In the same way the support of $G_A(b_1, b_2, b_3)$ is contained in the regions $D_2(b_1, b_2, b_3)$ $D_3(b_1, b_2, b_3)$, and $T_1(b_1, b_2, b_3)$. We note especially that in the common region of support $D_3(b_1, b_2, b_3)$ only the difference $[G_B(b_1, b_2, b_3) - G_A(b_1, b_2, b_3)]$ can be determined from $\mathfrak{L}(\Delta_3 G)$ (cf. Sec. 6).

3. THE BOUNDARY VALUES OF THE ABSORPTIVE PART IN TERMS OF THE FORM FACTORS

A. On the approach to the boundary regions

In this section we will discuss the appearance of the domain of analyticity for the vertex function in the neighborhood of the singularity surfaces. We will in particular employ the connection between the absorptive part $\Delta_3 G$ of the vertex function and the retarded and advanced functions from Eqs. (31) and (32). Our goal is to show[11] that each point in the integration region in Eq. (14) is surrounded by a complex neighborhood belonging to the domain of analyticity. It is then possible to reach the boundary points "from inside" along many different and equivalent directions.

We will define "the directions of approach" to be the ratio (ϵ'/ϵ'') between the limiting imaginary parts of the arguments in the boundary value $\Delta_3 G(b_1 + i\epsilon', b_2 + i\epsilon'', b_3)$. By the same methods that are employed in Sec. 2C (cf. the similar explicit constructions in SI, SII, and DI) we can express this boundary value in terms of boundary values of the retarded and advanced functions. The limit when ϵ' and ϵ'' approaches zero corresponds in that way to the limit when the imaginary part k_2 approaches the zero vector inside a definite light cone in accordance with Eqs. (31) and (32). The ratio (ϵ'/ϵ'') is then connected to the coordinates of the boundary point, i.e., (b_1, b_2, b_3), and the limiting direction of the vector k_2. Equivalence between two directions of approach means that there are vectors k_2 belonging to a definite light cone so that both the ratios can be attained. Some details of the explicit constructions have been gathered in Appendix B. We have in particular investigated the case of two-dimensional space–time. Conditions which are sufficient for that case are certainly also sufficient for four-dimensional space–time, where the "freedom of approach" is correspondingly larger. The resulting sufficient conditions on the ratio ϵ'/ϵ'' such that the boundary values $\sigma(\Delta_3 G)$ should be expressible in terms of advanced and retarded functions are the following:

Inside the regions $D_1(b_1, b_2, b_3)$ and $D_2(b_1, b_2, b_3)$ (which according to the investigation in the end of Sec. 2B are connected to the anomalous cut ACI) the ratio ϵ'/ϵ'' must be a nonzero, finite, and positive number. Inside the region $D_3(b_1, b_2, b_3)$ (in the same way connected to the anomalous cut ACII) the ratio ϵ'/ϵ'' must be a nonzero, finite, and negative number. Inside the regions T_1 and T_2 there are no restrictions on the *ratio* ϵ'/ϵ''. Inside, e.g., the region $T_1(b_1, b_2, b_3)$ it is, however, necessary that ϵ'' (the limiting imaginary part of Z_2) has a definite sign. In the same way inside $T_2(b_1, b_2, b_3)$ the sign of ϵ' is restricted.

The difference between the regions T and D is to be expected because in the D regions both b_1 and b_2 are positive real numbers, while in the T regions one of them is negative. The difference between the boundary value above and below the real axis is only of interest in case there are singularities on the axis. This is the case along the (normal) cut surfaces, i.e., only along the positive real axes.

We may then conclude that, subject to the restrictions above, there is a great freedom of choice in defining the limiting procedure for $\sigma(\Delta_3 G)$ in Eq. (17). The resulting boundary values are in "the allowed cases" unique in the same sense as the limits of the retarded and advanced functions are unique, i.e., in a distribution-theoretical sense.

B. The boundary values as functionals of the form factors

In this section we will employ the results of Sec. 3A, i.e., the fact that the boundary values which occur in Eq. (17) can be equivalently attained along any one of many different directions. We will in particular use this freedom to compute the boundary value of $\Delta_3 G$ on the intersection of the real axes and the anomalous cuts as a boundary value "along" the anomalous cut surface. Thus any boundary point which occurs in the integration range of Eq. (14) [for $\sigma_I(\Delta_3 G)$] can be reached "along" the anomalous cut surface in the following way.

Consider an arbitrary but fixed point (b_1, b_2, b_3) such that,

b_j real,

$$r_0 = \tfrac{1}{2}(b_1 + b_2 - b_3 + \sqrt{\lambda(b_1, b_2, b_3)}) > 0,$$
$$\lambda(b_1, b_2, b_3) = b_1^2 + b_2^2 + b_3^2 - 2b_1 b_2 - 2b_1 b_3 - 2b_2 b_3. \tag{38}$$

This condition is sufficient that the point (b_1, b_2, b_3) is on the intersection of the real axes and ACI with the parameter $r = r_0$.

This point can be reached from any complex point (Z_1, Z_2, b_3) with

$$Z_2 = r_0 + r_0 b_3/(r_0 - Z_1) \tag{39}$$

[i.e., (Z_1, Z_2, b_3) is on ACI] by the prescription $Z_1 \to b_1$.

In that way a particular boundary value $\Delta_3 G(b_1 + i\epsilon', b_2 + i\epsilon'', b_3)$ can be defined by

$$\lim_{Z_1 \to b_1} \Delta_3 G(Z_1, r_0 + r_0 b_3/(r_0 - Z_1), b_3) = \lim_{Z_1 \to b_1} \Delta_3 G_{r_0}(Z_1, Z_2, b_3). \tag{40}$$

[The notation $\Delta_3 G_{r_0}$ is introduced in Eq. (23).]

Note that the limiting imaginary parts ϵ' and ϵ'' have the same sign and in particular that

$$\epsilon'/\epsilon'' = (r_0 - b_1)^2/r_0 b_3. \tag{41}$$

According to the result of Sec. 3A all "directions of approach" ϵ'/ϵ'' are equivalent at least if ϵ'/ϵ'' is finite and nonzero. Consequently the limit according to Eq. (40) is well defined except possibly when

(i) $r_0 = b_1$ or $b_1 = \infty$,

(ii) $b_3 = 0$,

(iii) $r = 0$.

The first solution corresponds to asymptotic values of one of the arguments b_1 or b_2. The assumed boundedness properties of $\Delta_3 G$ excludes a singular contribution to Eq. (14) in that case.

The second situation may occur in case there are states in the theory with the quantum numbers of the C-field and vanishing mass [cf. Eq. (31)]. The third situation may correspond to a more singular situation and Eq. (23) has no meaning in that case. It is seen, however, that the point $r = 0$ only contributes to a low-dimensional part of the actual integral in Eq. (14). There are eventual difficulties only if $\Delta_3 G(b_1 + i\epsilon', b_2 + i\epsilon'', b_3)$ is singular at one of the points $b_1 = 0$ or $b_2 = 0$. In case this singularity is of integrable type in the mass variable, it is possible to define the limiting situation in Eq. (40) and (iii) by

$$\lim_{r_0 \to 0} \lim_{Z_1 \to b_1} \Delta_3 G_{r_0}(Z_1, Z_2, b_3). \tag{42}$$

By looking back upon the original derivation of the representation formula in DII, it is clear that this interpretation is the correct one.

If the singularity at the origin in mass space is of non-integrable type (pole type etc.) we must interpret the difficulty as connected to zero mass states with the quantum numbers of the A- and C-field (e.g., "infrared divergences"). For such a situation it is necessary to modify the representation formulas and explicitly exhibit such contributions. For the remainder the prescription of Eq. (42) is valid (cf Sec. 6).

There are similar difficulties associated with a limit along the surface ACII for the quantity $\sigma_{II}(\Delta_3 G)$. Thus Eq. (24) has no meaning in case $\alpha = 0$ or $\beta = 0$ and modifications similar to the case (iii) above may be necessary.

It can be seen, however, that the possible difficulties in using Eqs. (23) and (24) to compute the boundary values $\sigma_I(\Delta_3 G)$ and $\sigma_{II}(\Delta_3 G)$ are only connected to the eventual appearance of zero-mass states in the theory.

Neglecting such possibilities we may write for the boundary values $\sigma_I(\Delta_3 G)$ and $\sigma_{II}(\Delta_3 G)$ in Eqs. (17):

$$\begin{aligned}\sigma_I(\Delta_3 G_r) \equiv \sigma_{Ir}(b_1, b_2, b_3) &= 4\pi i \iint da_1 da_2 \\ &\times \delta((r - a_1)(r - a_2) + rb_3)\Big(G_B(a_1, a_2, b_3) \\ &\times \theta(a_1 - r)\frac{\frac{1}{2}(a_1 + b_1 - 2n)}{(a_1 - b_1)_p} \\ &+ G_A(a_1, a_2, b_3)\theta(a_2 - r)\frac{\frac{1}{2}(a_2 + b_2 - 2r)}{(a_2 - b_2)_p}\Big) \\ &= 4\pi i \iint da_1 da_2\ \delta((r - a_1)(r - a_2) + rb_3)\frac{|r - a_1|}{(a_1 - b_1)_p} \\ &\times [G_B(a_1, a_2, a_3)\theta(a_1 - r) + G_A(a_1, a_2, a_3)\theta(r - a_1)],\end{aligned} \tag{43}$$

$$\begin{aligned}\sigma_{II}(\Delta_3 G_{\alpha\beta}) \equiv \sigma_{II\alpha}(b_1, b_2, b_3) &= 4\pi i \iint da_1 da_2 \\ &\times \delta(\alpha a_2 + \beta a_1 - \alpha\beta b_3)\Big(G_B(a_1, a_2, b_3) \\ &\times \theta(a_1)\frac{\alpha}{(a_1 - b_1)_p} + G_A(a_1, a_2, b_3)\theta(a_2)\frac{\beta}{(a_2 - b_2)_p}\Big) \\ &= 4\pi i \iint da_1 da_2\ \delta(\alpha a_2 + \beta a_1 - \alpha\beta b_3)\frac{\alpha}{(a_1 - b_1)_p} \\ &\times [G_B(a_1, a_2, b_3) - G_A(a_1, a_2, b_3)].\end{aligned} \tag{44}$$

In the second line of Eq. (43) we have made use of the superconvergence relation in Eq. (25) and the fact that both (a_1, a_2, b_3) and (b_1, b_2, b_3) are on ACI with the same parameter r.

We note that the actual integration region for the arguments (a_1, a_2, b_3) of the form factors G_A and G_B in Eq. (43) are for the first term the regions $D_1(a_1, a_2, b_3)$ and $T_1(a_1, a_2, b_3)$ and for the second term $D_2(a_1, a_2, b_3)$ and $T_2(a_1, a_2, b_3)$. Thus the common region of support $D_3(a_1, a_2, b_3)$ (cf. Sec. 2C and Sec. 6) does not occur in Eq. (43). In the second line of Eq. (44) we have made use of the fact that both (a_1, a_2, b_3) and (b_1, b_2, b_3) are on ACII with the same parameters α and β. The actual integration region in Eq. (44) is $T_1(a_1, a_2, b_3)$ (for G_A) and $T_2(a_1, a_2, b_3)$ (for G_B) while in the common region, $D_3(a_1, a_2, b_3)$ only the difference $(G_B - G_A)$ occurs.

4. THE FORM FACTORS AS LIMITS OF THE REPRESENTATION FORMULA FOR THE ABSORPTIVE PART

In section 2B the form factors G_A and G_B are defined as particular limits of the absorptive part $\Delta_3 G$ [cf. Eqs. (33)–(35)].

In this section the corresponding limits will be computed from the representation formula for $\Delta_3 G$ in Eq. (14). In that way the form factors G_A and G_B will be expressed as functionals of the boundary values $\sigma_I(\Delta_3 G)$ and $\sigma_{II}(\Delta_3 G)$.

The procedure should be contrasted to the results of the foregoing section in which the boundary values $\sigma(\Delta_3 G)$ are expressed in terms of the form factors. The compatibility of the results is investigated in Sec. 5.

Some details of the limiting procedure have been gathered in Appendix C. The results for the quantities $\mathcal{L}(K^I)$ and $\mathcal{L}(K^{II})$ defined by Eq. (33) are:

$$\begin{aligned}\mathcal{L}(K^I) &= \iint dc_1 dc_2\ \delta(c_1 + p_1^2)\delta(c_2 + p_2^2)(l_1^I + l_2^I), \\ \mathcal{L}(K^{II}) &= \iint dc_1 dc_2\ \delta(c_1 + p_1^2)\delta(c_2 + p_2^2)(l_1^{II} + l_2^{II}).\end{aligned} \tag{45}$$

The quantities l in Eq. (45) are defined by

$$\begin{aligned}l_1^I &= 2\pi i \int dr\ \theta(r)(b_1 + b_2 - b_3 - 2r)\delta((r - b_1)(r - b_2) + rb_3) \\ &\times \left(\theta(b_1 - r)\epsilon(-p_1)\frac{\delta(b_1 - c_1)}{(b_2 - c_2)_p} + \theta(b_2 - r)\epsilon(p_2)\frac{\delta(b_2 - c_2)}{(b_1 - c_1)_p}\right),\end{aligned} \tag{46}$$

$$l_2^I = 2\pi i \int dr\ \theta(r)\delta((r - b_1)(r - b_2) + rb_3)\delta((r - c_1) \times (r - c_2) + rb_3)\,\epsilon(c_1 - c_2)\sqrt{\lambda(b_3, c_1, c_2)}\,\Lambda_2^I \tag{47}$$

$$\begin{aligned}\Lambda_2^I &= \theta(b_1 - r)\frac{1}{2}\frac{(b_1 + c_1 - 2r)}{(b_1 - c_1)_p} + \theta(b_2 - r)\frac{\frac{1}{2}(b_2 + c_2 - 2r)}{(b_2 - c_2)_p} \\ &\cong |b_1 - r|\left(\frac{1}{(b_1 - c_1)_p} - \frac{1}{2}\frac{1}{b_1 - r}\right) \\ &\cong |b_2 - r|\left(\frac{1}{(b_2 - c_2)_p} - \frac{1}{2}\frac{1}{b_2 - r}\right),\end{aligned}$$

$$l_1^{II} = 2\pi i\int\!\!\int d\alpha d\beta\ \delta(1-\alpha-\beta)\theta(\alpha)\theta(\beta)\delta(\alpha b_2+\beta b_1-\alpha\beta b_3)$$
$$\times\Big[[\theta(b_1)\theta(-b_2)+\tfrac{1}{2}\theta(b_1)\theta(b_2)]\epsilon(-p_1)$$
$$\times\Big(\frac{b_1}{\alpha}-\alpha b_3\Big)\frac{\delta(b_1-c_1)}{(b_2-c_2)_p}+[\theta(-b_1)\theta(b_2)$$
$$+\tfrac{1}{2}\theta(b_1)\theta(b_2)]\epsilon(p_2)\Big(\frac{b_2}{\beta}-\beta b_3\Big)\frac{\delta(b_2-c_2)}{(b_1-c_1)_p}\Big], \qquad (48)$$

$$l_2^{II} = 2\pi i\int\!\!\int d\alpha d\beta\ \delta(1-\alpha-\beta)\theta(\alpha)\theta(\beta)\delta(\alpha b_2+\beta b_1-\alpha\beta b_3)$$
$$\times\delta(\beta c_1+\alpha c_2-\alpha\beta b_3)\sqrt{\lambda(b_3,c_1,c_2)}\,\Lambda_2^{II},$$
$$\Lambda_2^{II} = [\theta(b_1)\theta(-b_2)+\tfrac{1}{2}\theta(b_1)\theta(b_2)]\frac{\alpha}{(b_1-c_1)_p}$$
$$-[\theta(-b_1)\theta(b_2)+\tfrac{1}{2}\theta(b_1)\theta(b_2)]\frac{\beta}{(b_2-c_2)_p}$$
$$\cong\frac{\alpha}{(b_1-c_1)_p}\cong\frac{\beta}{(b_2-c_2)_p}. \qquad (49)$$

Several different expressions have been given for the quantities Λ_2 in Eqs. (47) and (49). The equivalance sign $\cong$ between the expressions for, e.g., $\Lambda_2{}^{I}$ means that the expressions coincide if [as the δ functions multiplying $\Lambda_2{}^{I}$ in Eq. (47) imply] the points (b_1,b_2,b_3) and (c_1,c_2,b_3) are both on the anomalous cut ACI with the same parameter r. The same goes for the equivalence sign for $\Lambda_2{}^{II}$ for points on ACII. These results should be compared to the second lines of Eqs. (43) and (44).

With the definitions above it is possible to write for the limit

$$\mathfrak{L}(\Delta_3 G) = (2\pi i)^2\int\!\!\int dc_1 dc_2\ \delta(p_1^2+c_1)\delta(p_2^2+c_2)$$
$$\times\int\!\!\int db_1 db_2\{[l_1^{I}(c_1,c_2;b_3;b_1,b_2)$$
$$+l_2^{I}(c_1,c_2;b_3;b_1,b_2)]\,\sigma_I(\Delta_3 G)(b_1,b_2,b_3)$$
$$+[l_1^{II}(c_1,c_2;b_3;b_1,b_2)+l_2^{II}(c_1,c_2;b_3;b_1,b_3)]$$
$$\times\sigma_{II}(\Delta_3 G)(b_1,b_2,b_3)\}. \qquad (50)$$

5. CONSISTENCY RELATIONS

In this section we will investigate the compatibility between the formulas of Sec. 3B [expressing the boundary values $\sigma(\Delta_3 G)$ in terms of the form factors G_A and G_B] and Sec. 4 [expressing the form factors in terms of $\sigma(\Delta_3 G)$].

To that end, we note the following distribution equalities:

$$\pi^{-2}\int\!\!\int\!\!\int dr db_1 db_2\ \delta((r-b_1)(r-b_2)+rb_3)$$
$$\times\delta((r-c_1)(r-c_2)+rb_3)$$
$$\times\delta((r-a_1)(r-a_2)+rb_3)\frac{(a_1-r)}{(a_1-b_1)_p}$$
$$\times\theta(a_1-r)\frac{|b_1-r|}{(b_1-c_1)_p}\theta(r)$$
$$=-[\lambda(b_3,c_1,c_3)]^{-1/2}\,\delta(a_1-c_1)\delta(a_2-c_2)$$
$$\times[2\chi(D_1(c_1,c_2,b_3))+\chi(T_2'(c_1,c_2,b_3))], \qquad (51)$$

$$\pi^{-2}\int\!\!\int\!\!\int dr\,db_1 db_2\ \delta((r-b_1)(r-b_2)+rb_3)$$
$$\times\delta((r-c_1)(r-c_2)+rb_3)$$
$$\times\delta((r-a_1)(r-a_2)+rb_3)\frac{(r-a_1)}{(a_1-b_1)_p}$$
$$\times\theta(r-a_1)\frac{|b_1-r|}{(b_1-c_1)_p}\theta(r)$$
$$=-[\lambda(b_3,c_1,c_2)]^{-1/2}\,\delta(a_1-c_1)\delta(a_2-c_2)$$
$$\times[2\chi(D_2(c_1,c_2,b_3))+\chi(T_1(c_1,c_2,b_3))], \qquad (52)$$

$$\pi^{-2}\int\!\!\int\!\!\int\!\!\int d\alpha d\beta db_1 db_2\ \delta(1-\alpha-\beta)\delta(\alpha c_2+\beta c_1-\alpha\beta b_3)$$
$$\times\delta(\alpha b_2+\beta b_1-\alpha\beta b_3)\,\delta(\alpha a_2+\beta a_1-\alpha\beta b_3)$$
$$\times\frac{\alpha^2}{(b_1-c_1)_p(a_1-b_1)_p}\theta(a_1)\theta(\alpha)\theta(\beta)$$
$$=-[\lambda(b_3,c_1,c_2)]^{-1/2}\delta(a_1-c_1)\delta(a_2-c_2)$$
$$\times[2\chi(D_3(c_1,c_2,b_3))+\chi(T_2(c_1,c_2,b_3))], \qquad (53)$$

$$\pi^{-2}\int\!\!\int\!\!\int\!\!\int d\alpha d\beta db_1 db_2\ \delta(1-\alpha-\beta)$$
$$\times\delta(\alpha c_2+\beta c_1-\alpha\beta b_3)\delta(\alpha b_2+\beta b_1-\alpha\beta b_3)$$
$$\times\delta(\alpha a_2+\beta a_1-\alpha\beta b_3)$$
$$\times\frac{\alpha^2}{(b_1-c_1)_p(a_1-b_1)_p}\theta(a_2)\theta(\alpha)\theta(\beta)$$
$$=-[\lambda(b_3,c_1,c_2)]^{-1/2}\delta(a_1-c_1)\delta(a_2-c_2)$$
$$\times[2\chi(D_3(c_1,c_2,b_3))+\chi(T_1(c_1,c_2,b_3))]. \qquad (54)$$

We have in Eqs. (51)–(54) introduced the notation $\chi(D)$ and $\chi(T)$, respectively, for the characteristic functions of the regions D and T of Sec. 2C. The characteristic function is defined to be equal to 1 in the region in question and to vanish outside it. The proofs of Eqs. (51)–(54) are straightforward and will be given in Appendix D.

We now make use of these results to compute the following integrals from Eq. (50):

$$(2\pi i)^{-2}\int\!\!\int db_1 db_2\ l_2^{I}(c_1,c_2;b_3;b_1,b_2)\sigma_{I_r}(b_1,b_2,b_3)$$
$$=(2\pi)^2\{G_A(c_1,c_2,b_3)[\chi(D_2)+\tfrac{1}{2}\chi(T_1)]$$
$$-G_B(c_1,c_2,b_3)[\chi(D_1)+\tfrac{1}{2}\chi(T_2)]\}, \qquad (55)$$

$$(2\pi i)^{-2}\int\!\!\int db_1 db_2\ l_2^{II}(c_1,c_2;b_3;b_1,b_2)\sigma_{II\alpha}(b_1,b_2,b_3)$$
$$=(2\pi)^2\{G_A(c_1,c_2,b_3)\tfrac{1}{2}\chi(T_1)-G_B(c_1,c_2,b_3)\tfrac{1}{2}\chi(T_2)$$
$$+[G_A(c_1,c_2,b_3)-G_B(c_1,c_2,b_3)]\chi(D_3)\}. \qquad (56)$$

To reach the results in the right-hand sides of Eqs. (55) and (56), we use Eqs. (47) and (49) for l_2^{I} and l_2^{II} (with appropriate choice of Λ_2) and the second lines of Eqs. (43) and (44) for σ_{I_r} and $\sigma_{II\alpha\beta}$. We have further neglected all eventual contributions from zero-mass states in accordance with the discussion in Sec. 3B.

We now sum the contributions in Eqs. (55) and (56) and compare the results to the defining Eq. (35) for G_A and G_B. The support properties of $G_A(G_B)$ (cf. the discussion in Sec. 2B) imply that the sum of the characteristic functions multiplying $G_A(G_B)$ in Eqs. (55) and (56) is equal to the characteristic function of the support of $G_A(G_B)$ The regions T and D do not overlap and eventual difficulties are then only connected to common boundary regions, i.e., when one (or both) c_j, $= j = 1, 2$, vanish ("zero-mass states"). Consequently, we deduce that the sum of the contributions from the terms containing $l_2{}^{I}$ and $l_2{}^{II}$ in Eq. (50) are sufficient to account for the whole limit $\mathfrak{L}(\Delta_3 G)$ in Eq. (35).

In that way we have shown that Eqs. (43) and (44) (defining the boundary values σ_I and σ_{II} in terms of the form factors) is compatible with Eqs. (35) and (50) defining the form factors in terms of the boundary values) iff the remaining contributions to Eq. (50) vanish:

$$\int\!\!\int db_1 db_2(l_1^{I}(c_1,c_2;b_3;b_1,b_2)\sigma_I(b_1,b_2,b_3)$$
$$+l_1^{II}(c_1,c_2;b_3;b_1,b_2)\sigma_{II}(b_1,b_2,b_3))=0. \qquad (57)$$

This is indeed the case under the same conditions as we assumed in order to derive the representation formula in DII. In order to prove it, we rewrite Eq. (57) in the following way:

$$\iint db_1 db_2 \Big(\epsilon(-p_1)\frac{\delta(b_1-c_1)}{(b_2-c_2)_p}\gamma_1(b_1,b_2) + \epsilon(p_2)\frac{\delta(b_2-c_2)}{(b_1-c_1)_p}\gamma_2(b_1,b_2)\Big). \quad (58)$$

The quantities in Eq. (58) are then defined by

$$\gamma_j = \gamma_j^{\mathrm{I}} + \gamma_j^{\mathrm{II}}, \quad j = 1, 2,$$

$$\gamma_1^{\mathrm{I}} = \int dr\ \theta(r)(b_1+b_2-b_3-2r)\delta((r-b_1)(r-b_2)+rb_3) \times \theta(b_1-r)\sigma_{\mathrm{I}r}, \quad (59)$$

$$\gamma_1^{\mathrm{I}} = \iint d\alpha d\beta\ \theta(\alpha)\theta(\beta)\delta(1-\alpha-\beta)\delta(\alpha b_2+\beta b_1-\alpha\beta b_3) \times (b_1/\alpha-\alpha b_3)[\theta(b_1)\theta(-b_2)+\tfrac{1}{2}\theta(b_1)\theta(b_2)]\sigma_{\mathrm{II}\alpha}.$$

The corresponding quantities $\gamma_2{}^{\mathrm{I}}$ and $\gamma_2{}^{\mathrm{II}}$ can be found from Eq. (59) by exchanging the indices 1 and 2 (as well as the integration variables of α and β). We will now carry out the r integral and the α–β integrals in Eq. (59) and show that γ_1 vanishes identically. We will then rely upon the results of Sec. 3A and Appendix B that the boundary values σ_{I} and σ_{II} are (almost) independent of the limit procedure and can be given in terms of limits of the retarded and advanced functions. In particular they only depend upon the boundary point (b_1, b_2, b_3) and do not depend upon r, α, and β. These conditions were also used in the derivation of the representation formula Eq. (14), in DII.

Concerning the support properties of $\gamma_1{}^{\mathrm{I}}$ and $\gamma_1{}^{\mathrm{II}}$ we conclude (cf. the discussion of the support properties of the kernel functions K^{I} and K^{II} in Sec. 2B and also Appendix D) that

(i) inside the region $D_1(b_1, b_2, b_3)$ only $\gamma_1{}^{\mathrm{I}}$ may be nonvanishing,

(ii) inside the region $D_3(b_1, b_2, b_3)$ only $\gamma_1{}^{\mathrm{II}}$ may be nonvanishing,

(iii) inside the region $T_2(b_1, b_2, b_3)$, $(\gamma_1{}^{\mathrm{I}} + \gamma_1{}^{\mathrm{II}})$ may be nonvanishing.

Inside the region D_1 there are two roots $r_\pm$ contributing from the δ function to the r integral:

$$r_\pm > 0, \quad 2r_\pm + b_3 - b_1 - b_2 = \pm\sqrt{\lambda(b_1, b_2, b_3)}. \quad (60)$$

Thus the contribution to γ_1 inside D_1 according to (i) and Eq. (59) is

$$\chi(D_1)(\sigma_{\mathrm{I}r_-} - \sigma_{\mathrm{I}r_+}) = 0. \quad (61)$$

In the same way, inside the region D_3 there are two roots $\alpha_\pm$ contributing and

$$0 < \alpha_\pm = (2b_3)^{-1}[b_3 + b_1 - b_2 \pm \sqrt{\lambda(b_3, b_1, b_2)}] < 1,$$
$$b_1/\alpha_\pm - \alpha_\pm b_3 = \mp\sqrt{\lambda(b_3, b_1, b_2)}. \quad (62)$$

The contribution to γ_1 inside D_3 is then according to (ii) and Eq. (59):

$$\chi(D_3)(\sigma_{\mathrm{II}\alpha_+} - \sigma_{\mathrm{II}\alpha_-}) = 0. \quad (63)$$

Finally inside the region T_2 only the roots r_+ and α_+ in Eqs. (60) and (62) will give any contribution to γ_1. The sum is

$$\chi(T_2)(\sigma_{\mathrm{II}\alpha_+} - \sigma_{\mathrm{I}r_+}) = \chi(T_2)[\Delta_3 G(b_1+i\epsilon', b_2-i\epsilon'', b_3) + \Delta_3 G(b_1-i\epsilon', b_2+i\epsilon'', b_3) - \Delta_3 G(b_1+i\epsilon', b_2+i\epsilon'', b_3) - \Delta_3 G(b_1-i\epsilon', b_2-i\epsilon'', b_3)] = 0. \quad (64)$$

The vanishing of the left-hand side of Eq. (64) comes about not only because the boundary values are independent of α and r but also from the fact that inside the region T_2 (where $b_2 < 0$) the sign of the imaginary part of the second argument is of no significance.

The conditions in Eqs. (61) and (64) have immediate counterparts in conditions for the vanishing of the quantity γ_2:

$$\chi(D_2)(\sigma_{\mathrm{I}r_-} - \sigma_{\mathrm{I}r_+}) = 0, \quad (65)$$

$$\chi(T_1)(\sigma_{\mathrm{I}r_+} - \sigma_{\mathrm{I}\alpha_-}) = 0. \quad (66)$$

In all these equations the quantities $\sigma_{\mathrm{I}r}$ and $\sigma_{\mathrm{II}\alpha}$ are defined by Eqs. (43) and (44). Consequently Eqs. (61) and (63)–(66) are actually conditions on the form factors G_A and G_B in order that the formalism should be consistent.

In the next section we will further discuss this interpretation of the results. We will end this section by the following remarks. As of now we have only discussed requirements on the form factors G_A and G_B from the properties of the absorptive part $\Delta_3 G$. It is evident that a completely similar discussion can be carried through for the absorptive parts $\Delta_1 G$ and $\Delta_2 G$. It should be noted, however, that such an investigation has a direct bearing on the form factors G_A and G_B.

In connection with the absorptive part $\Delta_1 G$ we may define a limit like the one in Eq. (33). We will thereby get a formula for the form factors G_B and G_C, with G_C defined by

$$G_C(-p_1^2, -p_2^2, -(p_1+p_2)^2) = (2\pi)^6 \sum_{|n>|m>} [\theta(p_1)\theta(-p_2)\delta(p_1-p_n) \times \delta(p_2+p_m)\langle 0|B|m\rangle\langle m|C|n\rangle\langle n|A|0\rangle + \theta(-p_1)\theta(p_2)\delta(p_1+p_n) \times \delta(p_2-p_m)\langle 0|A|n\rangle\langle n|C|m\rangle\langle m|B|0\rangle]. \quad (67)$$

We will finally end up with a set of integral equations like the ones in Eqs. (25), (61) and (63)–(66) but this time involving (the difference of) the form factors $(G_C\text{-}G_B)$.

An investigation of $\Delta_2 G$ results in conditions on $(G_A\text{-}G_C)$. It is evident that the difference $(G_B\text{-}G_A)$ (which effectively occurs in all the integral equations derive so far) cannot be discussed independent of the "new" equations because

$$G_B - G_A = (G_B - G_C) + (G_C - G_A). \quad (68)$$

The "new" equations do, however, serve as a precise definition of G_B and G_A in "the overlap region" D_3, where only the difference can be defined by Eq. (33). But even outside the region D_3 will the form factors G_A and G_B in general be restricted by the analyticity properties of the absorptive parts $\Delta_1 G$ and $\Delta_2 G$. The reason is, of course, that there is one unique vertex function $G(Z_1, Z_2, Z_3)$

and the three quantities $\Delta_1 G$, $\Delta_2 G$, and $\Delta_3 G$ are different boundary values of the function G. The "new" equations can easily be found from Eqs. (25), (61), and (63)–(66) by permutation of indices.

6. CONCLUDING REMARKS

(1) In section 5 we have derived a set of conditions on the form factors G_A, G_B, and G_C which will guarantee the *existence* and *uniqueness* of the absorptive parts $\Delta_1 G$, $\Delta_2 G$, and $\Delta_3 G$ of the vertex functions.

If the form factors G_A and G_B fulfill the integral equations in Eqs. (25), (61), and (63)–(66), then we may compute the weight functions $\sigma_{\rm I}$ and $\sigma_{\rm II}$ according to Eqs. (43) and (44). We may afterwards insert these expressions in the representation formula Eq. (14) and compute the absorptive $\Delta_3 G$. The resulting analytic function $\Delta_3 G$ will then

(i) exhibit the analyticity properties required by KW,
(ii) have the "input" form factors as physical limits according to Eq. (35),
(iii) vanish in asymptotic directions.

The property (i) is true for any weight functions $\sigma_{\rm I}$ and $\sigma_{\rm II}$ (such that the integrals converge) and (iii) is a boundedness condition. The property (ii), is, however, a *uniqueness condition* on the form factors.

If the form factors G_A and G_B also fulfill the permuted equations involving G_C, then we may in the same way construct the absorptive parts $\Delta_1 G$ and $\Delta_2 G$. We are then also assured that properties (i)–(iii) are fulfilled for these functions.

The analyticity properties of KW do imply that the physical assumptions (1)–(3) in the Introduction are fulfilled for the appropriate matrix elements. Thus the analyticity properties of the absorptive part $\Delta_3 G$, which are implied by the representation formula, Eq. (14), do result in the vanishing of the quantity M outside the light-cone[14]:

$$M = \sum_{|n>} \langle 0 | [A(x_1), B(x_2)] | n \rangle \langle n | C | 0 \rangle \delta(p_3 - p_n). \qquad (69)$$

Therefore, the matrix element M, which is the Fourier transform of $\mathcal{L}(\Delta_3 G)$ [cf. Eq. (34)] fulfills the locality conditions *if the form factors G_A and G_B fulfill the integral equations*. We may consequently interpret Eqs. (25), (61), (63)–(66) as *a set of sufficient conditions on the form factors so that the appropriate matrix elements of the field theory fulfill the physical assumptions*, as well as the boundedness assumptions which are necessary in order that the representation formulas should make sense.

(2) We have during the investigation repeatedly changed the orders of integrations and performed certain limits inside the occurring integrals without further justification. All such formal operations can without doubt be discussed in a more rigorous mathematical setting. We will, however, not carry out any such detailed discussions but be satisfied with a few remarks to supplement those already made in due places of the main text.

The difficulties which are due to the unboundedness of some of the occurring integration ranges can always be solved along the lines indicated in Appendix A ("subtracted dispersion relations") as long as the main assumption on polynomial boundedness of the vertex function is fulfilled. The integral equations then involve the generalized form-factors $g_A^{m,n}$ and $g_B^{m,n}$ defined by [cf. eqs. (A4)–(A6) in Appendix A] ($z_1 < 0, z_2 < 0$)

$$g_A^{m,n}(b_1, b_2, b_3; Z_1, Z_2) = \left(\frac{\partial}{\partial Z_1}\right)^m \left(\frac{\partial}{\partial Z_2}\right)^n \times \frac{G_A(b_1, b_2, b_3) - G_A(Z_1, b_2, b_3)}{(b_1 - Z_1)(b_2 - Z_2)},$$

$$g_B^{m,n}(b_1, b_2, b_3; Z_1, Z_2) = \left(\frac{\partial}{\partial Z_1}\right)^m \left(\frac{\partial}{\partial Z_2}\right)^n \times \frac{G_B(b_1, b_2, b_3) - G_B(b_1, Z_2, b_3)}{(b_1 - Z_1)(b_2 - Z_2)}. \qquad (70)$$

It should also be understood that in case there are singularities in the finite part of the integration domains a similar procedure can sometimes by developed. Thus, in case there is a simple pole at the origin, in one of the variables, e.g., Z_1 (i.e., there is a state with vanishing mass in the theory with the quantum numbers of the field A), then we may investigate the generalized absorptive part $\Delta_3 G'$ defined by ($z_1 < 0$)

$$\Delta_3 G'(Z_1, Z_2, b_3; z_1) = [Z_1/(Z_1 - z_1)] \times [\Delta_3 G(Z_1, Z_2, b_3) - \Delta_3 G(z_1, Z_2, b_3)]. \qquad (71)$$

Equation (71) will result in the following generalized form factors g'_B and g'_A

$$g'_B(b_1, b_2, b_3; Z_1) = [b_1/(b_1 - Z_1)] G_B(b_1, b_2, b_3),$$
$$g'_A(b_1, b_2, b_3; Z_1) = [b_1/(b_1 - Z_1)][G_A(b_1, b_2, b_3) - G_A(z_1, b_2, b_3)] \qquad (72)$$

It is clear that the singular contribution from G_B at $b_1 = 0$ in an integral like Eq. (43) or Eq. (44) for $r = 0$ or $\alpha = 0$ is not present in g'_B.

The dispersion relations can then be applied to the generalized function $\Delta_3 G'$. We note that "the subtraction" at the point z_1 leads to the same asymptotic properties. The actual absorptive part $\Delta_3 G$ can be recovered by

$$\Delta_3 G\,(Z_1, Z_2, b_3) = \Delta_3 G(z_1, Z_2, b_3) + [(Z_1 - z_1)/Z_1] \Delta_3 G'(Z_1, Z_2, b_3; z_1). \qquad (73)$$

In Eq. (73) the pole term is explicitly exhibited. The generalization to other situations should be obvious.

(3) The possibility above is of some interest in the light of the results of the investigation in Sec. 3 on "the approach to the boundary." It is shown that, unless there are zero-mass states in the theory, the limit procedure used in this paper is unique.

The integral equations (61), (63)–(66), which are derived in Sec. 5 can be interpreted as requirements of just this uniqueness. Thus we have exploited the fact that it is possible to approach each point on the distinguished boundary *along different directions* and still get *the same boundary value* of the vertex function. In particular it is possible to choose as direction of approach, directions "along" the singularity-surfaces, called the anomalous cuts. All points on the distinguished boundary can as a matter of fact be reached in that way from *two* different directions. The integral equations are then requirements *that these two different limits should be equal*.

Consequently, in a theory with appropriate boundedness properties of the vertex function [Eq. (25)] and no zero-mass states, the conditions on the integral equations are also *necessary conditions*.

In case the boundedness requirements are not fulfilled but the vertex function is allowed to grow at most polynomially at infinity and (or) as an inverse power around the origin of mass space, then it is possible to define generalized form factors according to the remark (2) above. The conditions on these generalized form factors are correspondingly less restrictive.

(4) The formulation of the integral equation in Eqs. (61), (63)–(66) brings out the formal symmetry between the occurring mass variables. For practical computation it might be useful to reformulate the equations in terms of other variables. We will here briefly outline such a reformulation of Eq. (64).

A convenient set of variables for Eq. (64) [which is valid inside the domain $T_2(b_1, b_2, b_3)$] is $(b_1, r = r_+, b_3)$. We note the following relation between the parameter $\alpha_- = \alpha$ and r which follows from the definitions in Eqs. (60) and (62):

$$\alpha b_2 = b_1 - n > 0, \qquad \beta b_3 = r - b_1 + b_3 > 0. \tag{74}$$

If we use the second line of Eq. (44) as definition for $\sigma_{\mathrm{II}\alpha}$, we get with this result inserted

$$\begin{aligned}\sigma_{\mathrm{II}\alpha} &= 4\pi i \int \frac{da_1}{(a_1 - b_1)_p}\left[G_A\left(a_1 \beta b_3 - \frac{\beta}{\alpha} a_1, b_3\right)\right. \\ &\quad \left. - G_A\left(a_1 \beta b_3 - \frac{\beta}{\alpha} a_1, b_3\right)\right] \\ &= 4\pi i \iint da_1 da_2 \frac{(b_1 - r)}{(a_1 - b_1)_p}\delta([(a_2 - r)(b_1 - r) + r b_3] \\ &\quad + (a_1 - b_1)(r - b_1 + b_3))(G_B - G_A).\end{aligned} \tag{75}$$

For the remaining quantity in Eq. (64) we get from Eq. (43):

$$\begin{aligned}\sigma_{\mathrm{I}r} &= 4\pi i \iint da_1 da_2 \left(\frac{b_1 - r}{(a_1 - b_1)_p} + \frac{1}{2}\right) \\ &\quad \times \delta((a_2 - r)(a_1 - r) + r b_3)(G_B - G_A) \\ &= 4\pi i \iint da_1 da_2 \left(\frac{b_1 - r}{a_1 - b_1}\right)_p \delta((a_2 - r) \\ &\quad \times (a_1 - r) + r b_3)(G_B - G_A).\end{aligned} \tag{76}$$

Just as in connection with the two different versions of Eq. (43) we have made use of the fact that (b_1, b_2, b_3) and (a_1, a_2, b_3) are on ACI with the same parameter r. Note that, in going from the first to the second line of Eq. (76), we have also made use of the sum rule in Eq. (25).

Consequently, we may write Eq. (64) in terms of the variables (b_1, r, b_3) as

$$\begin{aligned}0 = (b_1 - r) \iint da_1 da_2 \frac{1}{(a_1 - b_1)_p}[\delta((a_2 - r)(b_1 - r) + r b_3 \\ + (a_1 - b_1)(r + b_3 - b_1)) - \delta((a_2 - r)(a_1 - r) + r b_3) \\ \times (G_B - G_A)].\end{aligned} \tag{77}$$

The region T_2 is in terms of the variables (b_1, r, b_3):

$$0 < r < b_1 < r + b_3. \tag{78}$$

Equation (77) contains a few features which are generally true for the integral equations:

(i) Only the difference of the form factors $(G_A - G_B)$ occur (cf. the remarks in Sec. 2A).
(ii) The principal value prescription is actually not necessary because in the points where the integration variable a_1 equals b_1 the arguments in the δ functions are equal too.

The remaining equations in Sec. 5 can also be reformulated into formulas similar to Eq. (77).

The simplification is bought, however, at the price of some obvious inconvenience in Eq. (77). Thus, the formula must be interpreted by appropriate limit procedures when the parameter b_1 approaches r, and the points $a_1 = r$ and $a_2 = r$ of the integration range.

(5) The corresponding relation for the two-point function is the well-known requirement of "weak local commutativity." This means that the two spectral functions G_{AB} and G_{BA} defined by

$$\begin{aligned}G_{AB}(-p^2) &= \sum_{|m\rangle} \delta(p - p_m)\langle 0|A|m\rangle\langle m|B|0\rangle, \\ G_{BA}(-p^2) &= \sum_{|m\rangle} \delta(p - p_m)\langle 0|B|m\rangle\langle m|A|0\rangle\end{aligned} \tag{79}$$

must fulfill

$$G_{AB}(a) - G_{BA}(a) = 0. \tag{80}$$

The conditions for the three point function, Eq. (73), which we have derived here, can be considered as a straightforward generalization of Eq. (76). The condition that the difference of the spectral functions of two commuting fields should vanish corresponds for the form factors to the condition that the difference $(G_A - G_B)$ should vanish when integrated over certain (one-dimensional) real surfaces.

(6) There are a set of obvious solutions to the integral equations (61), (63)–(66), which can be found from Eq. (50):

$$\begin{aligned}G_A(c_1, c_2, b_3) - G_B(c_1, c_2, b_3) = -(2\pi)^{-4} \iint db_1 db_2 \\ \times [l_2^{\mathrm{I}}(c_1, c_2; b_3; b_1, b_2)\sigma_{\mathrm{I}}(b_1, b_2, b_3) \\ + [l_2^{\mathrm{II}}(c_1, c_2; b_3; b_1, b_3)\sigma_{\mathrm{II}}(b_1, b_2, b_3)].\end{aligned} \tag{81}$$

Any set of weight functions σ_{I} and σ_{II} such that

(i) σ_{I} fulfills the superconvergence relation in Eq. (25),
(ii) $\sigma_{\mathrm{I}} = \sigma_{\mathrm{II}}$ inside the regions T_1 and T_2

can easily be seen to fulfill the equations. We will not give any details as the computations are very similar to the ones above. The following results can also be proven by straightforward computations:

(a) Insertion of (the difference of) the form factors according to Eq. (81) in Eqs. (43) and (44) will after some few computations give back the "input functions" σ_{I} and σ_{II}.

(b) Insertion of (the difference of) the form factors according to Eq. (81) in Eqs. (23) and (24) will also give back Eq. (14).

There is consequently a one-to-one correspondence between the form factors $G_A - G_B$ (fulfilling the integral equations) and the weight functions σ_{I} and σ_{II} [fulfilling

properties (i) and (ii) above]. This relation closely resembles the connection between "the real and imaginary parts" for a function analytic in the cut plane according to the example in the Introduction.

The solution in Eq. (81) is in general nonzero in the whole "physical region" T_1, T_2, D_1, D_2, and D_3. In a theory with a "realistic" mass spectrum there are then further restrictions on the weight functions σ_I and σ_{II}. If, e.g., there are no states in the theory with squared mass less than m^2, then the left-hand side of Eq. (81) vanishes in appropriate parts of the (c_1, c_2, b_3) space. The resulting condition on σ_I and σ_{II} are integral equations which can be seen to closely resemble Eqs. (61), (63)–(66) for G_A and G_B.

ACKNOWLEDGMENT

I would like to express my gratitude toward Professor A. S. Wightman for several enlightening discussions (the examples in the Introduction were pointed out by him), as well as for making my stay at Princeton possible.

APPENDIX A

In the formulas of the main text we have assumed that the vertex function in momentum space vanishes in asymptotic directions inside the analyticity domain derived in KW. In this appendix we will extend the formalism to cover also the cases when the vertex function grows at most as a polynomial when one or more of the variables approaches infinity.[1]

To that end it is useful to introduce the "associated" vertex functions $g_3^{n_1 n_2}(Z_1, Z_2, b_3; z_1, z_2)$ defined by

$$
\begin{aligned}
g_3^{00} &= \Delta_3 G, \\
g_3^{10} &= [\Delta_3 G(Z_1, Z_2, b_3) - \Delta_3 G(z_1, Z_2, b_3)](Z_1 - z_1)^{-1}, \\
g_3^{01} &= [\Delta_3 G(Z_1, Z_2, b_3) - \Delta_3 G(Z_1, z_2, b_3)](Z_2 - z_2)^{-1}, \\
g_3^{11} &= (\Delta_3 G(Z_1, Z_2, b_3) - \Delta_3 G(z_1, Z_2, b_3) - \Delta_3 G(Z_1, z_2, b_3) \\
&\quad + \Delta_3 G(z_1, z_2, b_3)) \times (Z_1 - z_1)^{-1}(Z_2 - z_2)^{-1}, \\
g_3^{n_1+1\, n_2+1} &= \left(\frac{\partial}{\partial z_1}\right)^{n_1}\left(\frac{\partial}{\partial z_2}\right)^{n_2} g_3^{11}.
\end{aligned} \tag{A1}
$$

The following two properties of the functions $g_3^{n_1 n_2}$ are immediate consequences of the definitions and the analyticity properties of the vertex function proved in KW.

(1) If the parameters (z_1, z_2) are arbitrary *negative* numbers, then for any positive integers n_1 and n_2 the functions $g_3^{n_1 n_2}$ are analytic in the same domain as the absorptive part $\Delta_3 G$ of the vertex function.

(2) When the absorptive part $\Delta_3 G$ is at most polynomially increasing in asymptotic direction inside the analyticity domain then by choosing the integers n_1 and n_2 sufficiently large we can make the associated function $g_3^{n_1 n_2}$ to vanish asymptotically.

Consequently, even if the dispersion relations for $\Delta_3 G$ do not make sense due to lack of damping in the integrals, the representation formulas may be meaningful for one of the associated functions. By induction it is easy to prove the following relation:

$$
\begin{aligned}
\Delta_3 G(Z_1, Z_2, b_3) &= (Z_1 - z_1)^{n_1+1}(Z_2 - z_2)^{n_2+1}\frac{1}{n_1!}\frac{1}{n_2!} \\
&\quad \times g_3^{n_1 n_2}(Z_1, Z_2, b_3; z_1, z_2) \\
&\quad + \sum_{j=0}^{n_1} \frac{1}{j!}(Z_1 - z_1)^j \left(\frac{\partial}{\partial z_1}\right)^j \frac{1}{n_2!}(Z_2 - z_2)^{n_2+1} \\
&\quad \times g_3^{0 n_2}(z_1, Z_2, b_3; z_1, z_2) \\
&\quad + \sum_{k=0}^{n_2} \frac{1}{k!}(Z_2 - z_2)^k \left(\frac{\partial}{\partial z_2}\right)^k \frac{1}{n_1!}(Z_1 - z_1)^{n_1+1} \\
&\quad \times g_3^{n_1 0}(Z_1 z_2, b_3; z_1, z_2) \\
&\quad - \sum_{j=0}^{n_2}\sum_{k=0}^{n_2} \frac{1}{j!}\frac{1}{k!}(Z_1 - z_1)^j (Z_2 - z_2)^k \\
&\quad \times \left(\frac{\partial}{\partial z_1}\right)^j \left(\frac{\partial}{\partial z_2}\right)^k \Delta_3 G(z_1, z_2, b_3).
\end{aligned} \tag{A2}
$$

The functions $g^{0 n_2}(z_1, Z_2, b_3; z_1, z_2)$ and $g^{n_1 0}(Z_1, z_2, b_3; z_1, z_2)$ are actually analytic in the whole complex plane cut along the positive real axis of the corresponding Z variable when (z_1, z_2) are negative. It is as a matter of fact simple to prove that

$$
\begin{aligned}
&\frac{1}{n_2!}(Z_2 - z_2)^{n_2+1} g_3^{0 n_2}(z_1, Z_2, b_3; z_1, z_2) = \Delta_3 G(z_1, Z_2, b_3) \\
&\quad - \sum_{j=0}^{n_2} \frac{1}{j!}(Z_2 - z_2)^j \left(\frac{\partial}{\partial z_2}\right)^j \Delta_3 G(z_1, z_2, b_3) \\
&\quad = -2\pi i (Z_2 - z_2)^{n_2+1} \int_0^\infty \frac{db_2}{(b_2 - Z_2)} \frac{G_A(z_1, b_2, b_3)}{(b_2 - Z_2)^{n_2+1}}.
\end{aligned} \tag{A3}
$$

To achieve this result, we have made use of the representation formula for the absorptive part in terms of the advanced and retarded functions in Eqs. (31)–(35) of the main text. A very similar result is valid for the quantity $g^{n_1 0}$. Assuming that n_1 and n_2 have been chosen so large that the associated function $g_3^{n_1 n_2}$ vanishes in asymptotic directions, then we may represent it by means of Eq. (14) of the main text. The weight functions in that representation $\sigma_I(g_3^{n_1 n_2})$ and $\sigma_{II}(g_3^{n_1 n_2})$ can be defined as functionals of the form factors by means of the same procedure as in Sec. 3, i.e., by Eqs. (42) and (44) of the main text:

$$
\begin{aligned}
\sigma_I(g_3^{n_1 n_2}) &= 4\pi i \int da_1 da_2 \delta((r - a_1)(r - a_2 + r b_3) \\
&\quad \times \left(\frac{1}{2}\frac{a_1 + b_1 - 2r}{(a_1 - b_1)_p} g_B^{n_1 n_2}\theta(a_1 - r) \right. \\
&\quad \left. + \frac{1}{2}\frac{a_2 + b_2 - 2r}{(a_2 - b_2)_p} g_A^{n_1 n_2}\theta(a_2 - r)\right).
\end{aligned} \tag{A4}
$$

A similar formula can be given for the weight function $\sigma_{II}(g_3^{n_1 n_2})$. The "generalized form factors" $g_B^{n_1 n_2}$ and $g_A^{n_1 n_2}$ are defined by

$$
\begin{aligned}
g_B^{n_1 n_2} &= \frac{n_1!}{(a_1 - z_1)^{n_1+1}}\left(\frac{\partial}{\partial z_2}\right)^{n_2}[G_B(a_1, a_2, b_3) \\
&\quad - G_B(a_1, z_2, b_3)](a_2 - z_2)^{-1}, \\
g_A^{n_1 n_2} &= \frac{n_2!}{(a_2 - z_2)^{n_2+1}}\left(\frac{\partial}{\partial z_2}\right)^{n_1}(G_A(a_1, a_2, b_3) \\
&\quad - G_A(z_1, a_2, b_3))(a_1 - z_1)^{-1}.
\end{aligned} \tag{A5}
$$

Then we find by means of Eq. (14):

$$
\begin{aligned}
g_3^{n_1 n_2} &= \frac{1}{(2\pi i)^2}\iint db_1 db_2 [K^I(Z_1, Z_2, b_3; b_1, b_2)\sigma_I(g_3^{n_1 n_2}) \\
&\quad + K^{II}(Z_1, Z_2, b_3; b_1, b_2)\sigma_{II}(g_3^{n_1 n_2})].
\end{aligned} \tag{A6}
$$

Using eqs. (A6), (A4), (A5), and (A3), we find a representation formula for the absorptive part $\Delta_3 G$ in Eqs. (A2).

We will now assume that the generalized form factors $g_A^{n_1 n_2}$ and $g_B^{n_1 n_2}$ in eq. (A5) fulfill the integral equations derived in Sec. 5 of the main text. Then we find by an application of the limit in Eq. (35) from the first line in Eq. (A2) [cf. Eq. (A6)]

$$\begin{aligned}
&\mathfrak{L}\left((Z_1 - z_1)^{n_1+1}(Z_2 - z_2)^{n_2+1}\frac{1}{n_1!\,n_2!}g_3^{n_1 n_2}\right)\\
&= (2\pi)^2\theta(p_3)\left(\theta(-p_2)\frac{(b_2 - z_2)^{n_2+1}}{n_2!}\frac{(b_1 - z_1)^{n_1+1}}{n_1!}g_A^{n_1 n_2}\right.\\
&\quad\left. - \theta(-p_1)\frac{(b_1 - z_1)^{n_1+1}}{n_1!}\frac{(b_2 - z_2)^{n_2+1}}{n_2!}g_B^{n_1 n_2}\right)\\
&= (2\pi)^2\theta(p_3)\left\{\theta(-p_2)\left[G_A(b_1, b_2, b_3)\right.\right.\\
&\quad\left. - \sum_{j=0}^{n_1}\frac{1}{j!}(b_1 - z_1)^j\left(\frac{\partial}{\partial z_1}\right)^j G_A(z_1, b_2, b_3)\right]\\
&\quad - \theta(-p_1)\left[G_B(b_1, b_2, b_3) - \sum_{k=0}^{n_2}\frac{1}{k!}(b_2\right.\\
&\quad\left.\left. - z_2)^k\left(\frac{\partial}{\partial z_2}\right)^k G_B(b_1, z_2, b_3)\right]\right\}. \qquad \text{(A7)}
\end{aligned}$$

In the same limit we get the following contribution from the second line in Eq. (A2) [cf. Eq. (A3)]:

$$(2\pi)^2\theta(p_3)\theta(-p_2)\sum_{j=0}^{n_1}\frac{1}{j!}(b_1 - z_1)^j\left(\frac{\partial}{\partial z_1}\right)^j G_A(z_1, b_2, b_3). \qquad \text{(A8)}$$

A similar term (with the indices 1 and 2 exchanged) comes from the third line. The fourth line does not give any contribution in this limit. Consequently, we find by summing the different contributions that

$$\mathfrak{L}(\Delta_3 G) = (2\pi)^2[\theta(p_3)\theta(-p_2)G_A - \theta(p_2)\theta(-p_1)G_B]. \qquad \text{(A9)}$$

Thus the only change in the formalism in this paper is that the generalized form factors $g_A^{n_1 n_2}$ and $g_B^{n_1 n_2}$ occur instead of G_A and G_B, in case the absorptive part $\Delta_3 G$ "needs subtractions."

APPENDIX B

We will in this appendix investigate the properties of the analyticity region close to the boundary region inside which the integration in Eq. (14) is performed. We will explicitly construct (in two-dimensional space–time) the limiting arguments $(b_1 + i\epsilon', b_2 + i\epsilon'', b_3)$ in terms of Lorentz squares of vectors p_j and k_j according to Eq. (1). To that end we choose light cone coordinates (s_j, t_j), $j = 1, 2$, for the vectors p_1 and p_2. The limiting imaginary parts of the vectors k_1 and $k_2 = -k_1$ have the light cone coordinates $\epsilon(\mathcal{K}, \lambda) = k_2$. The limit parameter is ϵ.

Then we have

$$b_1 = -p_1^2 = s_1 t_1, \qquad \text{(B1)}$$

$$b_2 = -p_2^2 = s_2 t_2, \qquad \text{(B2)}$$

$$b_3 = -p_3^2 = -(p_1 + p_2)^2 = (s_1 + s_2)(t_1 + t_2), \qquad \text{(B3)}$$

$$\epsilon' = k_2 \cdot p_1 = -\tfrac{1}{2}\epsilon(s_1\lambda + t_1\mathcal{K}), \qquad \text{(B4)}$$

$$\epsilon'' = -k_2 p_2 = \tfrac{1}{2}\epsilon(s_2\lambda + t_2\mathcal{K}). \qquad \text{(B5)}$$

The condition that k_2 is inside a definite light cone is that $\mathcal{K}$ and λ have the same sign:

$$-k_2^2 = \epsilon^2\mathcal{K}\lambda > 0. \qquad \text{(B6)}$$

We will in particular investigate the neighborhood of the intersection between the anomalous cut ACI and the real axes, and we therefore assume that

$$(r - b_1)(r - b_2) + rb_3 = 0, \quad r > 0. \qquad \text{(B7)}$$

The two possible roots for the parameter r in this equation is in terms of light cone coordinate above:

$$r = r_1 = -s_1 t_2, \quad r = r_2 = -s_2 t_1. \qquad \text{(B8)}$$

Assuming that $r_1 > 0$ and solving for s_1, t_1 and s_2 in terms of t_2 from Eqs. (B1), (B2), and (B8), we get

$$s_1 = -r_1/t_2, \quad s_2 = b_2/t_2, \quad t_1 = -b_1 t_2/r_1. \qquad \text{(B9)}$$

Equation (B3) will only serve as a consistency requirement in this connection; insertion of the result of Eq. (B9) gives back the eq. (B7) for ACI.

From Eqs. (B4), (B5), and (B9) we find for the ratios

$$\epsilon'/\hat{\epsilon} = r_1\lambda + b_1 t_2^2/r_1,$$
$$\epsilon''/\hat{\epsilon} = b_2\lambda + t_2^2, \quad \hat{\epsilon} = \epsilon/2t_2. \qquad \text{(B10)}$$

Thus we find that the ratios have definite signs iff the corresponding quantity b_j is positive, but can take on both signs depending upon the ratio of $\mathcal{K}/\lambda$ and t_2^2 elsewhere. In particular the ratio ϵ'/ϵ'' can take on any positive real value (nonzero, finite) if both b_1 and $b_2 > 0$.

Inside the region $D_1(b_1, b_2, b_3)$ the explicit expressions for $\sigma_I(\Delta_3 G)$ in terms of the retarded and advanced functions in Eqs. (B1) and (B2) is (note that if $p_3 \in V^+$, then inside $D_1, p_3 \in V^+$ and $p_1 = -p_2 - p_3 \in V^-$)

$$\begin{aligned}
\sigma_I(\Delta_3 G) &= \Delta_3 G(b_1 + i\epsilon', b_2 + i\epsilon'', b_3)\\
&\quad + \Delta_3 G(b_1 - i\epsilon', b_2 - i\epsilon'', b_3)\\
&= \lim_{k_2 \to 0}\left(\Delta_3 G\,|_{k_2 \in V^+} + \Delta_3 G\,|_{k_2 \in V^-}\right)\\
&= (2\pi)^4 \int dx\, e^{ip_1 x_1 + ip_2 x_2}\sum_{|n\rangle}\delta(p_3 - p_n)\\
&\quad \times \langle 0\,|\epsilon(12)[B(x_2), A(x_1)]\,|n\rangle\langle n|C\,|0\rangle.
\end{aligned}$$

Similar formulas can be given for the weight function σ for other parts of the integration region.

APPENDIX C: LIMIT VALUES

In this appendix we will give a few details of the limiting procedure leading to Eqs. (45)–(49) in Sec. 4 of the main text. The quantities to be investigated are $\mathfrak{L}(K^{\mathrm{I}})$ and $\mathfrak{L}(K^{\mathrm{II}})$ defined by e.g. [cf. Eq. (33)]

$$\begin{aligned}
\mathfrak{L}(K^{\mathrm{I}}) = \lim_{k \in V^+ \to 0}&[K^{\mathrm{I}}(-(p_1 - ik)^2, -(p_2 + ik)^2)\\
&- K^{\mathrm{I}}(-(p_1 + ik)^2, -(p_2 - ik)^2)] \qquad \text{(C1)}
\end{aligned}$$

for the kernel function $K^{\mathrm{I}}(Z_1, Z_2)$ in Eq. (15). We have not indicated the dependence on the mass variables b_j, but note that $b_3 = -p_3^2 = -(p_1 + p_2)^2$. We will frequently make use of the following well-known relations

$$\lim_{\epsilon \to +0}[1/(X \pm i\epsilon)] = 1/X_p \mp i\pi\delta(X). \qquad \text{(C2)}$$

It is useful to define the quantities j_1 and j_2 by

$$K^{\mathrm{I}} = \int dr\ \theta(r)\delta((r-b_1)(r-b_2)+rb_3) \times[\theta(b_1-r)j_1 + \theta(b_2-r)j_2]. \quad \text{(C3)}$$

In the indicated limit we find for, e.g., j_1 defined by

$$j_1 = \frac{\theta(b_1-r)}{b_1-Z_1}\left(1+\frac{b_1+Z_1-2r)}{2} \times\frac{(2r+b_3-Z_1-Z_2)}{(r-Z_1)(r-Z_2)+rb_3}\right), \quad \text{(C4)}$$

$$\mathfrak{L}(j_1) = \theta(b_1-r)\int dc_1 dc_2\ \delta(p_1^2+c_1)\delta(p_2^2+c_2)\mathfrak{K}_1, \quad \text{(C5)}$$

$$\mathfrak{K}_1 = -2\pi i\epsilon(p_1)\delta(b_1-c_1) \times\left(1+\frac{(b_1+c_1-2r)}{2}\frac{(2r+b_3-c_1-c_2)}{(r-c_1)(r-c_2)+rb_3}\right) + 2\pi i\epsilon(n_{\mathrm{I}})\delta((r-c_1)(r-c_2)+rb_3) \times\frac{1}{2}\frac{(b_1+c_1-2r}{(b_1-c_1)_p}(2r+b_3-c_1-c_2). \quad \text{(C6)}$$

The sign function ϵ with a vector argument is positive (negative) when the vector belongs to V^+ (V^-). We note that due to the positivity of b_1 the vector p_1 is timelike in the first term and therefore the sign function is well defined.

The vector argument in the sign function of the second term is

$$n_{\mathrm{I}} = (r-c_1)p_2 - (r-c_2)p_1. \quad \text{(C7)}$$

Making use of the restriction on (c_1, c_2, b_3) according to the δ function, we find

$$-n_{\mathrm{I}}^2 = (1/r)(r^2-c_1c_2)^2 = r(2r+b_3-c_1-c_2)^2 \geq 0. \quad \text{(C8)}$$

The vector n_{I} is consequently timelike or lightlike. To investigate which light cone the vector belongs to, we make use of the assumption that $p_3 = -p_1-p_2$ is in V^+ [cf. Eq. (28)]. We find

$$n_{\mathrm{I}}\cdot p_3 = \tfrac{1}{2}(c_2-c_1)(2r+b_3-c_1-c_2). \quad \text{(C9)}$$

Thus n_{I} changes light cone at the point $2r+b_3-c_1-c_2 = 0$. The argument of the δ function can be written as [cf. Eq. (38)]

$$(r-c_1)(r-c_2)+rb_3 = \tfrac{1}{4}[(2r+b_3-c_1-c_2)^2 - \lambda(b_3,c_1,c_2)]. \quad \text{(C10)}$$

We conclude that this condition corresponds to the configuration when $\lambda(b_3, c_1, c_2) = 0$, i.e.,

$$b_3 = (\sqrt{c_1} \pm \sqrt{c_2})^2. \quad \text{(C11)}$$

This can only happen when the vectors p_1 and p_2 are parallel. We note that the term $\epsilon(n_{\mathrm{I}})$ occurs multiplied by the factor $(2r+b_3-c_1-c_2)$ and we may therefore write

$$\epsilon(n_{\mathrm{I}})\delta((r-c_1)(r-c_2)+rb_3)(2r+b_3-c_1-c_2) = \epsilon(c_1-c_2)\delta((r-c_1)(r-c_2)+rb_3)\sqrt{\lambda(b_3,c_1,c_2)}. \quad \text{(C12)}$$

In this way the result for the quantity l_2^{I} is evident.

The first term of $\mathfrak{K}_1$ can also be simplified by use of the δ functions $\delta(b_1-c_1)$ and $\delta((r-b_1)(r-b_2)+rb_3)$:

$$1+\frac{(b_1+c_1-2r)}{2}\frac{(2r+b_3-c_1-c_2)}{((r-c_1)(r-c_2)+rb_3)_p} = 1+(b_1-r)\frac{2r+b_3-b_1-c_2}{((r-b_1)(r-c_2)+rb_3)_p} = 1+(2r+b_3-b_1-c_2)\frac{1}{(c_2-b_2)_p} = \frac{2r+b_3-b_1-b_2}{(c_2-b_2)_p}. \quad \text{(C13)}$$

Applying very similar considerations to the term j_2 in Eq. (C3) we arrive at the Eq. (45) for $\mathfrak{L}(K^{\mathrm{I}})$. In the same way Eq. (45) for $\mathfrak{L}(K^{\mathrm{II}})$ can also be derived.

APPENDIX D

In this appendix we will give a few details of the proof of Eqs. (51)–(54) of the main text.

We begin by considering Eq. (51) and applying the left-hand side to a suitable test function $g(a_1, a_2)$:

$$\pi^{-2}\int dr\ db_1 db_2\delta((r-b_1)(r-b_2)+rb_3) \times\delta((r-c_1)(r-c_2)+rb_3)\delta((r-a_1) \times(r-a_2)+rb_3)\frac{(a_1-r)}{(a_1-b_1)_p} \times\theta(a_1-r)\frac{(b_1-r)}{(b_1-c_1)_p}\theta(r)g(a_1,a_2)da_1da_2. \quad \text{(D1)}$$

We note that the effective integration range in a_1 and a_2 is limited by the requirements in the step function and other δ functions. Thus the following inequalities must be fulfilled:

$$\begin{aligned} &a_1>0, b_3>0,\\ &a_2 = r+\frac{rb_3}{r-a_2}\\ &\quad = a_1+b_3+(r-a_1)+\frac{a_1b_3}{r-a_1} \leq (\sqrt{a_1}-\sqrt{b_3})^2;\\ &(a_1-r)(a_2-r)\leq 0,\\ &a_1 = r+\frac{rb_3}{r-a_2} = a_2+b_3+r-a_2+\frac{a_2b_3}{r-a_2}\\ &\leq(\sqrt{a_2}+\sqrt{b_3})^2 \quad \text{if } a_2>0 \end{aligned} \quad \text{(D2)}$$

Comparing the defining Equation (21) for the regions D_j and T_j, we note that the inequalities in Eq. (D2) correspond to the regions D_1 and T_2. We may consequently introduce the characteristic functions $\chi(D_1)$ and $\chi(T_2)$ for these regions without changing anything in the integral. The integrals over a_2 and b_2 may be carried out by means of the δ functions, and we are left with

$$\pi^{-2}\int dr\,db_1da_1\ \delta((r-c_1)(r-c_2)+rb_3)\theta(r)[\chi(D_1)+\chi(T_2)] \times\frac{1}{(a_1-b_1)_p}\frac{1}{(b_1-c_1)_p}g\left(a_1, r+\frac{rb_3}{r-a_1}\right) \quad \text{(D3)}$$

We further note the well-known distribution equality:

$$\frac{1}{\pi^2}\int\frac{db}{(a-b)_p(c-b)_p} = \delta(a-c). \quad \text{(D4)}$$

Application of (D4) to the b_1 integral leads to

$$-\int dr\, da_1\ \delta(a-c)\delta((r-c)(r-c)+rb_3)\theta(r)$$
$$[\chi(D)+\chi(T_2)]\, g\left(a_1, r+\frac{rb_3}{r-a_1}\right)$$
$$=-\int dr\ \theta(r)\delta((r-c_1)(r-c_2)+rb_3)[\chi(D_1(c_1,c_2,b_3))$$
$$+\chi(T_2(c_1,c_2,b_3))]g(c_1,c_2). \tag{D5}$$

The final step is to recognize that in the region $D_1(c_1, c_2, b_3)$ there are two positive solutions to the quadratic equation for r in the δ function, while in T_2 there is one positive and one negative solution. The result is then what results from an application of the right side of Eq. (60):

$$-\frac{1}{\sqrt{\lambda(b_3,c_1,c_2)}}[2\chi(D_1(c_1,c_2,b_3))$$
$$+\chi(T_2(c_1,c_2,b_3))]g(c_1c_2)$$

where $\lambda(b_3, c_1, c_2)$ is defined in Eq. (38) (note that $\lambda > 0$ in all the regions D_j and T_j).

The difference between Eq. (51) and (52) is only the step function argument. From the inequalities in Eq. (D2) we deduce that the support in this case is in the regions D_2 and T_1. The remaining steps then follow from Eqs. (D3)–(D6) with trivial modifications.

The proof of Eqs. (53) and (54) follows the same lines. We note that the step functions imply

$$b_3 = (1/\beta)a_2 + (1/\alpha)a_1$$
$$= a_1 + a_2 + (\alpha/\beta)a_2 + (\beta/\alpha)a_1 \geq (\sqrt{a_1}+\sqrt{a_2})^2$$
$$\text{if } a_1 > 0, a_2 > 0. \tag{D7}$$

This explains the appearance of the characteristic function for D_3 and T_1, T_2, respectively.

[1] B. Andersson, Commun. Math. Phys. **25**, 283 (1972);Commun. Math. Phys. **25**, 308 (1972) (DI, DII).

[2] B. Andersson, Nucl. Phys. B **30**, 413 (1971);Nucl. Phys. B **30**, 429 (1971); Nucl. Phys. B **30**, 453 (1971) (SI, SII, SII).

[3] G. Källén and A. S. Wightman, Kgl. Dan. Vidensk. Selsk. Mat.-Fys. Medd. **1**, 6 (1958).

[4] R. F. Streater and A. S. Wightman, *PCT, spin and statistics and all that* (Benjamin, New York, 1964); R. Jost, *The general theory of quantized fields* (Amer. Math. Soc., Providence, R. I., 1965)..

[5] For more details cf. the original article: A. S, Wightman, Phys. Rev. **107**, 860 (1956).

[6] A. S. Wightman, lectures, in *Relations des dispersion et particules élémentaires* (Les Houches, 1960); L. Hörmander, *An introduction to complex analysis in several variables* (Princeton U. P., Princeton, N. J., 1966).

[7] G. Källén, lectures, in *Relations des dispersion et particules élémentaires* (Les Houches, 1960).

[8] G. Källén, Nucl. Phys. **25**, 568 (1961).

[9] In appendix A the formalism is extended to the case when "subtracted dispersion relations" are necessary.

[10] H. Lehmann, K. Symanzik, and W. Zimmermann, Nuovo Cimento Lett. **1**, 205 (1955).

[11] Similar constructions have been given by, e.g., J. Bros, H. Epstein, and V. Glaser, Commun. Math. Phys. **1**, 240 (1965).

[12] G. Källén and J. Toll, Helv. Phys. Acta **33**, 753 (1960).

[13] J. Bros, H. Epstein, and V. Glaser, Commun. Math. Phys. **6**, 77 (1967).

[14] B. Andersson (to be published).

Differential solutions of the biharmonic Poisson and first order Stokes equations

J. D. Love
Department of Physics, University of Toronto, Toronto 5, Ontario, Canada
(Received 31 March 1972; revised manuscript received 6 September 1972)

For coordinate systems with rotational symmetry it is shown that particular solutions of the biharmonic Poisson and first order Stokes equations exist and can be expressed in terms of simple derivatives and algebraic functions of the corresponding solutions of the Laplace and Stokes equations.

INTRODUCTION

In mathematical physics, problems frequently arise in which it is necessary to solve either the biharmonic Poisson equation, or an equation we shall call the first order Stokes equation, together with prescribed boundary conditions. These two equations may be written as

$$\nabla^2\Phi = \phi, \quad \Delta^2\Psi = \psi, \tag{1}$$

where

$$\nabla^2\phi = 0, \quad \Delta^2\psi = 0. \tag{2}$$

Here ∇^2 and Δ^2 are the Laplace and Stokes operators, respectively. The first equation occurs, for example, in elasticity and in the theory of slow asymmetric electromagnetic waves; the latter in the Stokes flow of a viscous fluid and in the theory of slow symmetric electromagnetic waves.

Regardless of whether the boundary conditions for such problems are satisfied by eigenfunction expansions, or whether they are included in a Green's function solution, it is necessary to find particular solutions of Eqs. (1) that form complete sets in the coordinate system used. Our original motivation was the need to find such solutions in the system of toroidal coordinates. This system has one degree of symmetry, but is such that the equations of mathematical physics cannot be solved by simple separation of variables. We have, however, been able to generalize our results to include all coordinate systems possessing rotational symmetry, regardless of their separability properties.

For such symmetry systems we show how particular solutions of Eqs. (1) can be written down explicitly in terms of linear operators, involving only derivatives and simple algebraic expressions, which act on the solutions of the Laplace or Stokes equations, respectively. These solutions are such that the dependence, or independence, of the solutions of the Laplace or Stokes equations on the symmetry coordinate is preserved. Further, from a complete set of solutions of the Laplace or Stokes equations, our procedures generate complete sets of particular solutions of the respective Eqs. (1).

METHOD OF SOLUTION

The particular solutions of Eq. (1) for a general rotational coordinate system are derived in two stages. Firstly, we consider the rotational system of cylindrical polar coordinates, and construct solutions for Φ and Ψ in terms of algebraic functions and derivatives only that act linearly on ϕ and ψ, respectively. A simple modification to these particular solutions then enables us to show how they can be interpreted as particular solutions for any rotational system.

Cylindrical polar coordinates (ρ, θ, z) have the z axis as the axis of symmetry and θ as the azimuthal angle, or symmetry coordinate, that is common to all rotational systems. The solutions of the Laplace equation in these coordinates can be written, for the purposes of subsequent manipulations, in the form

$$\phi(\rho, \theta, z) \equiv \phi_0(\rho, z, \alpha)e^{i\alpha\theta}, \tag{3}$$

where α is an arbitrary constant of separation. It is the independence of the metric coefficients on θ that allows this separation and the existence of solutions of the Stokes equation, which by definition are independent of θ.

The solutions of the Laplace equation as given by Eq. (3) allow the biharmonic Poisson equation to be written as

$$\nabla^2\Phi \equiv \left(\frac{\partial^2}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} + \frac{\partial^2}{\partial z^2} - \frac{\alpha^2}{\rho^2}\right)\Phi(\rho, \theta, z) = \phi_0(\rho, z, \alpha)e^{i\alpha\theta}. \tag{4}$$

We look for a solution to Eq. (4) of the form

$$\Phi(\rho, \theta, z) = \left(A(\rho, z, \alpha)\frac{\partial}{\partial\rho} + B(\rho, z, \alpha)\frac{\partial}{\partial z} + C(\rho, z, \alpha)\right)\phi_0(\rho, z, \alpha)e^{i\alpha\theta}, \tag{5}$$

where $A, B,$ and C are arbitrary functions of ρ, z and the parameter α. Substituting Eq. (5) into Eq. (4) and recalling that $\nabla^2\phi = 0$ we obtain

$$\begin{aligned} &A_{\rho\rho}\phi_\rho - 2A_\rho\phi_{\rho\rho} + (A/\rho^2)\phi_\rho + (A_\rho/\rho)\phi_\rho + A_{zz}\phi_\rho \\ &\quad + 2A_z\phi_{\rho z} - 2\,\alpha^2/\rho^3\,A\phi_\rho + B_{\rho\rho}\phi_z + 2B_\rho\phi_{\rho z} \\ &\quad + (B_\rho/\rho)\phi_z + B_{zz}\phi_z - 2B_z\phi_{\rho\rho} + 2(\alpha^2/\rho^2)B_z\phi \\ &\quad - 2(B_z/\rho)\phi_\rho + C_{\rho\rho}\phi + 2C_\rho\phi_\rho + (C_\rho/\rho)\phi + C_{zz}\phi \\ &\quad + 2C_z\phi_z \equiv \phi, \end{aligned}$$

where the subscript ρ denotes partial differentiation with respect to ρ etc. This relationship is an identity as it must hold for all functions ϕ. Thus equating coefficients of ϕ and its derivatives we find that the functions $A, B,$ and C must satisfy the set of equations

$$C_{\rho\rho} + (C_\rho/\rho) + C_{zz} + 2(\alpha^2/\rho^2)B_z - 2\alpha^2(A/\rho^3) = 1, \tag{6}$$

$$A_{\rho\rho} + (A_\rho/\rho) + (A/\rho^2) + A_{zz} - 2(B_z/\rho) + 2C_\rho = 0, \tag{7}$$

$$B_{\rho\rho} + (B_\rho/\rho) + B_{zz} + 2C_z = 0, \tag{8}$$

$$A_z + B_\rho = 0, \tag{9}$$

$$B_z - A_\rho = 0. \tag{10}$$

We need only obtain a particular solution to this set of equations.

Equations (9) and (10) are Cauchy–Riemann and, therefore, imply

$$A_{\rho\rho} + A_{zz} = B_{\rho\rho} + B_{zz} = 0. \tag{11}$$

With this simplification Eqs. (7) and (8) reduce to

$$\frac{\partial}{\partial\rho}\left(2C - \frac{A}{\rho}\right) = \frac{\partial}{\partial z}\left(2C - \frac{A}{\rho}\right) = 0,$$

whence

$$A = \rho(2C + p), \tag{12}$$

where p is an arbitrary constant. Substituting for A in Eqs. (6) and (11) yields

$$\begin{aligned} &C_{\rho\rho} + [(4\alpha^2 + 1)/\rho]C_\rho + C_{zz} = 1, \\ &C_{\rho\rho} + 2(C_\rho/\rho) + C_{zz} = 0. \end{aligned} \tag{13}$$

Upon subtracting and integrating with respect to ρ

$$C = \frac{1}{(4\alpha^2 - 1)}\frac{\rho^2}{2} + h(z).$$

The arbitrary function $h(z)$ is determined by substituting in Eq. (13) and integrating with respect to z. Finally,

$$C = \frac{1}{(4\alpha^2 - 1)}\frac{(\rho^2 - 3z^2)}{2} + qz + r, \tag{14}$$

where q and r are arbitrary constants. From Eq. (12)

$$A = [\rho/(4\alpha^2 - 1)](\rho^2 - 3z^2) + 2q\rho z + (p + 2r)\rho. \tag{15}$$

The function B is deduced by integrating Eqs. (9) and (10) with the above form for A and comparing the two resulting expressions:

$$B = [z/(4\alpha^2 - 1)](3\rho^2 - z^2) + q(z^2 - \rho^2) + (p + 2r)z. \tag{16}$$

If we substitue Eqs. (14)–(16) in Eq. (5), it is readily verified that the following terms are harmonic:

$$[2q\rho z + (p + 2r)\rho]\frac{\partial\phi}{\partial\rho} + [q(z^2 - \rho^2) + (p + 2r)z]\frac{\partial\phi}{\partial\rho} + [qz + r]\phi.$$

Thus a particular solution of the biharmonic Poisson equation is

$$\Phi(\rho, \theta, z) = \frac{1}{(4\alpha^2 - 1)}\left(\rho(\rho^2 - 3z^2)\frac{\partial}{\partial\rho} + z(3\rho^2 - z^2)\frac{\partial}{\partial z} + \frac{(\rho^2 - 3z^2)}{2}\right)\phi_0(\rho, z, \alpha)e^{i\alpha\theta}. \tag{17}$$

This solution preserves the θ dependence of ϕ, but is invalid in the limits $\alpha \to \pm\frac{1}{2}$. For these values of α, however, the function ϕ is double-valued, and in most physical situations such solutions of the Laplace and biharmonic Poisson equations would be of no interest. Therefore, we shall not pursue solutions in these limits.

The solutions of the first order Stokes equation are obtained in cylindrical polar coordinates by applying the same technique to the equation

$$\Delta^2\Psi \equiv \left(\frac{\partial^2}{\partial\rho^2} - \frac{1}{\rho}\frac{\partial}{\partial\rho} + \frac{\partial}{\partial z^2}\right)\Psi(\rho, z) = \psi(\rho, z),$$

and assuming a form for Ψ as given by Eq. (5) with Φ and ϕ replaced by $\Psi(\rho, z)$ and $\psi(\rho, z)$, respectively. The final particular solution is then shown to be

$$\Psi(\rho, z) = \frac{1}{3}\left(\rho(\rho^2 - 3z^2)\frac{\partial}{\partial\rho} + z(3\rho^2 - z^2)\frac{\partial}{\partial z} - \frac{(\rho^2 - 3z^2)}{2}\right)\psi(\rho, z). \tag{18}$$

The particular solutions given by Eqs. (17) and (18) have been obtained in cylindrical polar coordinates. In order to extend these results to a general rotational system, we define q_1 and q_2 to be the two coordinates of the rotational system orthogonal to θ. The metric coefficients are then independent of θ, and, consequently, the whole of the above calculation is valid if $\phi(\rho, z, \alpha)$ and $\psi(\rho, z)$ are replaced by $\phi(q_1, q_2, \alpha)$ and $\psi(q_1, q_2)$, respectively. Both of the operators in Eqs. (17) and (18) are transformed from (ρ, z) to (q_1, q_2) variables by utilizing the geometrical relationships that exist between the coplanar coordinates (ρ, z) and (q_1, q_2). Consequently, the particular solutions $\Phi(q_1, q_2, \theta)$ and $\Psi(q_1, q_2)$ are expressible in terms $\phi(q_1, q_2, \theta)$ and $\psi(q_1, q_2)$, respectively, through operators containing the variables q_1 and q_2 only.

We now give an example of the application of this procedure to a coordinate system in which both the Laplace and Stokes equations cannot be solved by separation of variables.

TOROIDAL COORDINATES

The system of toroidal coordinates (η, τ, θ) is an example of a rotational system, since the metric coefficients

$$h_\eta = h_\tau = \frac{d}{(\cosh\eta - \cos\tau)}, \qquad h_\theta = \frac{d\sinh\eta}{(\cosh\eta - \cos\tau)},$$

where d is a constant, are independent of θ. Complete sets of solutions of the Laplace and Stokes equations are, respectively,

$$\begin{aligned} \phi &= (\cosh\eta - \cos\tau)^{1/2}P^m_{n-1/2}(\cosh\eta)e^{in\tau + im\theta}, \\ \psi &= \frac{\sinh\eta}{(\cosh\eta - \cos\tau)^{1/2}}P^1_{n-1/2}(\cosh\eta)e^{in\tau}, \end{aligned} \tag{19}$$

where $P^m_{n-1/2}$ is an associated Legendre function of the first kind, of degree m and half-odd integral order $n - \frac{1}{2}$. The geometrical relationships between (η, τ) and (ρ, z) coordinates are

$$\rho = \frac{d\sinh\eta}{(\cosh\eta - \cos\tau)}, \qquad z = \frac{d\sin\tau}{(\cosh\eta - \cos\tau)}.$$

From these relationships the operators in Eqs. (17) and (18) can be expressed in terms of (η, τ) coordinates. Operating on ϕ and ψ, as given by Eq. (19), yields expressions for Φ and Ψ that can be simplified by dropping complementary functions. We find

$$\begin{aligned} \Phi(\eta, \tau, \phi) = {} & [d^2/(4m^2 - 1)(\cosh\eta - \cos\tau)^{1/2}]e^{in\tau + im\phi} \\ & \times [2\sinh^2\eta P^{m\prime}_{n-1/2}(\cosh\eta) + (\cosh\eta - 2in\sin\tau) \\ & \times P^m_{n-1/2}(\cosh\eta), \end{aligned}$$

$$\Psi(\eta,\tau)=\frac{d^2\sinh^2\eta}{6}\Big(\{2\cosh\eta+\cosh\eta\cos\tau(\cosh\eta-\cos\tau)$$

$$+2in\sin\tau\,[2-3\cosh\eta$$

$$\times(\cosh\eta-\cos\tau)]\}P'_{n-1/2}(\cosh\eta)$$

$$-6\sinh^2\eta(n^2-\tfrac{1}{4})\,[2+\cos\tau(\cosh\eta-\cos\tau)]$$

$$\times P_{n-1/2}(\cosh\eta))\,e^{in\tau}/(\cosh\eta-\cos\tau)^{3/2}$$

where prime denotes differentiation with respect to $\cosh\eta$.

Position operators in a (3+1) de Sitter space

R. L. Mallett and G. N. Fleming
Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802
(Received 7 June 1972; revised manuscript received 28 July 1972)

Earlier studies of covariant position operators in special relativistic quantum theory are generalized here to the case of a de Sitter universe of positive curvature. The hyperplane formalism previously employed undergoes a natural generalization in this case to a spacelike hypersphere formalism. In the analysis of the underlying geometry, the minimal pseudo-Euclidean embedding space for the de Sitter universe plays a dominant role and suggests an intrinsic coordinate system of special interest for the representation of geodesics. The de Sitter analogs of the Minkowski space center of energy, center of spin, and center of inertia are constructed, and we find that de Sitter-center of spin to retain commuting components in our intrinsic coordinate system.

1. INTRODUCTION

Within the general formalism of the quantum theory accompanied by the specification of a particular space–time symmetry group, we define the localization problem to be the identification of those operators which can be associated with measurements designed to locate or localize some dynamical property that can be manifested by a physical system in an appropriate experimental environment. Some of the steps involved in the solution of this problem are the determination of the transformation properties of the operators under the specified symmetry group and the determination of the dynamical or structural property that is, in fact, localized. Ultimately the determination of the role these operators play in the description of interactions will be of great importance. Any more or less complete set of hypothetical answers to these aspects of the localization problem constitutes a *quantum theory of localization*. At the present, with the exception of the case in which the symmetry group is the Galilean group, there is no such theory generally accepted. This paper is intended as a contribution to a quantum theory of localization in a space–time of constant positive curvature.

There are a number of different approaches which one might take to this problem. We note, in particular, the investigations by Philips and Wigner of states localized on a circle in a (2 + 1) de Sitter space.[1] More recently, Hannabuss put forward the view, motivated by the Iwasawa decomposition, that particles should be localized with respect to horospheres in a (4 + 1) de Sitter space.[2] However, the approximations employed, or the restricted set of cases considered, or the problematic character of auxiliary hypotheses do not permit a final judgment to be reached. For us particular interest lies in determining the admissibility of analogs for finite radius of curvature R to each of the position operators studied by one of us[3] and others[4] in Minkowski space–time. The method we propose to use is based on an appeal to the minimal embedding space of the physical de Sitter universe. The embedding geometrical manifold is the five-dimensional Minkowski space $\mathfrak{M}_{1,4}$ with metric $(1, -1, -1, -1, -1)$. The group of isometries is the inhomogeneous de Sitter group $ISO(1, 4) \equiv T_5 \times SO(1, 4)$. Here T_5 is the embedding space–time translation group and $SO(1, 4)$ the homogeneous de Sitter group[5] which is the group of isometries of the *physical* de Sitter universe $\mathfrak{D}_{1,3}$.

From the outset of this study we were concerned that we associate Hermitian operators with coordinates that in the classical limit have a relation to the group of metric automorphisms on the de Sitter universe. We strongly expect, and will present arguments for our position, that an observer in a de Sitter universe evolving along a timelike geodesic and free to play with light rays and mirrors and watch otherwise non interacting particles, would *experience* the passage of time in accordance with the development of a one-parameter subgroup of the group of metric automorphisms. We have attempted to choose our coordinates accordingly. As far as the directional characteristics of our coordinates go, we have fared well. A subgroup of the metric automorphisms *does* take $x_0 = \text{const}$ slices into $x_0 = \text{const}$ slices. Furthermore we have a family of equivalent coordinate systems in which *all geodesics satisfy linear equations of the form*

$$x_i = x_i(0) + \dot{x}_i(0)x_0, \qquad i = 1, 2, 3.$$

We are far from exhaustively familiar with the relevant cosmological or even group-theoretical literature; but we have not seen these coordinates used before, and one thing that seems to mitigate against their popularity in the minds of many writers is that they lead to a non-diagonal form for the metric tensor. A *natural*[6] clock at the spatial origin of coordinates would not measure x_0 which through all eternity ranges only from $-R$ to $+R$. The proper time of the natural clock ranges from $-\infty$ to $+\infty$. Also our system has the very common property of assigning the same coordinates to two distinct points of the universe.

In Sec. 2 we discuss the classical one-sheeted de Sitter universe from the point of view of the five-dimensional Minkowski embedding space $\mathfrak{M}_{1,4}$.[7] We discuss the isometries, set up our geodesic coordinates, and note a very neat generalization of the hyperplane parameters used by one of us to parameterize observables in special relativistic quantum theory.[3]

In the last section a unitary representation of the group $ISO(1, 4)$ in a Hilbert space is introduced, and the problem of constructing the analog of the manifestly Poincaré covariant center of energy, center of spin, and center of inertia[3] is attacked. Transformation properties and commutation relations involving these operators are discussed as well as the relation between $SO(1, 4)$ and $ISO(1, 3)$ via group contraction.[8]

2. THE GEOMETRY OF THE (3+1) DE SITTER UNIVERSE

The (3 + 1) de Sitter universe is a space–time manifold of constant positive curvature originally presented by de Sitter as providing a solution to Einstein's gravitational field equations which possessed structure (non-vanishing curvature), but was yet empty of matter and energy. We are not concerned with the sometimes controversial questions surrounding this historical motivation for the de Sitter universe. For us the space–time

manifold is strictly a given model which retains the homogenous and isotropic character of Minkowski space-time while introducing a minimal element of cosmological structure, viz. curvature.

The structure of the five-dimensional Minkowski space $\mathfrak{M}_{1,4}$ is characterized by the line element

$$ds^2 = \delta_{ab} dy^a dy^b, \quad a, b = 0, 1, 2, 3, 4 \tag{2.1}$$

with

$$\delta_{ab} = (1, -1, -1, -1, -1). \tag{2.2}$$

The metric automorphisms of $\mathfrak{M}_{1,4}$ constitute the inhomogeneous de Sitter group of coordinate transformations

$$y'_a = \Lambda^b_a y_b + c_a \tag{2.3}$$

with

$$\Lambda^c_a \delta^{ab} \Lambda^d_b = \delta^{cd} \tag{2.4}$$

and c_a a constant 5-vector.

The de Sitter universe $\mathfrak{D}_{1,3}$ is conveniently represented as a four-dimensional hyperboloid of one sheet embedded in $\mathfrak{M}_{1,4}$. $\mathfrak{D}_{1,3}$ then becomes the set of points, $y_a \in \mathfrak{M}_{1,4}$, satisfying

$$y_a y^a = y_0^2 - y_1^2 - y_2^2 - y_3^2 - y_4^2 = -R^2. \tag{2.5}$$

The invariant line element in $\mathfrak{D}_{1,3}$ is (2.1) and is greater than, equal to, or less than zero for timelike, null, or spacelike infinitesimal intervals, respectively.

The metric automorphisms of $\mathfrak{D}_{1,3}$ are the linear homogeneous de Sitter transformations

$$y'_a = \Lambda^b_a y_b \tag{2.6}$$

with (2.2) and (2.4).

We adopt the standard assumptions concerning the inertial behavior of free particles and light rays, i.e., the former evolve along timelike geodesics, the latter along null geodesics. The geodesics themselves are well known to be the intersections of $\mathfrak{D}_{1,3}$ with two-dimensional planes in $\mathfrak{M}_{1,4}$ that pass through the origin $y_a = 0$. The spacelike geodesics are the intersections with spacelike planes and are closed curves along which the integral of the line element yields

$$\oint |ds| = 2\pi R. \tag{2.7}$$

This defines the sense in which $\mathfrak{D}_{1,3}$ is said to be a spatially finite universe. The null geodesics are the intersections with those planes containing null intervals, but no timelike intervals, and are straight lines in $\mathfrak{M}_{1,4}$. They are, in fact, the straight line generators of the hyperboloid of revolution $\mathfrak{D}_{1,3}$. The timelike geodesics are the intersections with planes containing timelike intervals and are all open curves. As prototypes of the three kinds of geodesics we can take

$$y_0 = y_2 = y_3 = 0, \quad y_4^2 + y_1^2 = R^2, \tag{2.8}$$

for spacelike geodesics,

$$y_2 = y_3 = 0, \quad y_4 = R, \quad y_0 \pm y_1 = 0 \tag{2.9}$$

for null geodesics, and

$$y_1 = y_2 = y_3 = 0, \quad y_4 = (R^2 + y_0^2)^{1/2} \tag{2.10}$$

for timelike geodesics. Every spacelike geodesic can be transformed into (2.8) by metric automorphisms (2.6) and (2.4); every null geodesic can be so transformed into (2.9) and every timelike geodesic can be so transformed into (2.10).

Now consider an inertial observer in $\mathfrak{D}_{1,3}$ with worldline (2.10). We contend that such an observer will judge all points with $y_0 = 0$ and $y_4 > 0$ to be simultaneous since a light ray emitted from the observers world line at proper time $-|s|$ relative to the $y_0 = 0$ point and reflected back to the observer from a point with $y_0 = 0$ will reach the observer again at a relative proper time of $+|s|$. Furthermore, since metric automorphisms in $\mathfrak{D}_{1,3}$ leave the world function between two points invariant

$$S(y^{(2)}, y^{(1)}) = S(y^{(2)'}, y^{(1)'}), \tag{2.11}$$

we conclude that any metric automorphism that leaves the worldline (2.10) invariant takes the set of points with $y_0 = 0$ and $y_4 > 0$ into a set of points that again appear simultaneous to the observer. Among these automorphisms are the rotations in the (y_1, y_2) or (y_2, y_3) or (y_3, y_1) planes, which "appear" like spatial rotations about the point $y_0 = y_1 = y_2 = y_3 = 0$, $y_4 = R$ to the observer in $\mathfrak{D}_{1,3}$, and the pseudo-rotations in the (y_4, y_0) plane which "appear" like time translations.

The remaining planar metric automorphisms alter the world line of the observer; but they still transform the $y_0 = 0$, $y_4 > 0$ points into points that appear simultaneous to the inertial observer on the transformed world line. These automorphisms are the rotations in the planes (y_1, y_4) or (y_2, y_4) or (y_3, y_4) which "appear" like spatial translations and the pseudo-rotations in the planes (y_1, y_0) or (y_2, y_0) or (y_3, y_0) which "appear" like Lorentz transformations.

In the preceding discussion the original set of points $y_0 = 0$ was restricted by the condition $y_4 > 0$ because the points with $y_0 = 0$ and $y_4 < 0$ are beyond the well-known light horizon[7] of the world line (2.10). More generally any point in $\mathfrak{D}_{1,3}$ with

$$y_4 < -y_0 \tag{2.12}$$

is beyond the *active* light horizon of (2.10), i.e., no light ray can reach such a point from (2.10). Correspondingly any point in $\mathfrak{D}_{1,3}$ with

$$y_4 < y_0 \tag{2.13}$$

is beyond the *passive* light horizon of (2.10), i.e., no light ray can reach (2.10) from such a point.

Now the equation for the original set of simultaneous points for the world line (2.10)

$$y_0 = 0 \tag{2.14}$$

can be written in the invariant form

$$y^a \eta^{(0)}_a = 0, \tag{2.15}$$

where the 5-vector $\eta^{(0)}_a$ is

$$\eta^{(0)}_a \equiv (1, 0, 0, 0, 0). \tag{2.16}$$

Under the active metric automorphism Λ^b_a, this set of points is mapped into the set satisfying

$$y^a\eta_a = 0, \tag{2.17}$$

where

$$\eta_a = \Lambda_a^b \eta_b^{(0)}. \tag{2.18}$$

In $\mathfrak{M}_{1,4}$ Eqs. (2.14), (2.15), and (2.17) define four-dimensional spacelike hyperplanes, and in $\mathfrak{D}_{1,3}$ these same equations determine the intersections of these hyperplanes with $\mathfrak{D}_{1,3}$ which intersections are three-dimensional spacelike hyperspheres S_3 of radius R in $\mathfrak{D}_{1,3}$. The hypersphere (2.17) "appears" instantaneous to that inertial observer whose world line is the transform under Λ_a^b of (2.10). To the original observer on (2.10) this same hypersphere "appears" noninstantaneous; but if he likes he may still use this hypersphere as a domain for stipulating initial conditions. Since these hyperspheres are uniquely determined by the timelike unit 5-vector η_a that enters in the equation of the hypersphere, we shall henceforth refer to them as the η_a hyperspheres.[9] Notice that in the reduced model $\mathfrak{D}_{1,1}$ the instantaneous hypersphere S_3 becomes the circle S_1 where states may be localized in the sense of Philips and Wigner.[1]

With this discussion of the geometry and kinematics of $\mathfrak{D}_{1,3}$ viewed as embedded in $\mathfrak{M}_{1,4}$ behind us, we are ready to introduce intrinsic coordinates for $\mathfrak{D}_{1,3}$. We want these coordinates to be as intimately related to the geodesics in $\mathfrak{D}_{1,3}$ as possible since, as our discussion has emphasized, we regard the geodesic structure as all important in determining the way the de Sitter universe "appears" to an inertial observer. On this basis we have found the coordinates

$$x_\mu \equiv R(y_\mu/y_4), \qquad \mu = 0, 1, 2, 3 \tag{2.19}$$

to be very convenient. The main reason for this is seen by noting that the 4-vector equations for an arbitrary geodesic in $\mathfrak{D}_{1,3}$ are[10]

$$\mathbf{y}(y_0) = \mathbf{y}(0)\{1 + [1 - \dot{\mathbf{y}}(0)^2](y_0^2/R^2)\}^{1/2} + \dot{\mathbf{y}}(0)y_0, \tag{2.20}$$

where

$$\dot{\mathbf{y}} \equiv \frac{d\mathbf{y}}{dy_0}, \tag{2.21a}$$

$$\mathbf{y} = (y_1, y_2, y_3, y_4), \tag{2.21b}$$

$$\mathbf{y}(0)\cdot\dot{\mathbf{y}}(0) = 0, \tag{2.22a}$$

and

$$\mathbf{y}(0)^2 = R^2. \tag{2.22b}$$

Expressed in terms of the independent coordinates x_μ, (2.20) becomes

$$x_i(x_0) = \dot{x}_i(0)x_0 + x_i(0) \qquad i = 1, 2, 3, \tag{2.23}$$

where

$$x_i(0) = \frac{y_i(0)}{y_4(0)}R \tag{2.24a}$$

and

$$\dot{x}_i(0) \equiv \frac{dx_i(0)}{dx_0} = \left(\dot{y}_i(0) - y_i(0)\frac{\dot{y}_4(0)}{y_4(0)}\right). \tag{2.24b}$$

In short *all* the geodesics in $\mathfrak{D}_{1,3}$ satisfy the equation

$$\frac{d^2x_i(x_0)}{dx_0^2} = 0, \tag{2.25}$$

an equation of the same form as that for geodesics in Galilean or Minkowski space-time.

Under the metric automorphisms (2.6) the independent coordinates x_μ transform in accordance with

$$x'_\mu = \frac{\Lambda_\mu^\nu x_\nu + R\Lambda_\mu^4}{R^{-1}\Lambda_4^\nu x_\nu + \Lambda_4^4}. \tag{2.26}$$

In particular, the pseudo-rotations in the (y_0, y_4) plane take the points with $x_0 = 0$ into the points with

$$x'_0 = R\frac{\Lambda_0^4}{\Lambda_4^4} \tag{2.27}$$

and

$$x'_i = \frac{x_i}{\Lambda_4^4}. \tag{2.28}$$

In these coordinates the transformation appears like a time displacement accompanied by a space dilation. In fact the dilation is an apparent *contraction* since from (2.4)

$$\Lambda_4^4 = [1 + (\Lambda_0^4)^2]^{1/2} > 1. \tag{2.29}$$

But if we recall (2.11) we conclude that the invariant interval along an instantaneous geodesic between $x_\mu^{(1)} = (0, x_i^{(1)})$ and $x_\mu^{(2)} = (0, x_i^{(2)})$ is equal to the invariant interval along the corresponding instantaneous geodesic between $x_\mu'^{(1)} = (x'_0, x_i'^{(1)})$ and $x_\mu'^{(2)} = (x'_0, x_i'^{(2)})$. Consequently a free particle moving along the timelike geodesic $x_i = c_i$ *increases* its *invariant* separation from the origin during the time interval from $x_0 = 0$ to $x_0 = R\Lambda_0^4(\Lambda_4^4)^{-1}$. On this basis one can deduce the familiar *expansion* of the de Sitter universe.

Notwithstanding the simple form geodesics take in our intrinsic coordinate system, we should be careful in reading metrical properties of the space-time continuum from equations involving the coordinates. Thus while a clock at rest with our inertial observer ranges from a proper time reading of $-\infty$ to $+\infty$, the coordinate x_0 ranges only from $-R$ to $+R$. At the same time the spatial coordinates x_i range from $-\infty$ to $+\infty$ over one half of what we know to be a spatially finite universe. The important consequence of this whole discussion for us is that for any inertial observer the "evolution" of a physical system may be discussed in terms of the transitions of the spacelike configurations of the system as one considers first one spacelike hypersphere $\eta_a^{(1)}$ and then another $\eta_a^{(2)}$. The introduction of the 5-vector η_a enables the observer to treat all spacelike hyperspheres equivalently and, thereby, provides a manifestly covariant (under the group of metric automorphisms) description of the timelike evolution of a physical system. Upon the introduction of the dynamical variable the observer will parametrize them with the timelike unit vector η_a. Thus, η_a, or hypersphere, dependence becomes the appropriate generalization in the de Sitter universe of time dependence in the Galilean universe or hyperplane dependence in the Minkowski universe.[3]

Finally, although no use will be made of it here, we give the line element and metric tensor in terms of our coordinates x_μ

$$ds^2 = \left(1 - \frac{x^2}{R^2}\right)^{-1}\left(dx_\mu dx^\mu + \frac{(x_\nu dx^\nu)^2}{R^2 - x^2}\right), \tag{2.30}$$

$$g_{\mu\nu}(x) = \left(1 - \frac{x^2}{R^2}\right)^{-1}\left(\delta_{\mu\nu} + \frac{x_\mu x_\nu}{R^2 - x^2}\right). \tag{2.31}$$

3. ELEMENTARY SYSTEMS IN THE DE SITTER UNIVERSE AND POSITION OPERATORS

In the state space of a fictitious quantum mechanical system in $\mathfrak{M}_{1,4}$, the metric automorphisms on $\mathfrak{M}_{1,4}$ which constitute the group $ISO(1,4)$ are represented by unitary generators. Denoting the Hermitian generators of this unitary representation of $ISO(1,4)$ by P_a and J_{ab}, we have the commutation relations

$$[J_{ab}, J_{cd}] = i\hbar(\delta_{ad}J_{bc} - \delta_{ac}J_{bd} + \delta_{bc}J_{ad} - \delta_{bd}J_{ac}), \tag{3.1}$$

$$[P_c, J_{ab}] = -i\hbar(\delta_{cb}P_a - \delta_{ca}P_b), \tag{3.2}$$

$$[P_a, P_b] = 0, \tag{3.3}$$

where

$$J_{ab} = -J_{ba}, \quad a, b = 0, 1, 2, 3, 4. \tag{3.4}$$

For ease of comparison with the Poincaré group we redefine $J_{4\mu}$ and P_4 by

$$RB_\mu \equiv J_{4\mu}, \quad S/R \equiv P_4, \quad \mu = 0, 1, 2, 3. \tag{3.5}$$

The B_μ in $\mathfrak{D}_{1,3}$ are the curved space analogs of space-time translations. A partial interpretation of the operator S is achieved by observing that

$$S/R = (P_\mu P^\mu - \mathfrak{M}^2)^{1/2}, \tag{3.6}$$

where $\mathfrak{M}^2 \equiv P_a P^a$ is a Casimir invariant of $ISO(1,4)$. Now (3.1)–(3.3) can be rewritten as

$$[B_\mu, B_\nu] = i\hbar J_{\mu\nu}R^{-2}, \tag{3.7}$$

$$[B_\lambda, J_{\mu\nu}] = i\hbar(\delta_{\lambda\mu}B_\nu - \delta_{\lambda\nu}B_\mu), \tag{3.8}$$

$$[J_{\sigma\lambda}, J_{\mu\nu}] = i\hbar(\delta_{\sigma\nu}J_{\lambda\mu} - \delta_{\sigma\mu}J_{\lambda\nu} + \delta_{\lambda\mu}J_{\sigma\nu} - \delta_{\lambda\nu}J_{\sigma\mu}), \tag{3.9}$$

$$[P_\lambda, J_{\mu\nu}] = i\hbar(\delta_{\lambda\mu}P_\nu - \delta_{\lambda\nu}P_\mu), \tag{3.10}$$

$$[S, J_{\mu\nu}] = 0, \tag{3.11}$$

$$[B_\mu, P_\nu] = i\hbar\delta_{\mu\nu}SR^{-2}, \tag{3.12}$$

$$[B_\mu, S] = i\hbar P_\mu, \tag{3.13}$$

$$[P_\mu, P_\nu] = 0, \tag{3.14}$$

$$[P_\mu, S] = 0. \tag{3.15}$$

The metric automorphisms of the *physical* universe $\mathfrak{D}_{1,3}$ correspond to the subgroup $SO(1,4)$. The Lie algebra of the homogeneous de Sitter group $SO(1,4)$ is given by the relations (3.7)–(3.9). The generators P_μ of spacetime translations in the embedding space $\mathfrak{M}_{1,4}$ satisfy the relations (3.14), while $J_{\mu\nu}$ clearly generates a Lorentz subgroup of $ISO(1,4)$. Under the contraction defined by $R \to \infty$ the motions generated by the B_μ go over into those generated by the generators of translations in Minkowski space.

The difference between $SO(1,4)$ and the Poincaré group $ISO(1,3)$ is that in the present case space-time translations do not commute. This fact introduces a subtlety into the interpretation of the B_μ. While B_1, say, does generate a translation of x_1 starting from $x_2 = x_3 = x_0 = 0$ (a finite transformation of x_1 from this point yields

$$x_1' = \frac{\cos\theta x_1 + \sin\theta R}{\cos\theta - [(\sin\theta x_1)/R]} = \frac{x_1 + \tan\theta R}{1 - \tan\theta(x_1/R)}$$

$x_2' = x_3' = x_0' = 0$), the same transformation does not leave x_2, x_3, x_0 invariant if they are different from zero. Thus, if one first performs a translation along x_2 from the origin or translates the time variable x_0 by a finite amount, subsequent translations of x_1 are generated by a B_1' obtained from B_1 by appropriate transformations. This is all very familiar in the case of mixing rotations of Lorentz transformations with translations. In the case of translations themselves, however, it is a peculiarity of curved space.

We now turn to the problem of constructing within an irreducible representation of $ISO(1,4)$ a position operator which can be a quantum-mechanical representative for the coordinate vector y_a. The esential ingredients that go into our construction are the following.

(i) Within an irreducible representation space of $ISO(1,4)$ the operators at our disposal are the J_{ab} and P_a.

(ii) We consider building the Y_a from $ISO(1,4)$ generators as though they were spacelike hyperplane dependent position operators in the embedding space. The points in a spacelike hyperplane in the embedding space satisfy an equation of the form

$$\eta^a y_a = \tau, \quad a = 0, 1, 2, 3, 4, \tag{3.16}$$

where

$$\eta_a\eta^a = 1, \quad \eta_0 > 1. \tag{3.17}$$

Consequently each hyperplane is uniquely characterized by a pair (η_a, τ). Under an inhomogeneous de Sitter transformation the hyperplane parameters η_a and τ satisfy

$$\eta_a' = \Lambda_a^b\eta_b, \quad \tau' = \tau + c^a\Lambda_a^b\eta_b. \tag{3.18}$$

With these notational conventions adopted the construction of each of the 5-vector position operators is straightforward.[11]

(iii) The $Y_a(\eta, \tau)$ operators must satisfy the correspondence principle in the expectation value sense, i.e.,

$$\langle\Psi'|Y_a(\eta', \tau')|\Psi'\rangle = \Lambda_a^b\langle\Psi|Y_b(\eta, \tau)|\Psi\rangle + c_a\langle\Psi|\Psi\rangle. \tag{3.19}$$

(iv) Then as a consequence of (2.17) we set $\tau = 0$ or, equivalently, impose the constraint

$$\eta^a Y_a(\eta) = 0. \tag{3.20}$$

Motivated by this point of view, we now *define* the *hypersphere dependent* position operators in de Sitter space to be $\tilde{Y}_a(\eta) = (Y_\mu(\eta), \tilde{Y}_4(\eta))$, where the $Y_\mu(\eta)$ are simply determined by (i)–(iv). However, we take as the fifth

component the operator

$$\tilde{Y}_4(\eta) \equiv [R^2 + Y_\mu(\eta)Y^\mu(\eta)]^{1/2}. \tag{3.21}$$

This requires some comment. It should by now be reasonably clear why we have adopted this point of view, but what of the transformation properties of the operators $\tilde{Y}_a$ defined in this way? The transformation properties of the Y_μ follow at once from (iii) and (iv) and are given by

$$U^{-1}(\Lambda)Y_\mu U(\Lambda) = \Lambda^\nu_\mu Y_\nu + \Lambda^4_\mu(\eta^\rho Y_\rho/\eta_4), \tag{3.22}$$

where $U(\Lambda)$ denotes the unitary transformation associated with a homogeneous de Sitter transformation. In order to ascertain the $\tilde{Y}_4$ transform set $\tilde{Y}_4 = f(Y_\mu)$, then

$$\begin{aligned} U^{-1}(\Lambda)\tilde{Y}_4 U(\Lambda) &= U^{-1}(\Lambda)f(Y_\mu)U(\Lambda) \\ &= f(U^{-1}(\Lambda)Y_\mu U(\Lambda)) \\ &= f(\Lambda^\nu_\mu Y_\nu + \Lambda^4_\mu \eta^\rho Y_\rho/\eta^4), \end{aligned}$$

from which it immediately follows that for (3.21) we have

$$U^{-1}(\Lambda)\tilde{Y}_4 U(\Lambda) = [(\Lambda^\nu_\mu Y_\nu + \Lambda^4_\mu \eta^\rho Y_\rho/\eta_4)^2 + R^2]^{1/2}. \tag{3.23}$$

In order to avoid cumbersome notation we will drop the tilde for the remainder of the paper and assume it to be understood unless otherwise indicated.

The way is now clear, and we may proceed at once to a discussion of hypersphere dependent position operators. In order to make sure that the reader has a perfectly clear understanding of the rules according to which each of the position operators are to be constructed, we will give a brief sketch of the center of energy. It follows at once from (i)–(iii) that

$$Y^{\text{c.e.}}_a(\eta\tau) = \tau P_a/\eta P + J_{ab}\eta^b : (\eta P)^{-1}, \quad a, b = 0, 1, 2, 3, 4, \tag{3.24}$$

where

$$A : B = \tfrac{1}{2}(AB + BA) \tag{3.25}$$

and

$$\eta P = \eta^a P_a. \tag{3.26}$$

Then as a result of (iv) we readily obtain

$$Y^{\text{c.e.}}_a(\eta) = J_{ab}\eta^b : (\eta P)^{-1} \tag{3.27}$$

with the obvious result that

$$Y^{\text{c.e.}}_\mu(\eta) = J_{\mu b}\eta^b : (\eta P)^{-1}, \quad \mu = 0, 1, 2, 3. \tag{3.28}$$

This latter result is the one we wished to obtain and may be rewritten as

$$Y^{\text{c.e.}}_\mu(\eta) = \sigma B_\mu : (\eta P)^{-1} + J_{\mu\nu}\eta^\nu : (\eta P)^{-1}, \tag{3.29}$$

where

$$\sigma/R \equiv \eta_4 \tag{3.30}$$

and

$$\eta P = \eta^\mu P_\mu - \sigma S/R^2. \tag{3.31}$$

In the flat space limit the hypersphere parameter σ assumes the role of the time parameter on instantaneous (i.e., constant time) hyperplanes. The notation c.e. denotes that (3.29) is the desired analog of the center of energy in the flat space limit. However, $Y^{\text{c.e.}}_4(\eta)$ is now to be defined by (3.21).

Proceeding in a similar manner, we obtain the hypersphere generalization of the center of inertia

$$Y^{\text{c.i.}}_\mu(\eta) = Y^{\text{c.e.}}_\mu(\eta) - S_{\mu\nu}\eta^\nu/\eta P - \sigma\Sigma_\mu/\eta P \tag{3.32}$$

with

$$Y_4(\eta) \equiv [R^2 + Y_\mu(\eta)Y^\mu(\eta)]^{1/2}, \tag{3.33}$$

where the "spin tensor" $S_{\mu\nu}$ and "momentum spin" Σ_μ terms are defined by

$$\begin{aligned} S_{\mu\nu} \equiv J_{\mu\nu} &- J_{\mu\lambda} : (P^\lambda P_\nu/P^2) - B_\mu : (SP_\nu/P^2) \\ &+ B_\nu : (SP_\mu/P^2) + J_{\nu\lambda} : (P^\lambda P_\mu/P^2) \end{aligned} \tag{3.34}$$

and

$$\begin{aligned} \Sigma_\mu = S_{4\mu}/R \equiv B_\mu &- B_\lambda : (P^\lambda P_\mu/P^2) \\ &+ B_\mu/R^2 : (S^2/P^2) + J_{\mu\lambda}/R^2 : (P^\lambda S/P^2), \end{aligned} \tag{3.35}$$

with

$$P^2 = P_\mu P^\mu - S^2/R^2 \tag{3.36}$$

In the flat space limit the center of inertia is equal to the center of energy on instantaneous hyperplanes in the rest frame.

The hypersphere generalization of the center of spin is given by

$$Y^{\text{c.s.}}_\mu(\eta) = Y^{\text{c.i.}}_\mu(\eta) + S_{\mu\nu}\eta^\nu/(\eta P + \mathfrak{M}) + \sigma\Sigma_\mu/(\eta P + \mathfrak{M}) \tag{3.37}$$

with (3.33) and where

$$\mathfrak{M}^2 \equiv P_\mu P^\mu - S^2/R^2. \tag{3.38}$$

In the limit $R \to \infty$, $\mathfrak{M}^2$ reduces to the special relativistic mass-squared operator. Furthermore,

$$\lim_{R\to\infty} \Sigma_\mu = 0. \tag{3.39}$$

Consequently, in the flat space limit Eqs. (3.32) and (3.37) reduce to the hyperplane generalizations of the center of inertia and the unique self-commuting Newton–Wigner position operators respectively.[3,4] The commutators of these quantities are

$$\begin{aligned} [\Sigma_\mu, \Sigma_\lambda] = i\hbar[&(SP_\lambda/P^2)\Sigma_\mu/R^2 \\ &- (SP_\mu/P^2)\Sigma_\lambda/R^2 - (P_\sigma P^\sigma/P^2)S_{\lambda\mu}/R^2], \end{aligned} \tag{3.40}$$

$$\begin{aligned} [S_{\mu\nu}, S_{\lambda\rho}] = i\hbar[&(\delta_{\nu\rho} - P_\nu P_\rho/P^2)S_{\lambda\mu} \\ &- (\delta_{\nu\lambda} - P_\nu P_\lambda/P^2)S_{\rho\mu} + (\delta_{\mu\lambda} - P_\mu P_\lambda/P^2)S_{\rho\nu} \\ &- (\delta_{\mu\rho} - P_\mu P_\rho/P^2)S_{\lambda\nu}], \end{aligned} \tag{3.41}$$

$$\begin{aligned} [\Sigma_\mu, S_{\nu\lambda}] = i\hbar[&(\delta_{\mu\nu} - P_\mu P_\nu/P^2)\Sigma_\lambda \\ &- (\delta_{\mu\lambda} - P_\mu P_\lambda/P^2)\Sigma_\nu + (SP_\nu/P^2)S_{\mu\lambda}/R^2 \\ &- (SP_\lambda/P^2)S_{\mu\nu}/R^2]. \end{aligned} \tag{3.42}$$

Now using Eqs. (3.29)–(3.42) we find

$$\begin{aligned} [Y^{\text{c.e.}}_\mu(\eta), Y^{\text{c.e.}}_\lambda(\eta)] = -[i\hbar/(\eta P)^2][&S_{\mu\lambda} - S_{\mu\sigma}\eta^\sigma(P_\lambda/\eta P) \\ &- \sigma\Sigma_\mu(P_\lambda/\eta P) + S_{\lambda\sigma}\eta^\sigma(P_\mu/\eta P) \\ &+ \sigma\Sigma_\lambda(P_\mu/\eta P)], \end{aligned} \tag{3.43}$$

$$\begin{aligned} [Y^{\text{c.i.}}_\mu(\eta), Y^{\text{c.i.}}_\lambda(\eta)] = (i\hbar/P^2)[&S_{\mu\lambda} - S_{\mu\sigma}\eta^\sigma(P_\lambda/\eta P) \\ &- \sigma\Sigma_\mu(P_\lambda/\eta P) + S_{\lambda\sigma}\eta^\sigma(P_\mu/\eta P) \\ &+ \sigma\Sigma_\lambda(P_\mu/\eta P)], \end{aligned} \tag{3.44}$$

and

$$[Y_\mu^{c.s.}(\eta), Y_\lambda^{c.s.}(\eta)] = 0. \tag{3.45}$$

Note that the hypersphere generalization of the center of spin still retains commuting components. More will be said of this result later. Here we wish to state that while we recognize the mathematical utility of self-commuting position operators we would call for increased attention in the direction of "nonself-commuting" position operators such as the center of energy and center of inertia. There is, for example, no *a priori* reason for expecting that the independent measurements required for the localization of a dynamical property on any hypersphere will be compatible.

Let us now turn to a discussion of the equations of motion. The η_μ and σ derivatives which occur in these equations are complicated by the constraint (3.17). Hence we have found it convenient to define the restricted η_a derivative by

$$\frac{\delta}{\delta\eta^a} = \frac{\partial}{\partial\eta^a} - \eta_a\eta^c \frac{\partial}{\partial\eta^c}, \qquad a, c = 0, 1, 2, 3, 4. \tag{3.46}$$

Expressed in terms of η_μ and σ these derivatives are

$$\frac{\delta}{\delta\sigma} = \frac{\partial}{\partial\sigma} + \frac{\sigma}{R^2}\eta^\lambda \frac{\partial}{\partial\eta^\lambda} + \frac{\sigma^2}{R^2}\frac{\partial}{\partial\sigma} \qquad \lambda = 0, 1, 2, 3 \tag{3.47}$$

and

$$\frac{\delta}{\delta\eta^\mu} = \frac{\partial}{\partial\eta^\mu} - \eta_\mu\eta^\lambda \frac{\partial}{\partial\eta^\lambda} - \eta_\mu\sigma\frac{\partial}{\partial\sigma} \quad \mu,\lambda = 0, 1, 2, 3. \tag{3.48}$$

The constraint (3.17) can also be rewritten as

$$\eta_\mu\eta^\mu - \sigma^2/R^2 = 1. \tag{3.49}$$

Thus, in the limit $R \to \infty$ we have

$$\frac{\delta}{\delta\sigma} \to \frac{\partial}{\partial\sigma}, \tag{3.50}$$

$$\frac{\delta}{\delta\eta^\mu}\Big|_\sigma \to \frac{\partial}{\partial\eta^\mu} - \eta_\mu\eta^\lambda\frac{\partial}{\partial\eta^\lambda}, \tag{3.51}$$

and

$$\eta_\mu\eta^\mu \to 1, \tag{3.52}$$

so that we recover the usual hyperplane derivatives and constraint equation in Minkowski space.[3]

Now it will be recalled from (3.22) that $Y_\mu(\eta)$ transforms like

$$U(\Lambda)Y_\mu(\Lambda\eta)U^{-1}(\Lambda) = \Lambda_\mu^\nu Y_\nu(\eta) + a_\mu\eta^\rho Y_\rho(\eta)/\sigma \tag{3.53}$$

under a homogeneous de Sitter transformation where

$$\Lambda_\mu^4 = a_\mu/R. \tag{3.54}$$

Consequently for infinitesimal transformations

$$U(\delta_a^b + \delta\omega_a^b) = I - (i/2\hbar)J_{ab}\delta\omega^{ab} \tag{3.55}$$

each of the position operators satisfy the generalized Heisenberg equations of motion

$$[B_\mu, Y_\lambda(\eta)] = i\hbar\delta_{\mu\lambda}\frac{\eta^\rho Y_\rho(\eta)}{\sigma} - i\hbar\left(\eta_\mu\frac{\delta Y_\lambda(\eta)}{\delta\sigma} + \frac{\sigma}{R^2}\frac{\delta Y_\lambda(\eta)}{\delta\eta^\mu}\right) \tag{3.56}$$

and

$$[J_{\mu\nu}, Y_\lambda(\eta)] = i\hbar(\delta_{\nu\lambda}Y_\mu(\eta) - \delta_{\mu\lambda}Y_\nu(\eta)) + i\hbar\left(\eta_\nu\frac{\delta Y_\lambda(\eta)}{\delta\eta^\mu} - \eta_\mu\frac{\delta Y_\lambda(\eta)}{\delta\eta^\nu}\right). \tag{3.57}$$

The interpretation of terms such as $\delta Y_\lambda(\eta)/\delta\sigma$, $\delta Y_\lambda(\eta)/\delta\eta^\mu$, and $\eta^\rho Y_\rho(\eta)/\sigma$ depends on the dynamics of the system and the definition of $Y_\lambda(\eta)$. In all cases we find that

$$\lim_{R\to\infty}[\eta^\rho Y_\rho(\eta)/\sigma] = 1 \tag{3.58}$$

which is equivalent to the constraint satisfied by hyperplane dependent position operators in the flat space limit.[3]

Turning to (3.56) and taking the limit $R \to \infty$, we have

$$[B_\mu, Y_\lambda(\eta)] \to i\hbar\left(\delta_{\mu\lambda} - \eta_\mu\frac{\partial Y_\lambda(\eta)}{\partial\sigma}\right). \tag{3.59}$$

If we set

$$P_\mu = \lim_{R\to\infty} B_\mu \tag{3.60}$$

and

$$Y_\mu(\eta_\mu, \sigma) = \lim_{R\to\infty} Y_\mu(\eta_a), \tag{3.61}$$

then (3.59) can be rewritten as

$$[P_\mu, Y_\lambda(\eta_\mu, \sigma)] = i\hbar\left(\delta_{\mu\lambda} - \eta_\mu\frac{\partial Y_\lambda}{\partial\sigma}(\eta_\mu, \sigma)\right). \tag{3.62}$$

The physical interpretation of (3.62) is readily accomplished by noting that for instantaneous hyperplanes $\eta_\mu = (1, 0, 0, 0)$, (3.62) reduces to

$$[P_i, Y_j(\sigma)] = i\hbar\delta_{ij}, \qquad i, j = 1, 2, 3, \tag{3.63}$$

$$[P_0, Y_j(\sigma)] = i\hbar\frac{\partial Y_j(\sigma)}{\partial\sigma}, \tag{3.64}$$

$$[P_\mu, Y_0(\sigma)] = 0. \tag{3.65}$$

Thus, Eq. (3.63) is the canonical quantization of the spatial momentum and position, and Eq. (3.64) is the Heisenberg equation of motion which determines the change of the position in the Heisenberg representation with time.

As we mentioned earlier the de Sitter space analog of the center of spin retains commuting components. Consequently, if as the analog of the operator corresponding to the intrinsic coordinates (2.19) we take the operator function $RY_\mu(\eta) : Y_4^{-1}(\eta)$, then for the case of $Y_a^{c.s.}(\eta)$ we obtain the result

$$[X_\mu^{c.s.}(\eta), X_\nu^{c.s.}(\eta)] = 0. \tag{3.66}$$

Most of the effort that has gone into studies of the localization problem in Minkowski space have focused attention on finding localized states, which would be simultaneous eigenstates of the Cartesian components of the position operator, or in some other way singling out position operators with self-commuting components. Consequently, we felt the need to indicate the existence of such an operator in a coordinate system of de Sitter space for which the geodesics assume a particularly simple form.

4. CONCLUSION

Making use of the minimal pseudo-Euclidean embedding space of a Riemannian space–time of constant curvature and its associated group of rigid motions, we established the connection between the hypersphere parameters used to denote spacelike slices of de Sitter space and the hyperplane parameters used by one of us in studies of the localization problem in Minkowski space.[3] Arguments have been presented to justify the interpretation of spacelike hyperspheres in a de Sitter universe as appearing instantaneous to some inertial observer. We found a set of coordinate systems that were useful in the localization problem. They possess the property of describing all geodesics by linear equations. This is achieved at the expense of introducing a nondiagonal metric tensor. Finally, we identified and interpreted some position operators for a quantum mechanical system in a de Sitter space of positive curvature. The position operators considered, reduce to operators that have been extensively studied in the flat space limit. It was noted that there does exist a self-commuting position operator in de Sitter space.

We did not look for detailed quantitative differences with the corresponding results in Minkowski space since a plausible radius of curvature would be so large as to render such differences insensible. Rather we were interested in the qualitative conceptual differences that are required for interpreting the localization problem in a spatially finite universe of constant space–time curvature.

Note added in proof: In a paper by Aghassi, Roman, and Santilli, J. Math. Phys. **11**, 2297 (1970), the inhomogeneous de Sitter group ISO(3, 2) was used to generate position operators for special relativistic quantum theory by the contraction to G_5, a five-dimensional Galilean group. The principal differences with the present work lie in the use of the group of motions in a 3 + 2 embedding space, the imposition of physical interpretation only after contraction, and the use of proper time dependence in place of hyperplane dependence.

---

[1] T. O. Philips and E. P. Wigner, *Group theory and its applications*, edited by M. Loebl (Academic, New York, 1968), p. 631.

[2] K. C. Hannabuss, Proc. Camb. Philos. Soc. **70**, 283 (1971). In a reduced (i.e., three-dimensional) model of de Sitter space, a horosphere can be visualized as a constant time parabola characterized by a null vector.

[3] G. N. Fleming, Phys. Rev. B **137**, 188 (1965); G. N. Fleming, Phys. Rev. B **139**, 963 (1965); G. N. Fleming, J. Math. Phys. **7**, 1959 (1966).

[4] M. H. L. Pryce, Proc. R. Soc. A **195**, 62 (1948); T. D. Newton and E. P. Wigner, Rev. Mod. Phys. **21**, 400 (1949). Further references can be found in Ref. 3.

[5] The literature in this area is extensive. See, for example, F. Gürsey, in *Group theoretical concepts and methods in elementary particle physics*, edited by F. Gürsey (Gordon and Breach, New York, 1964), p. 365; O. Nachtman, Commun. Math. Phys. **6**, 1 (1967); G. Borner and H. P. Dürr, Nuovo Cimento A **64**, 669 (1969).

[6] By natural clock is meant one that measures proper time.

[7] E. Schrödinger, *Expanding universes* (Cambridge U. P., Cambridge, 1956). There $\mathfrak{M}_{1,4}$ is designated by the pseudo-Euclidean R_5.

[8] E. Inönü, in *Group theoretical concepts and methods in elementary particle physics*, edited by F. Gürsey (Gordon and Breach, New York, 1964), p. 365.

[9] We note that the same number of independent parameters are needed to specify a hyperplane in Minkowski space–time as are needed to specify a hypersphere in de Sitter space–time, see G. N. Fleming, Ref. 3.

[10] A similar equation may be found in J. L. Synge, *Relativity, the general theory* (North-Holland, Amsterdam, 1960), p. 263.

[11] The reader is referred to G. N. Fleming, Phys. Rev. B **137**, 188 (1965) for the detailed and completely analogous treatment of each of the position operators to be studied in the embedding space.

Kerr geometry as complexified Schwarzschild geometry

Menahem M. Schiffer
Stanford University, Stanford, California 94305

Ronald J. Adler*, James Mark†, and Charles Sheffield
The American University, Washington, D.C. 20016
(Received 6 June 1972; revised manuscript received 6 September 1972)

We present a simple derivation of the Schwarzschild and Kerr geometries by simplifying the Einstein free space field equations for the algebraically special form of metric studied by Kerr. This results in a system of two partial differential equations, the Laplace and eikonal equations, for a complex generating function. The metric tensor is a simple explicit functional of this generating function. The simplest solution generates the Schwarzschild geometry, while a displacement of the origin by *ia* in this solution generates the Kerr geometry.

1. INTRODUCTION

The Schwarzschild geometry and the Kerr geometry[1] are very fundamental in general relativity theory and have numerous applications in experimental relativity and astrophysics.[2] It is clearly desirable to have as simple a derivation as possible of these geometries and to understand fully their relation to each other. From the beginning it has been known that the parameter "a" of the Kerr geometry is a measure of the specific angular momentum of the source of the field and that for the case $a = 0$ the axially symmetric Kerr geometry reduces to the spherically symmetric Schwarzschild geometry.[1] In this work we will give a mathematical interpretation of the parameter a. We will first reduce the Einstein free space field equations for an algebraically special form of metric to a simple system of two partial differential equations, the Laplace and Eikonal equations, for a single complex function. The metric tensor is a simple explicit functional of this complex function, which we may call the generating function. The most obvious solution of the partial differential equations generates the Schwarzschild geometry. A displacement of the origin of the coordinate system by an imaginary amount ia in this solution leads to another obvious solution, which in turn generates the Kerr geometry. This result is similar to that of Janis and Newman,[3] who "complexify" a radial coordinate to turn a Schwarzschild geometry into a Kerr geometry; the difference is fundamental, however, since the "derivation" of Janis and Newman (their quotes) is an *ad hoc* procedure which may or may not yield a solution to the free space field equations. In our approach the complexification, which is performed on the generating function, is both well motivated and rigorous, in that it clearly must lead to a solution to the free space field equations.

Our purpose is twofold. In addition to demonstrating the mathematical relation of the Kerr and Schwarzschild geometries, we intend that our derivation should be pedagogically complete and as elementary as possible. As a bonus the structure of the equations reveals a very close similarity to classical Newtonian theory.

Our approach is algebraic and parallel to the geometric approach of Kerr and Schild[4] and of Kerr, Schild, and Debney.[5] These authors present a method of determining all solutions with the special metric form which we will consider. More recent work of Perjes,[6] and Parker, Ruffini, and Wilkins[7] on the problem of several charged Kerr–Newman[8] sources in equilibrium starts with a slightly different metric form; but it appears that a similar and interesting interpretation of their results should be possible.

2. EDDINGTON'S FORM OF THE SCHWARZSCHILD METRIC

In its original and most widely known form the Schwarzschild line element is[9]

$$ds^2 = [1 - (2m/r)]dt^2 - [1 - (2m/r)^{-1}dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2), \quad (2.1)$$

where m is the geometric mass of the source. In 1924, Eddington obtained an alternative form that is very useful for our purposes.[10] By introducing a new time coordinate

$$\bar{t} = t + 2m \log[(r/2m) - 1], \quad (2.2)$$

he obtained the line element

$$ds^2 = (d\bar{t})^2 - dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2) - (2m/r)(d\bar{t} + dr)^2, \quad (2.3)$$

or, in Cartesian coordinates

$$ds^2 = (d\bar{t})^2 - (dx^2 + dy^2 + dz^2) - \frac{2m}{r} \times \left(d\bar{t} + \frac{xdx + ydy + zdz}{r}\right)^2,$$

$$r^2 = x^2 + y^2 + z^2. \quad (2.4)$$

This is simply a flat space line element plus a term with interesting properties. We may write this line element as

$$ds^2 = g_{\alpha\beta}dx^\alpha dx^\beta = \eta_{\alpha\beta}dx^\alpha dx^\beta - 2m l_\alpha l_\beta dx^\alpha dx^\beta, \quad (2.5)$$

where $\eta_{\alpha\beta}$ is the Lorentz metric and l_α is a 4-vector given explicitly by

$$l_\alpha = (1/r)^{1/2}\ (1, x/r, y/r, z/r). \quad (2.6)$$

This 4-vector has the remarkable property that it is null with respect to both the Lorentz metric $\eta_{\alpha\beta}$ and the true metric $g_{\alpha\beta}$.

When the metric of an Einstein space can be cast in the form given in (2.5), it may be shown[4,11] that the space is not of Petrov type I, the most general form. This is equivalent to the statement that the separate Debever–Penrose principal null directions must number less than 4.[12] Such a space is termed algebraically

special or degenerate. Since the metric form (2.5) implies an algebraically special or degenerate space we will refer to it, for brevity, as degenerate; this does not imply that the metric of *any* algebraically special space can be cast in this form. In the remainder of this work we will assume a degenerate metric form. The purely algebraic problem of simplifying and solving the Einstein equations is thereby greatly facilitated.

3. DEGENERATE METRICS AND THE EINSTEIN EQUATIONS

Motivated by the preceding comments we now begin our formal development by assuming a degenerate metric, written as

$$g_{\alpha\beta} = \eta_{\alpha\beta} - 2ml_\alpha l_\beta, \tag{3.1}$$

where l_α is a null 4-vector with respect to $\eta_{\alpha\beta}$, the Lorentz metric,

$$l_\alpha l_\beta \eta^{\alpha\beta} = 0. \tag{3.2}$$

This is the form of metric originally assumed by Kerr,[2] and Kerr and Schild.[4] The constant m is an arbitrary constant, i. e., we demand that (3.1) be a solution to the Einstein free space field equations for any value of m. It is easily verified that the contravariant form of the metric tensor is

$$g^{\alpha\beta} = \eta^{\alpha\beta} + 2ml^\alpha l^\beta \tag{3.3}$$

and that indices on the 4-vector l_α may be raised and lowered with either the true metric tensor or the Lorentz metric; for example,

$$l^\alpha = g^{\alpha\tau} l_\tau = \eta^{\alpha\tau} l_\tau, \qquad l^\alpha l_\alpha = 0. \tag{3.4}$$

The 4-vector l_α has other interesting properties; since it is a null vector

$$l^\alpha l_{\alpha\,|\,\tau} = l_\alpha l^\alpha{}_{|\,\tau} = 0. \tag{3.5}$$

Also the covariant form of this relation holds. To see this, note that

$$\left\{{\alpha \atop \beta\mu}\right\} l^\mu = -m(l^\alpha l_\beta)_{|\,\mu} l^\mu, \tag{3.6}$$

from which it follows that

$$l^\alpha l_{\alpha\,\|\tau} = l_\alpha l^\alpha{}_{\|\tau} = 0. \tag{3.7}$$

The field equations themselves are much simplified by our choice of metric. As follows from definition (3.1) the metric determinant is $\|g\| = -1$, so that

$$\left\{{\alpha \atop \beta\alpha}\right\} = (\log\sqrt{-g})_{|\,\beta} = 0, \tag{3.8}$$

and the field equations contain only two terms

$$R_{\mu\nu} = -\left\{{\alpha \atop \mu\nu}\right\}_{|\,\alpha} + \left\{{\alpha \atop \beta\mu}\right\}\left\{{\beta \atop \alpha\nu}\right\} = 0. \tag{3.9}$$

Since the metric tensor is a first order polynomial in m, $R_{\mu\nu}$ is a fourth-order polynomial in m. Because m is arbitrary each order must vanish separately, which leads to the following four sets of field equations:

$$\eta^{\alpha\rho}[\mu\nu,\rho]_{|\,\alpha} = 0: \text{order } m, \tag{3.10a}$$

$$2m(l^\alpha l^\rho[\mu\nu,\rho])_{|\,\alpha} - \eta^{\alpha\sigma}\eta^{\beta\lambda}[\beta\mu,\sigma][\alpha\nu,\lambda] = 0: \text{order } m^2, \tag{3.10b}$$

$$l^\beta l^\lambda \eta^{\alpha\sigma}[\beta\mu,\sigma][\alpha\nu,\lambda] + l^\alpha l^\lambda \eta^{\beta\sigma}[\beta\mu,\lambda][\alpha\nu,\sigma] = 0: \text{order } m^3, \tag{3.10c}$$

$$l^\alpha l^\sigma l^\beta l^\lambda [\beta\mu,\sigma][\alpha\nu,\lambda] = 0: \text{order } m^4. \tag{3.10d}$$

The order m^4 equations form an identity by virtue of the properties of l_α already noted in (3.4) and (3.5), while the order m^3 equations lead us to an interesting new vector. Equation (3.10c) leads to

$$-ml_\mu l_\nu v^\alpha v_\alpha = 0, \tag{3.11}$$

where

$$v^\alpha = l^\beta l^\alpha{}_{\|\beta} = l^\beta l^\alpha{}_{|\,\beta}. \tag{3.12}$$

Thus v^α is a null vector; it is moreover easily seen to be orthogonal to the null vector l^α. From this we may infer that v^α and l^α are proportional at each point and, thus, are related by

$$v^\nu = l^\alpha l^\nu{}_{\|\alpha} = -Al^\nu, \tag{3.13}$$

where A is a scalar field. (This has the physical interpretation that the vector field l^μ is tangent to a family of geodesics.[13])

We will defer discussion of the order m^2 equations until later and proceed to simplify the order m equations. In terms of the D'Alembertian operator $\Box^2 = (\partial^2/\partial t^2) - \nabla^2$ and a scalar L defined by

$$L = -l^\alpha{}_{\|\alpha} = -l^\alpha{}_{|\,\alpha}, \tag{3.14}$$

the order m equations (3.10a) may be rewritten in convenient and concise form:

$$-\Box^2(l_\mu l_\nu) = [(L+A)l_\mu]_{|\,\nu} + [(L+A)l_\nu]_{|\,\mu}. \tag{3.15}$$

(The scalar L also occurs in the study of geometric optics in a Riemann space and is related to the expansion parameter θ in the notation of Pirani.[13])

Up to this point we have not assumed any symmetry of the metric. In our further development in the next section it will be convenient to consider only the case where the metric is stationary, or independent of t. In the stationary case any solution of the order m equations is automatically a solution of the order m^2 equations as we will show in Sec. 5.

4. SIMPLIFICATION OF THE FIELD EQUATIONS

In the stationary case it is possible to reduce the field equations to two simple partial differential equations for a single complex function. These differential equations are the Laplace and the Eikonal equations. The simplification is achieved by slightly lengthy, but elementary algebraic manipulation of the order m equations (3.15).

We begin by introducing a unit 3-vector λ_j to replace the space components of l_α.[9]

$$l_j = l_0\lambda_j, \qquad \lambda_j\lambda_j = 1. \tag{4.1}$$

The unit length of $\boldsymbol{\lambda}$ is a consequence of l^α being a null vector. The order m equations (3.15) then read

$$\nabla^2(l_0{}^2) = 0, \tag{4.2a}$$

$$\nabla^2(l_0{}^2\lambda_j) = [(L + A)l_0]_{|j}, \tag{4.2b}$$

$$\nabla^2(l_0{}^2\lambda_i\lambda_j) = [(L + A)l_0\lambda_i]_{|j} + [(L + A)l_0\lambda_j]_{|i}. \tag{4.2c}$$

These second-order differential equations can be manipulated to give a useful set of first-order differential equations to replace (4.2c). We expand (4.2c) and use (4.2a) and (4.2b) to cancel terms, and obtain

$$\lambda_{i|j} + \lambda_{j|i} = [2l_0/(L + A)]\lambda_{i|k}\lambda_{j|k} \equiv (1/p)\lambda_{i|k}\lambda_{j|k}. \tag{4.3}$$

It is possible to solve (4.3) for $\lambda_{i|j}$ in terms of λ_k by the use of matrix algebra. We denote $\lambda_{i|k}$ by the matrix M so (4.3) reads

$$M + M^T = (1/p)MM^T. \tag{4.4}$$

Moreover, from Eq. (3.13) and the unit length of $\boldsymbol{\lambda}$ we see that λ is in the null space of both M and M^T; that is

$$\begin{aligned} &\lambda_{k|j}\lambda_j = 0, \quad M\lambda = 0, \\ &\lambda_j\lambda_{j|k} = 0, \quad M^T\lambda = 0. \end{aligned} \tag{4.5}$$

It is remarkable that Eqs. (4.4) and (4.5) determine the matrix M, up to one real parameter, in terms of λ. By a rotation R of coordinates the vector λ may be placed along the x axis

$$\lambda' = R\lambda = \begin{pmatrix}1\\0\\0\end{pmatrix}. \tag{4.6}$$

In these rotated coordinates, λ' is in the null space of $M' = RMR^T$ and M'^T. It is, therefore, obvious that M' must have the form

$$M' = RMR^T = \left(\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & & \\ 0 & & N' \end{array}\right). \tag{4.7}$$

Due to the invariance of matrix algebra under rotations, Eq. (4.4) holds for the matrix M' and also for the 2×2 submatrix N':

$$N' + N'^T = (1/p)N'N'^T. \tag{4.8}$$

But this implies that N' is related to a 2×2 real unitary matrix U by

$$U = I - (1/p)N', \quad UU^+ = I. \tag{4.9}$$

If U is proper and has a positive determinant we can write it in terms of only one real parameter θ as

$$U = \begin{pmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{pmatrix}, \tag{4.10}$$

so M' may be written as

$$M' = p\begin{pmatrix}0 & 0 & 0 \\ 0 & 1-\cos\theta & \sin\theta \\ 0 & -\sin\theta & 1-\cos\theta\end{pmatrix}. \tag{4.11}$$

We have here assumed that U has a positive determinant, since if we allow the determinant to be negative, the resulting solutions to the field equations cannot represent an isolated point mass. We must now rotate back to the original coordinates to get $M = R^TM'R$. Using the orthonormality of the rows and columns of R we obtain

$$M_{ik} = p(1 - \cos\theta)(\delta_{ik} - R_{1i}R_{1k}) + p\,\sin\theta\epsilon_{ikl}R_{1l}. \tag{4.12}$$

It is fortunate that only the first row of the matrix R appears; we denote this by the vector $\mathbf{R}$. To relate $\mathbf{R}$ to $\boldsymbol{\lambda}$ we now impose the conditions in Eq. (4.5), and find that

$$\boldsymbol{\lambda} = \mathbf{R}(\mathbf{R}\cdot\boldsymbol{\lambda}), \quad \boldsymbol{\lambda}\times\mathbf{R} = 0. \tag{4.13}$$

Thus $\mathbf{R} = \pm\boldsymbol{\lambda}$; since the overall sign of the matrix R is arbitrary we may choose the plus sign. We may now write M, or $\lambda_{i|k}$, in terms of only λ_i and the arbitrary parameter θ as

$$\lambda_{i|k} = p(1 - \cos\theta)(\delta_{ik} - \lambda_i\lambda_k) + p\,\sin\theta\epsilon_{ikl}\lambda_l. \tag{4.14}$$

This represents a most useful simplification of Eq. (4.3). It replaces a nonlinear relation between the first derivatives of λ_i by an explicit expression for $\lambda_{i|k}$ in terms of λ_i. The price we pay is the introduction of the arbitrary parameter θ.

A consideration of the algebraic content of (4.14) will now lead directly to the simple differential equations discussed in the introduction. We first rewrite (4.14) with new parameters α and β replacing p and θ,

$$\lambda_{i|k} = \alpha(\delta_{ik} - \lambda_i\lambda_k) + \beta\epsilon_{ikl}\lambda_l. \tag{4.15}$$

This is more than a minor algebraic simplification since we will show that α and β conveniently determine the metric. From this expression we immediately obtain

$$\nabla\cdot\boldsymbol{\lambda} = 2\alpha \quad \nabla\times\boldsymbol{\lambda} = -2\beta\boldsymbol{\lambda} \tag{4.16}$$

The Laplacian of λ may be obtained in two ways; directly from (4.15) we get

$$\nabla^2\boldsymbol{\lambda} = \nabla\alpha - \boldsymbol{\lambda}(\nabla\alpha\cdot\boldsymbol{\lambda}) - 2(\alpha^2 + \beta^2)\boldsymbol{\lambda} + \nabla\beta\times\boldsymbol{\lambda}. \tag{4.17}$$

Alternatively we can take the curl of $\nabla\times\boldsymbol{\lambda}$ from (4.16), expand with a vector identity, and use (4.16) to simplify the result to

$$\nabla^2\boldsymbol{\lambda} = 2\nabla\alpha + 2(\nabla\beta\times\boldsymbol{\lambda}) - 4\beta^2\boldsymbol{\lambda}. \tag{4.18}$$

By equating these two expressions for $\nabla^2\boldsymbol{\lambda}$ we get an expression for $\nabla\alpha$

$$\nabla\alpha = -\nabla\beta\times\boldsymbol{\lambda} - \boldsymbol{\lambda}(\nabla\alpha\cdot\boldsymbol{\lambda}) - 2(\alpha^2 - \beta^2)\boldsymbol{\lambda}. \tag{4.19}$$

From this we obtain the important results

$$\nabla\alpha\cdot\boldsymbol{\lambda} = \beta^2 - \alpha^2, \quad \nabla\alpha = (\beta^2 - \alpha^2)\boldsymbol{\lambda} - (\nabla\beta\times\boldsymbol{\lambda}). \tag{4.20}$$

Analogous relations for β can also be obtained. From (4.16) we can calculate, similar to (4.20),

$$\nabla\beta\cdot\boldsymbol{\lambda} = -2\alpha\beta. \tag{4.21}$$

By forming $\nabla\alpha\times\boldsymbol{\lambda}$ from (4.19) and using (4.21) to simplify, we obtain the gradient of β:

$$\nabla\beta = -2\alpha\beta\boldsymbol{\lambda} + \nabla\alpha\times\boldsymbol{\lambda}. \tag{4.22}$$

The vector relations (4.19)–(4.22) can be reexpressed in concise fashion by introducing a complex function $\gamma = \alpha + i\beta$. In terms of this γ Eqs. (4.19)–(4.22) may be written

$$\nabla\gamma \cdot \boldsymbol{\lambda} = -\gamma^2, \nabla\gamma = -\gamma^2\boldsymbol{\lambda} + i(\nabla\gamma \times \boldsymbol{\lambda}). \tag{4.23}$$

The importance of viewing $\alpha + i\beta$ as a single complex function should be stressed. In terms of γ the Kerr and Schwarzschild solutions have a transparent relation, as discussed in the introduction. Moreover our further algebraic development is very easy in terms of γ.

Our remaining tasks are to obtain differential equations to determine γ, and to show that γ then uniquely determines λ_i and l_0, i.e., the complete metric. Thus the complex function γ is a generating function and contains full information on the geometry. To obtain the first differential equation we form the Laplacian of γ and use (4.16) and (4.23) to simplify. Then we obtain Laplace's equation for γ

$$\nabla^2\gamma = 0, \tag{4.24}$$

so that γ is a complex harmonic function. A second equation is nonlinear and follows from squaring $\nabla\gamma$ in (4.23)

$$(\nabla\gamma)^2 = \gamma^4. \tag{4.25}$$

An alternative and useful form of this is

$$(\nabla\omega)^2 = 1, \quad \omega = \gamma^{-1}, \tag{4.26}$$

which is the Eikonal equation, familiar in classical optics. The function γ is thus seen to be a complex harmonic function whose inverse satisfies the Eikonal equation. (Our γ is related to F_y of Ref. 4 and 5 by $\gamma = \sqrt{2}/F_y$.)

Now that we have a system of equations for γ it remains only to show that λ_i and l_0 are determined uniquely by γ. It is possible to solve (4.23) directly for $\boldsymbol{\lambda}$; we first rewrite it in terms of $\gamma^{-1} = \omega$

$$\boldsymbol{\lambda}\cdot\nabla\omega = \boldsymbol{\lambda}\cdot\nabla\omega^* = 1, \quad \nabla\omega = \boldsymbol{\lambda} + i(\nabla\omega \times \boldsymbol{\lambda}). \tag{4.27}$$

From this

$$\nabla\omega \times \nabla\omega^* = -i[\nabla\omega^* + \nabla\omega] + B\boldsymbol{\lambda}, \tag{4.28}$$

where B represents the coefficient of λ, which depends on λ and ω. However, we can solve (4.28) itself for B and thereby obtain $\boldsymbol{\lambda}$ very simply; we dot $\nabla\omega$ into (4.28) and use (4.27) to obtain

$$B = i[1 + \nabla\omega\cdot\nabla\omega^*] \tag{4.29}$$

and thus

$$\boldsymbol{\lambda} = [\nabla\omega + \nabla\omega^* - i(\nabla\omega \times \nabla\omega^*)]/(1 + \nabla\omega\cdot\nabla\omega^*), \tag{4.30}$$

which is manifestly real. Thus we see that a knowledge of γ specifies λ uniquely via an explicit equation; it is easy to verify this solution by substituting it into (4.15).

Our final task is to show that l_0 can also be obtained from γ. We will in fact invoke Eqs. (4.2a) and (4.2b) to show that a solution is $l_0{}^2 = \alpha = \mathrm{Re}\gamma$. Using this as a trial solution we first calculate the left side of (4.2b), $\nabla^2(\alpha\boldsymbol{\lambda})$, using the fact that α is harmonic. This Laplacian simplifies with the use of (4.15), (4,18), (4.20), and (4.22) to

$$\nabla^2(\alpha\boldsymbol{\lambda}) = \boldsymbol{\nabla}(\alpha^2 + \beta^2). \tag{4.31}$$

To get the right side of (4.2b) we return to (4.3). Setting $i = j$ in (4.3) and summing we see that the left side is twice the divergence of $\boldsymbol{\lambda}$, or 4α, while $\lambda_{i\,|k}\lambda_{i\,|k} = 2(\alpha^2 + \beta^2)$ from (4.15). Thus (4.3) leads to

$$(A + L) = (l_0/\alpha)(\alpha^2 + \beta^2), \quad (A + L)l_0 = \alpha^2 + \beta^2, \tag{4.32}$$

where $l_0{}^2 = \alpha$ is the trial solution. From the above two equations it is evident that Eq. (4.2b) is indeed satisfied by $l_0{}^2 = \alpha$. Since α is Reγ and harmonic, (4.2a) is also satisfied. It may be readily shown that this solution is unique up to a multiplicative constant, so that we have finally

$$l_0{}^2 = \kappa\alpha = \kappa\,\mathrm{Re}\gamma. \tag{4.33}$$

Our results can now be summarized. We have reduced the order m free space field equations to the pair of equations (4.24) and (4.26) and the recipes in (4.30) and (4.33). Thus to solve for the metric we find a harmonic function γ whose inverse satisfies the Eikonal equation. The metric components l_0 and λ_i, and thus l_α are generated by the explicit expressions (4.30) and (4.33). The complex function γ plays the role of a generalized Newtonian potential in that it obeys Laplace's equations and Re$\gamma = l_0{}^2$. (The function $l_0{}^2$ must be proportional to the Newtonian potential in the limit of small m, as is well known in the linearized theory.)

5. ORDER m^2 EQUATIONS

We have indicated that any stationary solution to the order m^3 and order m equations automatically satisfies the order m^2 equations. Now we will show this. The order m^2 equations (3.10b) reduce immediately to

$$[2(l^\alpha A)_{|\alpha} + l^\alpha{}_{|\beta}l^\beta{}_{|\alpha} - l^\alpha{}_{|\beta}l_\alpha{}^{|\beta} - A^2]l_\mu l_\nu = 0. \tag{5.1}$$

In the stationary case this implies the scalar equation

$$-2(l_iA)_{|i} + l_{i\,|j}l_{j\,|i} + l^\alpha{}_{|k}l_{\alpha\,|k} - A^2 = 0. \tag{5.2}$$

From the work of the preceding section it is easy to show that

$$\begin{aligned} l_{i\,|j}l_{j\,|i} &= [(A - L)l_i]_{|i} + L^2, \\ l^\alpha{}_{|k}l_{\alpha\,|k} &= A^2 - L^2. \end{aligned} \tag{5.3}$$

Thus (5.2) becomes

$$[(A + L)l_j]_{|j} = 0. \tag{5.4}$$

However if we set $i = j$ in the order m equation (4.2c) we find that

$$\nabla^2(l_0{}^2\boldsymbol{\lambda}^2) = \nabla^2(l_0{}^2) = 0 = 2[(A + L)l_j]_{|j} \tag{5.5}$$

so that (5.4) becomes an identity.

6. THE SCHWARZSCHILD AND KERR GEOMETRIES

We now obtain the simplest solution to Eqs. (4.24) and (4.26). In analogy with Newtonian theory we consider

$$\gamma = r^{-1} = [x^2 + y^2 + z^2]^{-1/2} \tag{6.1}$$

as a solution to the Laplace equation. It is easily checked that $\gamma^{-1} = r$ also obeys the Eikonal equation, so $\gamma = r^{-1}$ is a generating function. The metric functions generated by Eqs. (4.30) and (4.33) from $\gamma = r^{-1}$ are

$$l_0{}^2 = r^{-1}, \quad \lambda_1 = x/r, \quad \lambda_2 = y/r, \quad \lambda_3 = z/r, \tag{6.2}$$

which leads to the Eddington form of the Schwarzschild line element (2.4).

If we now ask for the simplest generalization of the above solution we are led to consider general displacement of the origin, i.e.,

$$\gamma = \tilde{r}^{-1} = [(x - a_1)^2 + (y - a_2)^2 + (z - a_3)^2]^{-1/2}. \tag{6.3}$$

This satisfies (4.24) and (4.26) for any choice of a_i. However, this solution merely represents a physical displacement of the coordinate origin if the a_i are real, which is of no physical interest. If the a_i are imaginary, however, we have a new result. By a suitable orientation of the three-dimensional coordinates we may, with no loss of generality, write an imaginary displaced γ function as

$$\gamma = [x^2 + y^2 + (z - ia)]^{-1/2}. \tag{6.4}$$

The metric functions generated from this by (4.30) and (4.33) are

$$l_0{}^2 = \frac{\rho^3}{\rho^4 + a^2z^2}, \quad \lambda_1 = \frac{\rho x + ay}{a^2 + \rho^2}, \quad \lambda_2 = \frac{\rho y - ax}{a^2 + \rho^2}, \quad \lambda_3 = \frac{z}{\rho}, \tag{6.5}$$

where ρ is the real part of $\omega = \gamma^{-1}$ and is a solution of

$$\rho^4 - \rho^2(r^2 - a^2) - a^2z^2 = 0. \tag{6.6}$$

Corresponding to the metric functions in (6.5) is the line element

$$ds^2 = dt^2 - (dx^2 + dy^2 + dz^2) - \frac{2m\rho^3}{\rho^4 + a^2z^2}\left(dt + \frac{\rho}{a^2 + \rho^2}(x\,dx + y\,dy) + \frac{a}{a^2 + \rho^2}(y\,dx - x\,dy) + \frac{z}{\rho}\,dz\right)^2. \tag{6.7}$$

This is the Kerr line element in a form given in Ref. (1).

7. SUMMARY AND FURTHER COMMENTS

The Einstein free space field equations simplify for the case of the degenerate metric to the Laplace and eikonal equations (4.24) and (4.26). Solutions to these equations generate the metric tensor via Eqs. (4.30) and (4.33). We have used these results to show that the generating function for the Schwarzschild geometry becomes a generating function for the Kerr geometry if the origin is displaced by an imaginary amount.

Further work is under way on the application of similar algebraic techniques to the spinning charged source geometry of Newman et al.,[8] the many source geometry of Perjes,[6] and the problem of the interior Kerr solution. The uniqueness conjectures of Carter and Israel [14,15] are also under consideration in the context of the degenerate metric form.

ACKNOWLEDGMENT

The authors wish to thank Professor J. A. Wheeler for his perceptive comments on this work.

---

*Supported in part by the National Science Foundation, GP-16565.
†Supported in part by the U.S Atomic Energy Commission.

[1] R. P. Kerr, Phys. Rev. Lett. **11**, 237 (1963).
[2] A useful summary of many facets of these fields may be found in "Relativistic Cosmology and Space Platforms," R. Ruffini and J. A. Wheeler's preprint (to be published).
[3] E. T. Newman and A. I. Janis, J. Math. Phys. **6**, 915 (1965). These authors indicate that R. P. Kerr has in a private communication, further justified their complexification process.
[4] R. P. Kerr and A. Schild, *Atti del convegno sulla relatività generale: Problemi dell' energia e onde gravitazionali (Anniversary Volume, Fourth Centenary of Galileo's Birth)* (G. Barbéra, Firenze, 1964), p. 173; *Applications of nonlinear partial differential equations in mathematical physics, Proceedings of symposia in applied mathematics* (Amer. Math. Soc., Providence, R.I., 1965), Vol. XVII, p. 199.
[5] G. C. Debney, R. P. Kerr, and A. Schild, J. Math. Phys. 10, 1842 (1969).
[6] Z. Perjes, Phys. Rev. Lett. **27**, 1668 (1971).
[7] L. Parker, R. Ruffini, and D. Wilkins, Bull. Am. Phys. Soc. **17**, 449 (1972).
[8] E. T. Newman, E. Couch, K. Chinnapared, A. Exton, A. Prakash, and R. Torrence, J. Math. Phys. **6**, 918 (1964).
[9] We use geometric units in which $c = G = 1$. Our metric signature is $(1, -1, -1, -1)$, Greek indices run from 0 to 3, and Latin indices run from 1 to 3. All repeated indices are to be summed over regardless of position.
[10] A. S. Eddington, Nature (Lond.) **113**, 192 (1924).
[11] A. Schild, "Lectures in General Relativity Theory" in *Lectures in applied mathematics*, edited by J. Ehlers (Amer. Math. Soc., Providence, R.I., 1967), Vol. 8.
[12] R. Penrose, Ann. Phys. (N.Y.) **10**, 171 (1960).
[13] F. A. E. Pirani, "Introduction to Gravitational Radiation Theory" in *Lectures on general relativity* (Prentice-Hall, Englewood Cliffs, N. J., 1965).
[14] W. Israel, Phys. Rev. **164**, 1776 (1967); Commun. Math. Phys. **8**, 245 (1968).
[15] B. Carter, Phys. Rev. **174**, 1559 (1968); Phys. Rev. Lett. **26**, 331 (1971).

Conformal group in a Poincaré basis. II. Principal discrete series

N. W. Macfadyen
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England
(Received 3 May 1971)

In the second of this series of papers we study the six principal discrete series of unitary irreducible representations of $SU(2,2)$. The same techniques are used as before, except that to examine the reducibility we are forced to use complexifications of all of our spaces. It is found that when restricted to the Poincaré group, two of the series contain only timelike momenta—with positive and negative masses, respectively — and a finite number of spins; the remaining four series contain spacelike momenta and spins which allow a certain helicity. A point of interest is that these two classes require entirely different scalar products.

I. INTRODUCTION

In a previous paper (Ref. 1, to which we shall refer as I), we constructed representations of the spin-covering group $SU(2,2)$ of the conformal group $SO(4,2)$ in the two principal continuous series, and showed how they reduced when restricted to the Poincaré group $\mathbb{P}$. We now consider the principal discrete series of representations. We shall assume acquaintance with the notation and conventions of I, and refer to the equations of that paper where necessary.

Recall then that we established our representations by operators on a space of functions defined over the six-dimensional manifold $Y = \{a, \bar{a}, b, c, \omega, \bar{\omega}\}$. In the series d_2 we found immediate irreducibility; but d_1 was reducible and we needed to introduce the two subspaces of functions analytic in $\mathrm{Re}\, b \gtrless 0$ for vanishing ω. This was a consequence of our still using Y, whereas a more natural treatment (as outlined for example by Graev[2]) would introduce a different manifold. The problem is greatly aggravated for the principal discrete series; for the natural treatment in this case requires a manifold that is different indeed from that which we have used. Let us remind ourselves briefly of this treatment.

The discrete series of representations then are induced by those of a compact Cartan subgroup, for instance that generated by $\{J_3, 2L_0 - P_0, 2L_3 + P_3\}$. Consider the space H of functions $f(\mathrm{Z}, \xi, \omega)$, where $\mathrm{Z}, \xi, \omega \in GL(2,c)$, which satisfy the following conditions.

1. For fixed ξ and ω, f is a polynomial in $\mathrm{Z}_{11}, \mathrm{Z}_{12}$, and $|\mathrm{Z}|$, homogeneous in the first two variables together and the third by itself; similarly for fixed Z and ω, and variable ξ.

2. For fixed Z and ξ, f is analytic in the elements of ω in the domain $\Omega: \omega\omega^+ < 1$ (the generalized unit disc).[3]

3. $$\int |f[u(1-\omega^+\omega)^{1/2}, v(1-\bar{\omega}\omega^T)^{1/2}, \omega]|^2\, |1-\omega\omega^+|^{-4} \times d\mu(u)\, d\mu(v)\, d\mu(\omega) \equiv \|f\|^2 < \infty,$$

where $u, v \in U(2)$, $d\mu(u)$ is the invariant measure, and $d\mu(\omega) = \prod d(\mathrm{Re}\,\omega_{ij})\, d(\mathrm{Im}\,\omega_{ij})$. The integration is over all unitary matrices u and v and over the manifold Ω.

Then a scalar product is defined in H to agree with this norm, and parametrising any $g \in U(2,2)$ by

$$g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix},$$

we set

$$T_g: f(\mathrm{Z}, \xi, \omega) = f[\mathrm{Z}(\alpha^+ + \gamma^+\omega)^{-1}, \xi(\rho^T\omega^T + \delta^T)^{-1}, (\omega\beta + \delta)^{-1}(\omega\alpha + \gamma)].$$

This defines a unitary representation of $U(2,2)$ which is specified by four integers—the degrees of homogeneity referred to above. Graev has shown[2] that there are five other representations with the same integer parameters (he calls the above "type I") and that these six are all irreducible and mutually inequivalent. There are no other unitary discrete nondegenerate series.

This then may be termed the canonical representation of the discrete series. It is clearly quite unsuitable for displaying the behavior under the Poincaré group [although extremely convenient for the reduction $U(2,2) \supset U(2) \times U(2)$] because of the manifold of matrices $\{\mathrm{Z}, \xi, \omega\}$ which appears as the basic parameter space; and indeed the remaining five irreducible components of the series d_0 are still more complicated. Obviously it is desirable to construct the representation instead over our manifold Y; this can be done by allowing the three parameters $(m, i\rho_1, i\rho_2)$ of (I. 21) to take integer values, and we are then left with the twin problems of reducibility and construction of a scalar product.

To examine the reducibility we are forced to consider not a space $\mathfrak{D}_0$ of functions over Y, but rather a space $\mathfrak{D}$ over the complexification Y_c of Y. It transpires that we can find six submanifolds of Y_c which are invariant under all transformations of the group, and so $\mathfrak{D}$ breaks down into a sum of six invariant subspaces $\mathfrak{D}^i$ of functions with certain analyticity properties. On these spaces we can then somewhat formally write a scalar product and thereby constitute our representations as unitary and irreducible.

Closer examination of the product, however, shows that the representations fall into two distinct classes, which we shall call $d_0^{\pm}$(timelike) and d_0^{0i}(spacelike). (We show in Sec. V that these names are justified.) The inner product actually takes entirely different forms on the two classes (the distinction was hidden earlier by a formal regularization), and we examine them separately in Sec. IV, where we find that for the timelike series the scalar product makes essential use of the complex nature of Y_c, whereas for the four spacelike series we can introduce a Hermitian form into the boundary-value space $\mathfrak{D}_0$ and ignore the "off-shell" continuation $\mathfrak{D}$. This situation can be understood better by examining the relationship between Graev's representation and our own. It can be shown quite straightforwardly—we do not do so in this paper—that Graev's space H for type I representations can be mapped onto our space $\mathfrak{D}^+$, which transforms under d_0^+, and that his transformation law becomes ours. The boundary $\omega\omega^+ = 1$ of his domain Ω maps onto the boundary Y of the manifold Y_c. Now his functions are analytic in certain domains; we can therefore specify them uniquely by their values on the boundaries of these regions (by the generalized Cauchy theorem), and so we

can construct the representation on the boundary-value space. But to define a scalar product, Graev needs the entire domains and not merely the boundaries: This is the situation we find in our treatment of $d_0^{\pm}$.

For the four spacelike series d_0^{0i}, which Graev does not discuss, the situation is different because we no longer have polynomial dependence on the elements of the matrices Z and ξ, but rather an analyticity requirement. We can therefore construct the scalar product with differential operators (which was impossible for $d_0^{\pm}$ since the requisite operator annihilated the entire space of polynomials) and genuinely restrict ourselves to the boundary manifold Y and functions thereon. The situation is very similar to the usual treatment[4] of the discrete series of representations of $SL(2,R)$, even up to the redefinition of our carrier spaces modulo the kernel of the differential operator, which is also the union of the pairwise intersections of the six irreducible subspaces of $\mathfrak{D}$.

The plan of the paper then is as follows. Section II introduces various complex manifolds and their transformation properties, and Sec. III constructs the unitary irreducible representations by means of operators on function spaces over these manifolds. It turns out that the three parameters (M, L, K) specifying the representations must not only be integers but also satisfy a parity relationship: $(L + K + M)$ must be even. Section IV discusses the scalar product in more detail, especially for the spacelike series, and shows that for them the representations can be extended to further values of the parameters.

Section V then shows how the representations reduce under the Poincaré group $\mathbb{P}$, and justifies our calling them timelike or spacelike. The series d_0^{+} is shown to contain only a finite number of spins, and all positive-mass momenta; d_0^{-} has only negative masses. The four spacelike series d_0^{0i} contain all "spins" that allow a certain helicity, exactly as in I. Finally, Sec. VI comments on our results. Two appendices contain material whose inclusion in the body of the text would be undesirable.

II. GROUP ELEMENTS AND COMPLEX MANIFOLDS

We recall[1] that our group $\mathcal{G}$, which is isomorphic to $SU(2,2)$, is defined by the antidiagonal metric tensor which corresponds to the quadratic form $\bar{Z}_1 Z_4 + \bar{Z}_2 Z_3 + Z_2 \bar{Z}_3 + Z_1 \bar{Z}_4$. We now define a standard set of elements g_i such that any $g \in \mathcal{G}$ can be written as a finite product of these elements; instead of examining an arbitrary element of the group it will then be sufficient to consider each g_i separately. We shall take as our set the translations Z and the dilation d:

$$z \in Z = \begin{pmatrix} 1 & \\ z & 1 \end{pmatrix}, \qquad d = \begin{pmatrix} d1 & \\ & d^{-1}1 \end{pmatrix} \tag{1}$$

together with the three block-diagonal matrices sufficient to cover the $SL(2,C)$ subgroup. Their (2, 2) blocks are

$$\sigma = \begin{pmatrix} \sigma & \\ & \sigma^{-1} \end{pmatrix}, \qquad \xi = \begin{pmatrix} 1 & \\ \xi & 1 \end{pmatrix}, \qquad I = \begin{pmatrix} & 1 \\ -1 & \end{pmatrix} \tag{2}$$

and the (1, 1) components are given by (I. 7); the remaining components vanish. Finally, we introduce two further elements:

$$J = \left(\begin{array}{cc:c} & & -1 \\ & 1 & \\ \hdashline & & \\ -1 & & \end{array}\right), \qquad B = \left(\begin{array}{cc:c} 1 & & \\ & i & \\ \hdashline & & \\ & & 1 \end{array}\right). \tag{3}$$

These are discrete operators of $\mathcal{G}$, either of which is sufficient for our purposes although sometimes one is more convenient than the other. That J with the others covers $\mathcal{G}$ follows since $JZJ = Z^T$; this is the operator considered by Castell.[5] The sufficiency of B now follows too because of the relation

$$B = z_1 J z_1 J z_1,$$

where z_1 is the element of Z specified by $a = 0 = c$, $b = i$. An analogous formula expresses J in terms of B. Notice that all these elements except d, J and B belong to the Poincaré group.

We now define the manifold Y_c which is the complexification of Y of (I. 10). This is the set of all complex matrices of the form

$$Y_c = \left(\begin{array}{cc:cc} 1 & & & \\ u & 1 & & \\ \hdashline a & b & 1 & \\ \omega a + c & \omega b + d & \omega & 1 \end{array}\right). \tag{4}$$

The relation $Y_c \in \mathcal{G}$ restricts Y_c to Y: We shall refer to this as the boundary. The parameters then satisfy

$$u = -\bar{\omega}, \qquad d = -\bar{a}, \qquad \mathrm{Re}b = 0 = \mathrm{Re}c. \tag{5}$$

Associated with Y_c is the manifold Λ_c which is the complexification of Λ:

$$\Lambda_c = \left(\begin{array}{cc:cc} \lambda_1^{-1} & -\mu_1 \Delta_1^{-1} & & \\ & \lambda_1 \Delta_1^{-1} & & L \\ \hdashline & & & \\ & & \lambda^{-1}\Delta & \mu \\ & & & \lambda \end{array}\right) \tag{6}$$

on its boundary $\Lambda \in \mathcal{G}$ we have (among others) the relations

$$\lambda_1 = \bar{\lambda}, \qquad \mu_1 = \bar{\mu}, \qquad \Delta_1 = \bar{\Delta}. \tag{7}$$

Since both Y_c and Λ_c define subgroups of $GL(4,C) \equiv \mathcal{G}_c$, which on their boundaries belong also to $\mathcal{G}$, we can define a transformation of Y_c under $\mathcal{G}_c$ which on the boundary reduces to the familiar form of (I. 13):

$$y_c g_c = \Lambda_c y_c'. \tag{8}$$

The reduction of the discrete series representations will then involve determining transitive domains of Y_c which are invariant under all $g \in \mathcal{G}$. The transformations under Z, d, and σ are almost trivial, but the remaining ones are less so; we give them here for convenience. Under ξ of (2) we obtain

$$\begin{aligned} a' &= a - b\bar{\xi}, & \omega' &= \omega + \xi, \\ b' &= b, & u' &= u - \bar{\xi}, \\ c' &= c - a\xi - d\bar{\xi} + b\xi\bar{\xi}, & \lambda &= 1 = \lambda_1, \\ d' &= d - b\xi, & \Delta &= 1 = \Delta_1, \end{aligned} \tag{9}$$

while under I,

$$\begin{aligned} a' &= -d, & \omega' &= -\omega^{-1}, \\ b' &= c, & u' &= -u^{-1}, \\ c' &= b, & \lambda &= \omega, \\ d' &= -a, & \lambda_1 &= -u, \\ & & \Delta &= 1 = \Delta_1, \end{aligned} \tag{10}$$

and under B,

$$\begin{aligned} a' &= -iab^{-1}, & \omega' &= i(\omega b + d), \\ b' &= b^{-1}, & u' &= i(ub - a), \\ c' &= (bc - ad)b^{-1}, & \lambda &= 1 = \lambda_1, \\ d' &= -idb^{-1}, & \Delta &= ib = \Delta_1. \end{aligned} \tag{11}$$

Notice that the new parameters are in each instance analytic functions of the old.

III. DISCRETE SERIES OF REPRESENTATIONS d_0

Consider the representation (I. 21) of the principal continuous series d_2 of $\mathcal{G}$; it is clear that upon setting $i\rho_1 = K$, $i\rho_2 = L$, where these are positive integers, we shall obtain another representation:

$$\begin{aligned} &T_g^0 : f(y) = \Delta^{K-1} |\lambda|^{L-K+m-2} \lambda^{-m} f(y') \\ &yg = \Lambda y'. \end{aligned} \tag{12}$$

(Notice that we have set $i\rho_2 = K$ and not $2k - 1$ as in I. The change is trivial.) The functions f belong to a certain topological vector space $\mathfrak{D}_0(K, L, m)$ of functions such that both they and their "inversions" $T_I f, T_B f$ are indefinitely differentiable in all variables, and this condition specifies the asymptotics of the functions for large values of their arguments; further details will be found in Appendix A.

Now let us extend (12) to a representation of $\mathcal{G}_c = GL(4, c)$ by writing it as

$$\begin{aligned} &T_g^0 : f(y_c) = \Delta^{K-1} \lambda^p \lambda_1^q f(y_c'), \\ &y_c g = \Lambda_c y_c', \end{aligned} \tag{13}$$

$$\begin{aligned} p &= \tfrac{1}{2}(L - K - m) - 1, \\ q &= \tfrac{1}{2}(L - K + m) - 1, \end{aligned} \tag{14}$$

where f now belongs to a space $\mathfrak{D}$ of functions over Y_c which satisfies similar requirements to $\mathfrak{D}_0$: It will be specified more closely when this is relevant. Suppose that both p and q are nonnegative integers; then since we know from the last section that not only the mapping $g: Y_c \to Y_c$ but also the parameters Δ, λ, and λ_1 are analytic in the six variables $\{a, b, c, d; u, \omega\}$, the transformations (13) preserve analyticity. Therefore the problem of subspace reducibility is just one of finding all submanifolds of Y_c which are $\mathcal{G}$ invariant. It is convenient to work with the functions of $\mathfrak{D}$ themselves rather than those of $\mathfrak{D}_0$ which are their boundary values, and so we shall henceforth remain in the complexified domains unless otherwise stated.

A. Reducibility

We start by examining the quantity T given by

$$T = (b + \bar{b})(c + \bar{c}) - (a + \bar{d})(\bar{a} + d). \tag{15}$$

On the boundary, $T = 0$; and off, it transforms under $\mathcal{G}$ by

$$g: T \to T' = |\Delta|^{-2} T \tag{16}$$

so that the two regions $T \gtrless 0$ are invariant under all $g \in \mathcal{G}$. We can reduce this still further, by noticing that $T > 0$ is divided by the surface $(b + \bar{b}) = 0$ into two parts which are themselves invariant; we shall call these two domains $T^{\pm}$:

$$T^{\pm} = \{Z_c \mid T > 0, -(b + \bar{b}) \gtrless 0\}. \tag{17}$$

The space $\mathfrak{D}$ can then be decomposed into three subspaces: of functions which are analytic in T^+ or T^-, and functions which are not. (Notice that we cannot have analyticity in the entire domain $T < 0$.) It is clear from the remarks above and the transformation law under $B \in \mathcal{G}$ given in (11), that the analyticity in $Z = \{a, b, c, d\}$ is preserved only if we also have analyticity in u and w; and hence, because there are no invariant regions in the space of these variables if $T > 0$, any dependence upon them must be polynomial if the analyticity is to be preserved. This therefore defines two subspaces $\mathfrak{D}^{\pm} \subset \mathfrak{D}$, of functions which are analytic in Z in $T^{\pm}$, and polynomial in u and ω.

Now let us return to the domain $T < 0$. We consider the two expressions

$$\begin{aligned} \Omega &\equiv (c + \bar{c}) + \omega(a + \bar{d}) + \bar{\omega}(\bar{a} + d) + \omega\bar{\omega}(b + \bar{b}), \\ U &\equiv (c + c) - u(a + d) - \bar{u}(a + \bar{d}) + u\bar{u}(b + \bar{b}), \end{aligned} \tag{18}$$

and find that

$$\begin{aligned} &g: \Omega \to \Omega' = |\lambda|^{-2} \Omega, \\ &g: U \to U' = |\lambda_1|^{-2} U, \end{aligned} \tag{19}$$

so that the regions $\Omega, U \gtrless 0$ are preserved too. It is trivial to verify that the conditions for the equations $\Omega, U = 0$ to have a solution for ω, u is indeed $T < 0$. We can therefore define four further invariant subspaces of $\mathfrak{D}$, of functions which are not analytic in z in either of the regions $T > 0$ (nor, of course, in $T < 0$) but which *are* analytic in u and ω in one of the four domains T^{0i}, where the notation is obvious; and we shall denote these by $\mathfrak{D}^{0i}$.

Then the union of the six subspaces $\mathfrak{D}^i = \{\mathfrak{D}^{\pm}, \mathfrak{D}^{0i}\}$ defines the space $\mathfrak{D}$. The pairwise intersections are

$$\mathfrak{D}^+ \cap \mathfrak{D}^- = M(z, u, \omega) = \mathfrak{D}^{\pm} \cap \mathfrak{D}^{0i}, \quad \mathfrak{D}^{0i} \cap \mathfrak{D}^{0j} = M(u, \omega), \tag{20}$$

where $M(z, u, \omega)$ is any multinomial in these variables and $M(u, \omega)$ is any multinomial in u and ω whose coefficients are nonanalytic in z in both domains $T^{\pm}$. We therefore redefine $\mathfrak{D}$ and all of its subspaces $\mathfrak{D}^i$ modulo the subspace $\mathcal{E}$, of functions which for fixed z are polynomial in u and ω, and for fixed u, ω are nonanalytic[6] in z in both domains $T^{\pm}$:

$$\mathcal{E} = \bigcup_{i \neq j} \mathfrak{D}^i \cap \mathfrak{D}^j. \tag{21}$$

We have then decomposed $\mathfrak{D}/\mathcal{E}$ into six disjoint invariant subspaces $\mathfrak{D}^i$ (we shall not explicitly write $\mathfrak{D}^i/\mathcal{E}$). These are the six irreducible subspaces of Graev; and on them the operators T_g^0 of (13) act irreducibly.

B. Scalar product

We assert that a possible scalar product for the series d_0^i is

$$\begin{aligned} (f, g)_i = \text{Reg.} \int_{T^i} \overline{f(y_c)} [\Omega(y_c)]^{-p-2} [T(z)]^{-K-1} \\ \times [U(y_c)]^{-q-2} g(y_c) d\mu(y_c), \end{aligned} \tag{22}$$

where Ω, U, and T are given by (18) and (15), and the integration is over the analyticity domain T^i appropriate to the representation d_0^i. (For example, over all ω and u, and $z \in T^+$ for the series d_0^+; or over $z \in T^0$ and the analyticity domain in the u, ω planes for one of the $d_0^{\mathcal{G}i}$).

Because of the transformation laws (16) and (19), this is invariant under all $g \in \mathcal{G}$; and because none of the quantities Ω, U, or T change sign in the domain T^i, the norm of a function $f \in \mathfrak{D}^i$ is of definite sign, provided the integral converges. Consider the asymptotic behavior of the integrand: Then we know from the appendix that this is

$$|f| \sim |z|^{-1} |\omega|^p |u|^q$$

(remember that $\mathfrak{D}$ is defined now modulo $\mathcal{E}$), so that the integral certainly converges at infinity. There remain only the divergences due to the zeros of T, Ω, and U, and the regularization is designed to remove these. We have not been able to cast this rather complicated functional into any more attractive form in the general case, but shall in the next section examine separately the two regions $T < 0$ and $T > 0$.

C. Discussion

We have now constructed the six unitary irreducible representations d_0^i of the principal discrete series of $\mathcal{G}$, by the operators T_g^Q acting on functions in one of the six spaces $\mathfrak{D}^i/\mathcal{E}$ with scalar product (22). Because the situation is a little involved, we make a few remarks on it here.

At first glance, the structure in ω and u of the spaces $\mathfrak{D}^\pm$ and $\mathfrak{D}^{0i}$ appears quite unrelated. But consider the two subspaces $\mathfrak{D}^{0i}$ of functions analytic in ω outside the circle $\Omega = 0$ in the ω plane, and suppose that T increases steadily from negative values. Then the circle shrinks to zero as T vanishes, and for positive values there is no trace of it left. If we follow the evolution of the spaces $\mathfrak{D}^{0i}$ in this process, we find that functions are eliminated with singularities progressively nearer the origin, until at last (i.e., for $T > 0$) the space embraces only functions which are analytic everywhere, that is, polynomials: The spaces have become $\mathfrak{D}^+$, and the subspaces of functions analytic within $\Omega = 0$ have disappeared entirely. This pictorial argument displays the similarity of structure of all the $\mathfrak{D}^i$.

It might be thought that since the surfaces $\Omega = 0 = U$ can be regarded as dividing the domain $T < 0$ rather than the u or ω planes, we could decompose $\mathfrak{D}^0$ into functions of z analytic in one of these four regions, for fixed ω and u. Irrespective however of the fact that only $\Omega U > 0$ define possible domains of analyticity, this decomposition cannot be invariant under $\mathcal{G}$ because the conjugates $\bar{u}$ and $\bar{\omega}$ must enter explicitly in order to specify the boundaries of the regions.

Now turn to the redefinition of $\mathfrak{D}$ modulo $\mathcal{E}$. This is entirely analogous to the well-known procedure[4] for the discrete series of $SL(2,R) \sim SU(1,1)$, where we take for carrier spaces of its representations the spaces of functions analytic in the upper or lower half of the complex plane, and then redefine them modulo their intersection—the space P of polynomials. Not only does this remove any common elements, but it also ensures convergence of the scalar product; indeed, P is the degeneracy subspace, and all elements thereof have vanishing norm. $\mathcal{E}$ plays a similar role.

This brings us to the scalar product (22). The form given there has the advantage of being applicable to all six subspaces $\mathfrak{D}^i$; but it is obviously extremely awkward to handle—in particular, we have not defined the exact regularization procedure. Indeed, we shall be able to extract several formally distinct functionals (we expect them to coincide on any irreducible subspace $\mathfrak{D}^i$, but this may not be immediately obvious) depending on our precise treatment of this problem: as a simple example of this, see Ref. 7, Chap. III, Sec. 2. Closer examination of the inner product reveals that a most important distinction must be drawn between the regions $T > 0$ and $T < 0$: In the latter case U and Ω have zeros, while in the former they do not. The regularization will therefore be quite different in these two regions, and we shall examine them separately in the next section, where we shall see too that $\mathcal{E}$ is indeed the degeneracy subspace and that the scalar product for $T < 0$ [and hence the representation (13)] can be extended to the case when p or q takes the value -1.

IV. ALTERNATIVE REALIZATIONS

A. The series $d_0^\pm$

We have not been able to introduce a nondegenerate scalar product into $\mathfrak{D}_0$ for the two series $d_0^\pm$, and so cannot discard altogether the spaces $\mathfrak{D}^\pm$. We can, however, show that the Hermitian functional defined by (22), when the "regularization" is effected by taking the residue of the generalized function at its singularity, is of definite sign; and also that it is degenerate upon the subspace of multinomials $M(z,u,\omega) = \mathfrak{D}^+ \cap \mathfrak{D}^-$. The demonstration is technical and we confine it to Appendix B.

B. The spaces D_0^{0i}

The situation here is of much greater interest, since a scalar product does exist upon the "boundary value" space $\mathfrak{D}_0^0$. We start by examining this space.

$\mathfrak{D}_0^0$ then is the boundary value of $\mathfrak{D}^0$. It is clear that the z dependence of a function in $\mathfrak{D}_0^0$ is restricted by the condition that it be not the boundary value of a function analytic in $T^\pm$; passing to the Fourier transform (as in I Sec. VB) we find that this implies[8] that $\tilde{f}(p)$ vanish for $p_b p_c + p_a \bar{p}_a < 0$. We shall return to this in the next section.

Now consider the dependence upon ω and $\bar{\omega}$, that is, upon ω and u. We have at first glance four regions $\Omega \gtrless 0$, $U \gtrless 0$ from which to take boundary values; but it is not difficult to see that when $u \to -\bar{\omega}$ and $z_c \to z$, then $\Omega U > 0$. This means that of the original four regions, only two survive, and we can decompose $\mathfrak{D}_0^0$ into two subspaces only: of functions which are boundary values of others analytic either in $\Omega > 0$ and $U > 0$, or in $\Omega < 0$ and $U < 0$.

This leaves a second reducibility to be sought; but we already know[1] of one such. Introduce the variables

$$\begin{aligned} &\alpha = a + b\bar{\omega}, \quad \bar{\alpha} \\ &\beta = b, \\ &\gamma = c + a\omega - \bar{a}\bar{\omega} + b\omega\bar{\omega}, \\ &\omega, \bar{\omega}; \end{aligned} \tag{23}$$

these parameters were denoted ρ, σ, and τ in I because the labels α, β, and γ were preempted, but the present

notation is preferable. Then we showed in I that transformations of the first principal continuous series d_1 preserved analyticity in $\beta + \bar{\beta} \gtrless 0$ and hence that $\mathfrak{D}_0(k, i\rho, m)$ reduced accordingly. Clearly the same is true for the representations (13) and the space $\mathfrak{D}_0^0(K, L, m)$, and so each of the two u, ω-boundary-value spaces can in turn be split into boundary values of functions analytic in $\mathrm{Re}\beta > 0$ or $\mathrm{Re}\beta < 0$.

This then gives us four subspaces of $\mathfrak{D}_0^0$. Notice that the β-reducibility did not appear in the last section, where we dealt with $\mathfrak{D}^0$ itself, because the regions $\beta + \bar{\beta} \gtrless 0$ are invariant only when all the remaining parameters take their boundary values; its appearance here is perhaps a consequence of the fact that the coefficient of the highest power of $|u|$ in U or $|\omega|$ in Ω is just $(\beta + \bar{\beta})$.

C. Scalar product upon D_0^{0i}

We introduce the four parameters N_i:

$$\begin{aligned} 2N_1 &= L - K + m, & 2N_3 &= L + K - m, \\ 2N_2 &= L - K - m, & 2N_4 &= L + K + m, \end{aligned} \tag{24}$$

and the notation $w^{A*B} \equiv \omega^A \bar{\omega}^B$, which will be a useful shorthand. Since $N_1 = q + 1$, $N_2 = p + 1$, all the N_i are nonnegative integers. We now consider the sesquilinear form

$$(f,g) \equiv \int \overline{f(y)}\, \partial_\omega^{N_1 * N_2}\, \partial_\beta^L \partial_\omega^{N_3 * N_4} \partial_\beta^K g(y)\, d\mu(y), \tag{25}$$

where the notation of the left-hand side anticipates its role of scalar product. The (partial) derivates are

$$\begin{aligned} \partial_\omega &= \frac{\partial}{\partial \omega}\Big|_{a,b,c} \\ -i\partial_\beta &= \omega \frac{\partial}{\partial \bar{a}} - \bar{\omega} \frac{\partial}{\partial a} + \frac{\partial}{\partial b} + \omega\bar{\omega} \frac{\partial}{\partial c} \\ &= \frac{\partial}{\partial \beta}\Big|_{\alpha,\gamma,\omega} \quad [\text{cf. (23)}] \end{aligned} \tag{26}$$

and do not commute; ∂_β is the operator denoted δ in (I. 28) sqq. Notice that these are just derivatives with respect to the parameters ω (I. 20) and x (I. 25) that do not belong to the manifolds Z and Z.

We assert then that (25) is invariant under all $g \in \mathfrak{G}$. To prove this, we remark that by using the set $\{g_i\}$ we can show that the operators (26) transform as

$$\begin{aligned} g : \partial_\omega &\to \partial'_\omega = \lambda^2 \Delta^{-1} \partial_\omega, \\ g : \partial_\beta &\to \partial'_\beta = \Delta^2 (\lambda\lambda_1)^{-1} \partial_\beta; \end{aligned} \tag{27}$$

then invariance of (25) under $g_i = z, d, \sigma$ and $\xi \in a$ follows almost trivially, and we are left with only the two elements $I \in a$ and B (or J). Now under $I, f \in \mathfrak{D}$ transforms as

$$\begin{aligned} T_I^0 : f(a,b,c,\omega) &= |\omega|^{m+L-K-2}\, \omega^{-m} f(\bar{a}, c, b, -\omega^{-1}) \\ &= \omega^{N_2 - 1}\, \bar{\omega}^{N_1 - 1} f(y') \end{aligned} \tag{28}$$

and $\partial'_\beta = |\omega|^{-2} \partial_\beta$, $\partial'_\omega = \omega^{-2} \partial_\omega$ from (27); hence the form (25) becomes

$$\begin{aligned} (T_I^0 f, T_I^0 g) = \int\ & \omega^{-N_1 - 1}\, \bar{\omega}^{-N_2 - 1}\, \overline{f(y)} \\ & \times [\omega^2 \partial_\omega]^{N_1 * N_2} [|\omega|^{-2} \partial_\beta]^L [\omega^2 \partial_\omega]^{N_3 * N_4} [|\omega|^{-2} \partial_\beta]^K \\ & \times \{\omega^{1 - N_2}\, \bar{\omega}^{1 - N_1} f(y)\}\, d\mu(y). \end{aligned} \tag{29}$$

[We have used the Jacobian of (I. 14), written this in terms of the transformed variables y', and dropped primes.] Observing that $N_2 + K = N_3$, etc., we use the identity

$$\left[x^2 \frac{d}{dx}\right]^p x^{1-\mu} g(x) = x^{1+\mu} g^{(\mu)}(x) \tag{30}$$

several times and deduce finally that

$$(T_I^0 f, T_I^0 g) = (f, g).$$

We can examine the action of B in a similar way, and once again we obtain invariance. Therefore (25) is invariant under all the g_i and hence under all $g \in \mathfrak{G}$. We must now show that this functional is both Hermitian and positive-definite, if we are indeed to use it as a scalar product. The first is simple: Mere commutation reveals that

$$G \equiv \partial_\omega^{N_1 * N_2}\, \partial_\beta^L\, \partial_\omega^{N_3 * N_4}\, \partial_\beta^K = \partial_\beta^K \partial_\omega^{N_4 * N_3} \partial_\beta^L \partial_\omega^{N_2 * N_1} = G^\dagger \tag{31}$$

and so the Hermiticity is manifest.

We have not, however, been able to find a general proof of the positivity. It is easy to show that if $L = K + |m|$ then the differential operator becomes

$$\begin{aligned} G &= \partial_\beta^K \partial_\omega^L \partial_\beta^{L-K} \partial_{\bar{\omega}}^L \partial_\beta^K, \quad m > 0, \\ &= \partial_\beta^K \partial_{\bar{\omega}}^L \partial_\beta^{L-K} \partial_\omega^L \partial_\beta^K, \quad m < 0; \end{aligned} \tag{32}$$

this defines a norm in accordance with (25),

$$\| f \|^2 = \int \overline{f(y)}\, G f(y)\, d\mu(y), \tag{33}$$

which is clearly of definite sign if $L - K$ is even, and is indeed definite too if $L - K$ is odd, by virtue of the analyticity of f in one of the two half-planes $\mathrm{Re}\beta \gtrless 0$. [Compare the treatment of $SL(2,R)$ in Ref. 4.] Notice that this special case corresponds to p or q having the extreme value of -1, and hence defines an extension of the representation (13) of the four series d_0^{0i} to those values of the parameters. Similarly, if $K = 0$, we obtain

$$G = \partial_\omega^{(L+m)/2}\, \partial_{\bar{\omega}}^{(L-m)/2}\, \partial_\beta^L\, \partial_{\bar{\omega}}^{(L-m)/2}\, \partial_{\bar{\omega}}^{(L+m)/2}, \tag{34}$$

which is manifestly of definite sign; but for other combinations of L, K, and m we cannot put forward a general proof. We have, however, examined a very large number of special cases, and in every one the norm turns out to be definite upon the $\mathfrak{D}_0^{0i}$, involving sums of terms including powers up to order K of the second order operator $(\partial_a \partial_{\bar{a}} + \partial_b \partial_c)$ when we rearrange the derivatives in G. Some typical examples are

$$\begin{aligned} &L = 6, K = 2, m = 0, \quad p = 1 = q, \\ &G = \omega^2 \bar{\omega}^2 \beta^2 [\beta^2 \omega^2 \bar{\omega}^2 \beta^2 - 16(a\bar{a} + bc)\beta\omega\bar{\omega}\beta \\ &\qquad + 72(a\bar{a} + bc)^2] \beta^2 \omega^2 \bar{\omega}^2, \\ &\qquad L = 5, K = 2, m = -1, \quad p = 0, q = 1, \\ &G = \bar{\omega}^2 \omega\beta [\beta^2 \omega^2 \beta \bar{\omega}^2 \beta^2 - 12(a\bar{a} + bc)\beta\omega\beta\bar{\omega}\beta \\ &\qquad + 36(a\bar{a} + bc)^2 \beta] \beta\bar{\omega}\omega^2, \end{aligned} \tag{35}$$

where we have strictly temporarily written the argument instead of the derivative, eg., $\partial_\beta \to \beta$ etc., in order to emphasize the essential features. We remark that the definiteness of sign (whether positive or negative depends upon the parameters and upon which space $\mathfrak{D}_0^{0i}$

we are considering) rests upon both the analyticity in the variable β and the fact that the Fourier transform (with respect to z) $\tilde{f}(p,\omega)$ of f vanishes when $(p_a\bar{p}_a + p_b p_c) < 0$, so that $-(\partial_a\partial_{\bar{a}} + \partial_b\partial_c)f$ is always of the same sign as f itself.[9] Notice that this differential operator commutes with all the others in the above expressions. While the investigation of particular instances does not furnish a general proof, nonetheless we feel that the special cases we have considered do not have any specially restrictive features, and that the norm (33) is indeed of definite sign upon any one irreducible subspace $\mathfrak{D}_0^{0i}$. In this respect our investigation is incomplete, and a correct proof would be welcome.

Finally, consider the kernel of G: This will be the degeneracy subspace. A full description of it is to be found in Appendix A; here we only note that it does indeed contain the space $\mathcal{E}$ introduced in the last section (more correctly, it contains its boundary-value space from $\mathfrak{D}$ onto $\mathfrak{D}_0$), so that redefining $\mathfrak{D}$ modulo $\mathcal{E}$ is indeed just equivalent to redefining it modulo those elements with norm zero. It is appropriate to point out here that we cannot define a scalar product for $\mathfrak{D}_0^{\pm}$ in this way because G annihilates that entire space.

V. REDUCTION UNDER $\mathbb{P}$

There are apparently two problems involved in the reduction of the six series d_0^i when restricted to $\mathbb{P}$: the structure of the carrier spaces $\mathfrak{D}^i$ (or of their Hilbert space completions under the scalar product, $\mathfrak{K}_0^i$) gives rise to restrictions upon the functions involved, and the scalar product itself is quite different from any that we are accustomed to. For the first principal continuous series $d_1^{\pm}$, we explicitly removed this latter difficulty in I (Sec. VB) by essentially using the square root of the differential operator in the scalar product (I. 28) to map $\mathfrak{K}_1(k, i\rho, m)$ onto a space isomorphic to $\mathfrak{K}_2(-k-1, i\rho, m)$; since the whole process was invariant under $g \in \mathbb{P} \subset \mathcal{G}$, we were left with an inner product without derivatives, which coincided in value with the original one and was invariant under $\mathbb{P}$ (but not under arbitrary $g \in \mathcal{G}$). The situation with the discrete series is less simple, and we cannot explicitly define such an operator; however, the problem is essentially superficial, since the representation is actually specified by the operators T_g^0 and the spaces $\mathfrak{D}_0^i/\mathcal{E}$, and so it is in fact unnecessary to concern outselves with it.

The first problem however is very pertinent, and in this section we investigate how the representations of $\mathbb{P}$ which occur are governed by the structure of the $\mathfrak{D}_0^i$. The appearance of $\mathbb{P}$ in this matrix realization was studied in I, Sec. 5, and we shall not discuss it further here; we do however remind ourselves of the alternative parametrization (I. 15) of Z:

$$a = ix_1 - x_2, \quad b = i(x_0 + x_3), \quad c = i(x_0 - x_3), \quad d = ix_1 + x_2 \tag{36}$$

and notice that the complexified manifold Z_c can be parametrized by letting the four real variables x_μ become complex, $x_\mu \to x_\mu + iy_\mu$.

A. The series $d_0^{\pm}$

Recall that $\mathfrak{D}^{\pm}$ is characterized by analyticity in $T^{\pm}$ and polynomial behavior in u and ω. Consider first the former restriction, and introduce the parametrization (36) of Z: Then the $T^{\pm}$ become

$$T^{\pm} = \{x_\mu + iy_\mu \mid y_0^2 - y_1^2 - y_2^2 - y_3^2 > 0,\ y_0 \gtrless 0\}. \tag{37}$$

Since the variables x_μ are indeed just the space–time coordinates, the $T^{\pm}$ are the forward and backward tube domains. The condition of analyticity in $T^{\pm}$ then tells us[8] that the Fourier transform[10] $\tilde{f}(p_\mu, \omega)$ of $f(x_\mu, \omega) \in \mathfrak{D}_0^+$ is concentrated in the region $p_\mu p^\mu > 0$, $p^0 > 0$: That is, only timelike momenta with positive energies can occur. For T^- we can have only negative energies.

Now turn to the ω dependence of $f \in \mathfrak{D}^{\pm}$. This is polynomial, of degree p in ω and q in $\bar{\omega}$; and so f transforms under a nonunitary finite-dimensional representation of $SL(2,C) \subset \mathbb{P}$ when regarded as a function of $\omega, \bar{\omega}$ alone.[11] Indeed, introduce the usual notation[12] (j_1, j_2) to specify such a finite-dimensional representation of $SL(2,C)$; then f transforms under $(j_1 = \frac{1}{2}p, j_2 = \frac{1}{2}q)$, which is just $(p+1)(q+1)$-dimensional. The spin content of such a representation is well known: All spins are included between $|j_1 - j_2|$ and $(j_1 + j_2)$. We therefore have

Theorem 3: When restricted to $\mathbb{P}$, the representation d_0^+ (d_0^-) of the principal discrete series of $\mathcal{G}$ associated with the parameters (K,p,q) contains a direct sum and integral over all the principal series representations of $\mathbb{P}$ with real and positive (negative) rest mass and spins s satisfying

$$|p - q| \leq 2s \leq (p + q).$$

Each representation enters with unit multiplicity.

B. The series d_0^{0i}

We consider the boundary-value spaces $\mathfrak{D}_0^{0i}$ of Sec. IVB rather than their complexifications, since it is the manifold Y rather than Y_c which has the physical significance. We see immediately that the lack of analyticity of $f \in \mathfrak{D}^{0i}$ in either of the tube domains $T^{\pm}$ implies[8] that $\tilde{f}(p_\mu, \omega)$ is concentrated[10] upon $p_\mu p^\mu < 0$—that is, that only spacelike[4] momenta can occur. As in I, analyticity in β [c.f. (23)] in the half-plane $\mathrm{Re}\beta > 0$ implies that for vanishing $\omega, \bar{\omega}$ the transform $\tilde{f}(p_\mu, 0)$ vanishes too if $(p_0 + p_3) < 0$: This does not impose any restraints upon the representations of the Poincaré group which occur, but only upon the basis functions.

We are left with the analyticity in $\Omega, U \gtrless 0$, and once again it is clear that no restrictions upon the actual representations of $\mathbb{P}$ are implied. There remains only the covariance implied by the representation label $m = q - p$ (see I, section Va), and this is dealt with as before. We obtain

Theorem 4: When restricted to $\mathbb{P}$, each representation d_0^{0i} of the principal discrete series of $\mathcal{G}$ associated with the parameters (K,p,q) contains a direct sum and integral over all the principal series representations of $\mathbb{P}$ with imaginary rest mass (i.e., spacelike 4-momenta) which allow a helicity of $\frac{1}{2}(q-p)$. Each representation occurs with unit multiplicity.

VI. COMMENTS

We have now constructed and reduced all the six representations d_0^i of the principal discrete series, but certain problems still remain and have not been completely solved in this paper. Among these are the following:
(i) What is the precise connection between the four

domains $\Omega, U \gtrless 0$ of Y_c for $T < 0$ and the introduction of analyticity in β when we restrict ourselves to Y?
(ii) The representations d_0^{0i} are defined for p and/or q taking the value -1 (see Sec. IVC); apparently the $d_0^{\pm}$ are not—we cannot have polynomials of this degree.
(iii) We have not found a proof that the norm defined by (25) is actually of definite sign on $\mathfrak{D}_0^{0i}/\mathcal{E}$ for general parameter values.[13]
(iv) Exactly what restrictions upon the matrix elements of $\mathbb{P}$ needed to expand $f \in \mathfrak{D}_0^{0i}$ are implied by the analyticity in Ω, $U \gtrless 0$?

Resolutions of these questions would be of interest, and we hope to return to them at a future date. Finally, we summarize the results we have obtained in these two papers on the principal series of representations of $\mathcal{G}$:

Series	Analyticity	K, L, m	Reduces to
d_2	—	$i\rho_1, i\rho_2, m$	all p_μ
d_1^+	Analytic in $\mathrm{Re}\beta > 0$	$K \geq -1, i\rho, m$	$p_0 + p_3 > 0$[14]
d_1^-	Analytic in $\mathrm{Re}\beta < 0$	$K \geq -1, i\rho, m$	$p_0 + p_3 < 0$[14]
d_0^+	Analytic in T^+ Polynomial in $\omega, \bar{\omega}$	$K + L + m$ even $L - K - \lvert m\rvert \geq 2$	$p^2 > 0$ $p > 0$ $m \leq 2S \leq L-K-2$
d_0^-	As for d_0^+ but analytic in T^-		$p^2 > 0$ $p_0 < 0$ $m \leq 2S \leq L-K-2$
d_0^{0i}	Nonanalytic in $T^\pm$ See Secs IIIA, IVB	$K + L + m$ even $L - K - \lvert m\rvert \geq 0$	$p^2 < 0$.

Where no other restriction is indicated, the spins S which occur are all those allowing a helicity of $\frac{1}{2}m$.

APPENDIX A: STRUCTURE OF CARRIER SPACE D_0

We consider the three transformations I, B, and J introduced in Sec. 2, and find that in the complexified space $\mathfrak{D}(y_c)$,

$$T_J : f = \Delta^{K-1}(\omega a + c)^p (ud - c)^q f\left(\frac{a}{\Delta}, \frac{c}{\Delta}, \frac{b}{\Delta}, \frac{d}{\Delta}; -\frac{\omega b + d}{\omega a + c}, \frac{ub - a}{ud - c}\right), \tag{A1}$$

$$T_B f = (ib)^{K-1} f\left(\frac{a}{ib}, \frac{1}{b}, \frac{bc - ad}{b}, \frac{d}{ib}; i(\omega b + d), i(ub - a)\right).$$

$$T_I f(a, b, c, d; \omega, u) = \omega^p u^q f(-d, c, b, -a; -\omega^{-1}, -u^{-1}),$$

where $\Delta = bc - ad$. Then all these transformed functions must be C^∞; in particular, finite at all finite values of the arguments. This implies that asymptotically, f behaves as

$$f \sim \{Z, (ad - bc)\}^{K-1} \{u, (a - bu), (c - du)\}^q \times \{w, (\omega a + c), (\omega b + d)\}^p, \tag{A2}$$

where, e.g.,

$$\{x, y, z\}^r = \sum_{j+k+l=r} C_{jkl}\, x^j y^k z^l. \tag{A3}$$

For the two spaces $\mathfrak{D}^\pm$, the latter two brackets are multinomials; the first is an asymptotic series, which by virtue of its definition can actually be differentiated term-by-term to yield such a series for the derivate of the function itself. For the remaining four spaces all these are asymptotic series. We observe that the fastest that a function can grow is

$$f \sim Z^{2K+p+q-2} u^q \omega^p. \tag{A4}$$

Now let us consider the kernel of the differential operator G introduced in (31). By a process of repeated integration one can show that this consists of all functions of the form

$$f_0 = M_1^{L-1}(\omega, \bar{a}, c) + M_2^{L-1}(\bar{\omega}, a, c) + M_3^{L+K-1}(Z) + P_1^p(\omega) + P_2^q(\bar{\omega}), \tag{A5}$$

where M_1 is a (nonhomogeneous) multinomial in $\omega, \bar{a}, c$ of total degree $L - 1$, whose coefficients are quite arbitrary functions of $\bar{\omega}, a$, and b; P_1 is a polynomial of degree p in ω, whose coefficients are analytic in $\bar{\omega}$ (i.e., in u) and arbitrary in z. The remaining terms are similarly defined. (We have restricted ourselves to the manifold Y here so as not to have extra, irrelevant terms in $\bar{\omega}$ and $\bar{u}$ separately).

The kernel of G is clearly not contained entirely in $\mathfrak{D}_0$. If we restrict it to $\mathfrak{D}_0$, however, we see that it contains multinomials in z and polynomials in ω and $\bar{\omega}$, and hence contains the intersection subspace $\mathcal{E} \subset \mathfrak{D}_0$ of (21); unfortunately, it also contains the entire subspaces $\mathfrak{D}_0^\pm$, and so G cannot be used in the definition of an inner produce for those spaces.

APPENDIX B: POSITIVITY OF SCALAR PRODUCT FOR $d_0^\pm$

We make the substitutions

$$\begin{aligned} (a + \bar{d}) &= y_1 + iy_2, \quad & y_0 &= r, \\ (b + \bar{b}) &= y_0 + y_3, \quad & \mathbf{y} &= rt\hat{\mathbf{e}}, \\ (c + \bar{c}) &= y_0 - y_3, \end{aligned} \tag{B1}$$

where $0 \leq t < 1$ and $\hat{\mathbf{e}}$ is a unit 3-vector. Then (22) implies that the norm of a function $f \in \mathfrak{D}^+$ is given by (we omit irrelevant numerical factors)

$$\|f\|^2 = \int d^4z \int_0^\infty r^{-2K-p-q-3}\, dr \times \int_0^1 (1 - t^2)^{-K-1} t^2 \Phi_1(y, r, t)\, dt, \tag{B2}$$

where

$$\begin{aligned} \Phi_1(y, r, t) = \int \{&(\omega\bar{\omega} + 1) + t[e_1(\omega + \bar{\omega}) \\ &+ ie_2(\omega - \bar{\omega}) + e_3(\omega\bar{\omega} - 1)]\}^{-p-2} \quad \{(u\bar{u} + 1) \\ &- t[e_1(u + \bar{u}) + ie_2(u - \bar{u}) - e_3(u\bar{u} - 1)]\}^{-q-2} \\ &\times |f(z, y, w, u)|^2 D\omega\, Du\, d^2\Omega_e. \end{aligned} \tag{B3}$$

Since there are no singularities in the region of integration, $\Phi_1 < \infty$, we suppose further that f is such that $(1 - t^2)^{-K-1}\Phi_1$ is nonsingular at $t = 1$; then it is certainly regular at all other points, and the integral in (B2) converges. (This assumption is essentially technical). The r integration has a simple pole when K is an integer, and we can take its residue as the integrand of the final expression, to obtain

$$\|f\|^2 = \int d^4z \left(\frac{d}{dr}\right)^{L+K} \left(\int_0^1 (1 - t^2)^{-K-1} t^2 \Phi_1(y, r, t)\, dt\right)\Big|_{r=0}. \tag{B4}$$

The only quantity depending upon r is $|f|^2$. Therefore this becomes

$$\int d^4z \int_0^1 (1 - t^2)^{-K-1} t^2\, dt \int \Omega_1^{-p-2} U_1^{-q-2} \times \left[\left(\frac{d}{dr}\right)^{L+K} \left|\tilde{f}(z, \hat{\mathbf{e}}, r, t, \omega, u)\right|^2\right]_{r=0} d^2\Omega_e D\omega Du, \tag{B5}$$

where Ω_1 and U_1 are the expressions in braces in (B3), and $\tilde{f}$ is f expressed as a function of the new parameters.

We see at once that if $\tilde{f}$ is polynomial in r, that is, if f is multinomial in z_c, this vanishes; and this is independent of our assumption of the regularity of $(1-t^2)^{-K-1}\Phi_1(y,r,t)$. Therefore $M(z,\omega,u)$ is indeed the degeneracy subspace. To show that the form is of definite sign, we recall that $f(z_c)$ is analytic in T^+. It is then easy to show by considering the Fourier transform that $(d/dr)^N|f|^2$ is itself of definite sign (whether positive or negative depends upon N), and this solves the problem.

---

[1]N. W. Macfadyen, J. Math. Phys. (N.Y.) **12**, 1436 (1971).
[2]M. I. Graev, Tr. Mosk. Mat. Obshch. 7, 335 (1958); Am. Math. Soc. Transl. 66, 1 (1968).
[3]If A is a self-adjoint matrix, by $A > 0$ we mean that A is positive definite, that is, its eigenvalues are all positive.
[4]I. M. Gel'fand, M. I. Graev, and N. Ya. Vilenkin, *Generalized Functions* (Academic, New York, 1966) Vol. 5.
[5]L. Castell, J. Math. Phys. (N.Y.) 11, 2999 (1970).
[6]We include in the nonanalytic terms all functions which are polynomial in one variable with the others fixed, even though this analyticity is actually preserved under the group.
[7]I. M. Gel'fand and G. E. Shilov, *Generalized functions* (Academic, New York, 1964), Vol. 1.
[8]H. Bremermann, *Distributions, complex variables and Fourier transforms* (Addison-Wesley, Reading, Mass. 1965).
[9]The minus signs here and in (35) arise because ∂_β was defined in (26) to be $i\,\partial/\partial\beta$; whereas $\partial_\alpha = \partial/\partial\alpha$, etc.
[10]The transform exists because $\mathfrak{D}$ is defined only modulo $\mathcal{E}$.
[11]Our scalar product restores unitarity by introducing the elements of Z.
[12]I. M. Gel'fand, R. A. Minlos, and Z. Ya. Shapiro, *Representations of the rotation and Lorentz groups* (Pergamon, New York, 1963).
[13]See, however, paper III in this series, J. Math. Phys. (to be published).
[14]This is to be understood, of course, only in a frame where $\omega = 0$.

Coordinate independent dyadic formulation of wave normal and ray surfaces of general anisotropic media

Ismo V. Lindell
*Department of Electrical Engineering, Helsinki University of Technology, Otaniemi, Finland**
(Received 13 December 1971; revised manuscript received 9 May 1972)

Dyadic algebra has been applied to the problem of propagation of an electromagnetic discontinuity in a general lossless anisotropic medium ($\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ positive definite symmetric dyadic quantities). It is seen that the wave normal and ray equations and, hence, the equations of the wave normal and ray surfaces, take a coordinate-independent explicit form. The optical axes of the medium are also discussed.

INTRODUCTION

Dyadics were introduced by Gibbs[1] some 70 years ago as an extension of the vector notation to mappings in 3-space. Since then, not being restricted to a vector space of three dimensions, matrix and tensor notations have outweighed dyadic notation in physics. The power of the dyadic notation however lies in its independence of any coordinate system and in a proper use of the multiple products as introduced by Gibbs. Among the areas in which dyadic algebra appears well suited are electromagnetic field problems in anisotropic media.

The purpose of this paper is to present a dyadic formulation of the well-known problem of propagation of an electromagnetic discontinuity in a general anisotropic medium. If $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ are the constitutive dyadics, the earlier considerations known to this author have applied matrix calculus in a coordinate system in which the matrix $[\bar{\bar{\mu}}]^{-1/2}[\bar{\bar{\epsilon}}][\bar{\bar{\mu}}]^{-1/2}$ is diagonal.[2] Hence, the resulting equations of wave-normal and ray surfaces do not show the functional dependence on the constitutive parameters but instead, give us an algorithm to calculate the equations for any given $[\bar{\bar{\epsilon}}]$ and $[\bar{\bar{\mu}}]$. The dyadic presentation which follows, is seen to result in an explicit expression on $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$, which is independent of any coordinate system.

WAVE NORMAL AND RAY EQUATION

As a starting point we apply the equations for a propagating electromagnetic discontinuity[2]

$$\nabla\psi \times \bar{H}^* = -\bar{D}^*, \tag{1}$$

$$\nabla\psi \times \bar{E}^* = \bar{B}^*, \tag{2}$$

where $\psi(\bar{r})$ is the wavefront function and $\bar{E}^*, \bar{H}^*, \bar{D}^*, \bar{B}^*$ are field vectors at the time $t = \psi(\bar{r})$. Denoting $\nabla\psi = \bar{p}$ and defining a vector $\bar{s}$, called the ray vector, by $\bar{s} \times (\bar{E}^* \times \bar{H}^*) = 0$ and $\bar{s}\cdot\bar{p} = 1$, we have from Eqs. (1) and (2)

$$\bar{p} \times \bar{H} = -\bar{D}^*, \tag{3}$$

$$\bar{p} \times \bar{E}^* = \bar{B}^*, \tag{4}$$

$$\bar{s} \times \bar{B}^* = -\bar{E}^*, \tag{5}$$

$$\bar{s} \times \bar{D}^* = \bar{H}^*. \tag{6}$$

The constitutive relations are supposed to be of the form

$$\bar{D} = \bar{\bar{\epsilon}}\cdot\bar{E}, \tag{7}$$

$$\bar{B} = \bar{\bar{\mu}}\cdot\bar{H}, \tag{8}$$

where $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ are symmetric positive definite dyadics.[3] Because this treatment uses the geometrical optics approximation, the dyadics $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ must be considered as algebraic quantities, i.e., they do not contain differential operations (nondispersive medium).[4] Now, from Eqs. (3), (4), (7), and (8), the field quantities $\bar{B}^*, \bar{D}^*$ and either $\bar{E}^*$ or $\bar{H}^*$ can be eliminated, giving two equations:

$$[\bar{\bar{\epsilon}} - \bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{p}\,\bar{p}]\cdot\bar{E}^* = 0, \tag{9}$$

$$[\bar{\bar{\mu}} - \bar{\bar{\epsilon}}^{-1} \overset{\times}{\times} \bar{p}\,\bar{p}]\cdot\bar{H}^* = 0. \tag{10}$$

It is to be noted that the existence of $\bar{\bar{\epsilon}}^{-1}$ and $\bar{\bar{\mu}}^{-1}$ follows from the assumed positive definite property of $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$. Correspondingly, from Eqs. (5)-(8) we have

$$[\bar{\bar{\epsilon}}^{-1} - \bar{\bar{\mu}} \overset{\times}{\times} \bar{s}\,\bar{s}]\cdot\bar{D}^* = 0, \tag{11}$$

$$[\bar{\bar{\mu}}^{-1} - \bar{\bar{\epsilon}} \overset{\times}{\times} \bar{s}\,\bar{s}]\cdot\bar{B}^* = 0. \tag{12}$$

The conditions for $\bar{p}$ and $\bar{s}$ for the nontrivial case arising from Eqs. (9)-(12) can be evaluated using dyadic identities (see Appendix), whence it is seen that (9) and (10) give us the same relation

$$\det(\bar{\bar{\epsilon}} - \bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{p}\,\bar{p}) = \det(\bar{\bar{\mu}} - \bar{\bar{\epsilon}}^{-1} \overset{\times}{\times} \bar{p}\,\bar{p}) = 0, \tag{13}$$

and that, we must also have

$$\det(\bar{\bar{\epsilon}}^{-1} - \bar{\bar{\mu}} \overset{\times}{\times} \bar{s}\,\bar{s}) = \det(\bar{\bar{\mu}}^{-1} - \bar{\bar{\epsilon}} \overset{\times}{\times} \bar{s}\,\bar{s}) = 0. \tag{14}$$

These equations are of sixth degree in $\bar{p}$ and $\bar{s}$, but use of the dyadic identity

$$\det(\bar{\bar{A}} + \bar{\bar{B}}) = \det\bar{\bar{A}} + \tfrac{1}{2}\,\bar{\bar{A}} \overset{\times}{\times} \bar{\bar{A}} : \bar{\bar{B}} + \tfrac{1}{2}\,\bar{\bar{A}} : \bar{\bar{B}} \overset{\times}{\times} \bar{\bar{B}} + \det\bar{\bar{B}} \tag{15}$$

and others (see Appendix) leaves us with equations of fourth degree:

$$\frac{(\bar{\bar{\mu}} : \bar{p}\bar{p})(\bar{\bar{\epsilon}} : \bar{p}\bar{p})}{(\det\bar{\bar{\mu}})(\det\bar{\bar{\epsilon}})} - \bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{\bar{\epsilon}}^{-1} : \bar{p}\,\bar{p} + 1 = 0, \tag{16}$$

$$(\bar{\bar{\mu}}^{-1} : \bar{s}\,\bar{s})(\bar{\bar{\epsilon}}^{-1} : \bar{s}\,\bar{s})(\det\bar{\bar{\mu}})(\det\bar{\bar{\epsilon}}) - \bar{\bar{\mu}} \overset{\times}{\times} \bar{\bar{\epsilon}} : \bar{s}\,\bar{s} + 1 = 0. \tag{17}$$

Equations (16) and (17) constitute the wave normal and ray equations for a propagating discontinuity in a general anisotropic medium in dyadic notation. They are seen to be coordinate independent.

From Eq. (16), the dispersion equation of a time-harmonic plane wave propagating in the anisotropic medium can be easily deduced, replacing $\bar{p}$ by $\bar{k}/\omega$, where $\bar{k}$ denotes the wavevector. Equation (16), then, is valid for temporally and spatially dispersive media, i.e., $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ may be

functions of ω as well as $\bar{k}$. The following conclusions, however, do not apply for spatially dispersive media.

A dyadic dispersion equation corresponding to Eq. (16) for a plane wave propagating in a medium with $\bar{\bar{\mu}} = \mu_0 \bar{\bar{I}}$ and $\bar{\bar{\epsilon}} = \epsilon_0 \bar{\bar{I}}$, has been presented before.[5] (It has been attained through an indirect attack via tensor calculus.)

Also, plane wave propagation in a medium with both $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ dyadics has been considered.[6] The resulting dispersion equation differs from our Eq. (16) with $\bar{p}$ replaced by $\bar{k}/\omega$ mainly only in the definition of the double dot product which is different from the usual [see Appendix (A4)].

WAVE NORMAL AND RAY SURFACES

Denoting $\bar{p} = p\bar{u}$ and $\bar{s} = s\bar{v}$, where $\bar{u}$ and $\bar{v}$ are real unit vectors, we may write (16) and (17) in the form

$$(1/p)^4 - (\bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{\bar{\epsilon}}^{-1} : \bar{u}\bar{u})(1/p)^2 + \tfrac{1}{4}(\bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{\bar{\mu}}^{-1} : \bar{u}\bar{u})(\bar{\bar{\epsilon}}^{-1} \overset{\times}{\times} \bar{\bar{\epsilon}}^{-1} : \bar{u}\bar{u}) = 0, \quad (18)$$

$$(1/s)^4 - (\bar{\bar{\mu}} \overset{\times}{\times} \bar{\bar{\epsilon}} : \bar{v}\bar{v})(1/s)^2 + \tfrac{1}{4}(\bar{\bar{\mu}} \overset{\times}{\times} \bar{\bar{\mu}} : \bar{v}\bar{v})(\bar{\bar{\epsilon}} \overset{\times}{\times} \bar{\bar{\epsilon}} : \bar{v}\bar{v}) = 0. \quad (19)$$

Now it can be shown[7] that if the dyadics $\bar{\bar{A}}$ and $\bar{\bar{B}}$ are real, symmetric and positive definite, then $\bar{\bar{A}} \overset{\times}{\times} \bar{\bar{B}}$ is symmetric and positive definite and the following inequality,

$$(\bar{\bar{A}} \overset{\times}{\times} \bar{\bar{B}} : \bar{w}\bar{w})^2 \geq (\bar{\bar{A}} \overset{\times}{\times} \bar{\bar{A}} : \bar{w}\bar{w})(\bar{\bar{B}} \overset{\times}{\times} \bar{\bar{B}} : \bar{w}\bar{w}), \quad (20)$$

is valid for all real vectors $\bar{w}$. [For $B \neq 0$ the equality sign implies the existence of a scalar α such that $(\bar{\bar{A}} + \alpha\bar{\bar{B}}) \overset{\times}{\times} \bar{w}\bar{w} = 0$)]. Applying this to Eqs. (18) and (19) we see that we are able to write

$$1/p = \{\tfrac{1}{2}(\bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{\bar{\epsilon}}^{-1} : \bar{u}\bar{u}) \pm \tfrac{1}{2}[(\bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{\bar{\epsilon}}^{-1} : \bar{u}\bar{u})^2 - (\bar{\bar{\mu}}^{-1} \overset{\times}{\times} \bar{\bar{\mu}}^{-1} : \bar{u}\bar{u})(\bar{\bar{\epsilon}}^{-1} \overset{\times}{\times} \bar{\bar{\epsilon}}^{-1} : \bar{u}\bar{u})]^{1/2}\}^{1/2} \quad (21)$$

$$1/s = \{\tfrac{1}{2}(\bar{\bar{\mu}} \overset{\times}{\times} \bar{\bar{\epsilon}} : \bar{v}\bar{v}) \pm \tfrac{1}{2}[(\bar{\bar{\mu}} \overset{\times}{\times} \bar{\bar{\epsilon}} : \bar{v}\bar{v})^2 - (\bar{\bar{\mu}} \overset{\times}{\times} \bar{\bar{\mu}} : \bar{v}\bar{v})(\bar{\bar{\epsilon}} \overset{\times}{\times} \bar{\bar{\epsilon}} : \bar{v}\bar{v})]^{1/2}\}^{1/2} \quad (22)$$

and $1/p, 1/s$ are seen to have two positive real roots $1/p_-, 1/p_+, 1/s_-, 1/s_+$ [the subscripts referring to the $\pm$ signs in Eqs. (21) and (22)]. From (20) it also follows that $p_- \geq p_+$, $s_- \geq s_+$, whence the surfaces $p_+(\bar{u})$, $s_+(\bar{v})$ have no points outside the surfaces $p_-(\bar{u})$, $s_-(\bar{v})$, respectively.

If the surfaces p_+ and p_- have points in common, the discriminant in Eq. (21) must vanish, which corresponds to the equality sign in the inequality (20). It can be shown[7] that the equality sign is valid in (20) only if there exist a scalar α and a vector $\bar{w}$ such that $(\bar{\bar{A}} + \alpha\bar{\bar{B}}) \overset{\times}{\times} \bar{w}\bar{w} = 0$. So, there exist optical axes $\bar{u}_0, \bar{v}_0$ only if there exist scalars α, β such that

$$(\bar{\bar{\mu}}^{-1} + \alpha\bar{\bar{\epsilon}}^{-1}) \overset{\times}{\times} \bar{u}_0\bar{u}_0 = 0, \quad (23)$$

$$(\bar{\bar{\mu}} + \beta\bar{\bar{\epsilon}}) \overset{\times}{\times} \bar{v}_0\bar{v}_0 = 0. \quad (24)$$

It can also be shown that there exist vectors $\bar{u}_0, \bar{v}_0$ only if α and β satisfy

$$\det(\bar{\bar{\mu}}^{-1} + \alpha\bar{\bar{\epsilon}}^{-1}) = 0, \quad (\bar{\bar{\mu}}^{-1} + \alpha\bar{\bar{\epsilon}}^{-1}) \overset{\times}{\times} (\bar{\bar{\mu}}^{-1} + \alpha\bar{\bar{\epsilon}}^{-1}) : \bar{\bar{I}} \leq 0, \quad (25)$$

$$\det(\bar{\bar{\mu}} + \beta\bar{\bar{\epsilon}}) = 0, \quad (\bar{\bar{\mu}} + \beta\bar{\bar{\epsilon}}) \overset{\times}{\times} (\bar{\bar{\mu}} + \beta\bar{\bar{\epsilon}}) : \bar{\bar{I}} \leq 0. \quad (26)$$

If $\alpha = \alpha_1$ and $\beta = \beta_1$ satisfy (25) and (26), the vectors $\bar{u}_0$ and $\bar{v}_0$ can be written out. In fact, because $\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1}$ is planar and symmetric, the dyadic $\bar{\bar{D}} = (\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1}) \overset{\times}{\times} (\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1})$ is linear and symmetric, and so there exists a vector $\bar{a}$ such that $\bar{\bar{D}} = -\bar{a}\bar{a}$ [the minus sign arising from the inequality (25)]. Now different cases can be separated:

(1) $\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1} = 0$ whence every vector $\bar{u}_0$ satisfies Eq. (23), i.e., the two wave normal surfaces coincide.

(2) $\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1} \neq 0$, $\bar{\bar{D}} = 0$. Then, $\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1}$ is linear, and there exists a vector $\bar{b}$ such that $\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1} = \pm\bar{b}\bar{b}$. Now there only exists one optical axis, namely, the one corresponding to the directions $\bar{u}_0 = \pm\bar{b}$ which both satisfy Eq. (23). The material is uniaxial.

(3) $\bar{\bar{D}} \neq 0$, whence the vector $\bar{a}$ defined above is nonzero. It can be shown that the four vectors $\bar{u}_0 = \pm[\bar{a} \times \bar{\bar{I}} \pm \sqrt{2}(\bar{\bar{\mu}}^{-1} + \alpha_1\bar{\bar{\epsilon}}^{-1})] \cdot \bar{d}$, where $\bar{d}$ is any vector which gives a unit vector $\bar{u}_0$, satisfy Eq. (23). The material, therefore, is biaxial.

ACKNOWLEDGMENT

The author is grateful to the anonymous referee for drawing the attention to Ref. 6, for showing the need of an appendix, and also for a discussion of the validity of the equations for dispersive media.

APPENDIX: DYADICS

A dyadic product $\bar{a}\bar{b}$ of two vectors $\bar{a}$ and $\bar{b}$ is defined by the associative law

$$(\bar{a}\bar{b}) \cdot \bar{c} = \bar{a}(\bar{b} \cdot \bar{c}) \quad \text{for all } \bar{c}. \quad (A1)$$

Dyadics $\bar{\bar{A}}, \bar{\bar{B}}$ are polynomials of dyadic products of vectors

$$\bar{\bar{A}} = \sum_{i=1}^{M} \bar{a}_i\bar{b}_i, \quad \bar{\bar{B}} = \sum_{j=1}^{N} \bar{c}_j\bar{d}_j. \quad (A2)$$

The following bilinear products between two dyadics are defined:

$$\bar{\bar{A}} \cdot \bar{\bar{B}} = \sum_{i,j} \bar{a}_i(\bar{b}_i \cdot \bar{c}_j)\bar{d}_j, \quad (A3)$$

$$\bar{\bar{A}} : \bar{\bar{B}} = \sum_{i,j} (\bar{a}_i \cdot \bar{c}_j)(\bar{b}_i \cdot \bar{d}_j), \quad (A4)$$

$$\bar{\bar{A}} \overset{\times}{\times} \bar{\bar{B}} = \sum_{i,j} (\bar{a}_i \times \bar{c}_j)(\bar{b}_i \times \bar{d}_j). \quad (A5)$$

If $\bar{e}_i$, $i = 1, 2, 3$ is a basis and $\bar{e}'_j = \epsilon_{ijk}[(\bar{e}_j \times \bar{e}_k)/(\bar{e}_i \cdot \bar{e}_j \times \bar{e}_k)]$ is the reciprocal basis, for every dyadic $\bar{\bar{A}}$ there exists a unique set of nine scalars α_{ij}, such that we can write

$$\bar{\bar{A}} = \sum_{i=1}^{3}\sum_{j=1}^{3} \alpha_{ij}\bar{e}_i\bar{e}'_j. \quad (A6)$$

Hence, there exists a correspondence between the dyadic $\bar{\bar{A}}$ and a matrix $[\alpha_{ij}]$ defined by the basis $\bar{e}_i$.

We denote by $\bar{\bar{A}}_T$ the transpose of $\bar{\bar{A}}$, i.e., the dyadic satisfying $\bar{\bar{A}}_T \cdot \bar{a} = \bar{a} \cdot \bar{\bar{A}}$ for all $\bar{a}$, and $\bar{\bar{I}}$ is the identity dyadic, which can be represented in a basis $\bar{e}_i$ as follows:

$$\bar{\bar{I}} = \sum_{i=1}^{3} \sum_{j=1}^{3} \delta_{ij} \bar{e}_i \bar{e}'_j = \sum_{i=1}^{3} \bar{e}_i \bar{e}'_i. \tag{A7}$$

Using vector algebra and definitions (A3)–(A5) we may prove the following identities:

$$\bar{\bar{A}} : \bar{\bar{B}} = \bar{\bar{B}} : \bar{\bar{A}} = \bar{\bar{A}}_T : \bar{\bar{B}}_T = (\bar{\bar{A}} \cdot \bar{\bar{B}}_T) : \bar{\bar{I}}; \tag{A8}$$

$$\bar{\bar{A}} \times\!\!\times \bar{\bar{B}} = \bar{\bar{B}} \times\!\!\times \bar{\bar{A}} = (\bar{\bar{A}}_T \times\!\!\times \bar{\bar{B}}_T)_T; \tag{A9}$$

$$\bar{\bar{A}} \times\!\!\times \bar{\bar{I}} = (\bar{\bar{A}} : \bar{\bar{I}})\bar{\bar{I}} - \bar{\bar{A}}_T; \tag{A10}$$

$$\bar{\bar{A}} \times\!\!\times (\bar{\bar{B}} \times\!\!\times \bar{\bar{C}}) = (\bar{\bar{A}} : \bar{\bar{C}})\bar{\bar{B}} + (\bar{\bar{A}} : \bar{\bar{B}})\bar{\bar{C}} - \bar{\bar{B}} \cdot \bar{\bar{A}}_T \cdot \bar{\bar{C}} - \bar{\bar{C}} \cdot \bar{\bar{A}}_T \cdot \bar{\bar{B}}; \tag{A11}$$

$(\bar{\bar{A}} \times\!\!\times \bar{\bar{B}}) : \bar{\bar{C}}$ is invariant in all permutations of $\bar{\bar{A}}, \bar{\bar{B}}, \bar{\bar{C}}$; (A12)

$$(\bar{\bar{A}} \times\!\!\times \bar{\bar{B}}) \cdot (\bar{\bar{C}} \times\!\!\times \bar{\bar{D}}) = (\bar{\bar{A}} \cdot \bar{\bar{C}}) \times\!\!\times (\bar{\bar{B}} \cdot \bar{\bar{D}}) + (\bar{\bar{A}} \cdot \bar{\bar{D}}) \times\!\!\times (\bar{\bar{B}} \cdot \bar{\bar{C}}); \tag{A13}$$

$$\bar{\bar{A}}^3 - (\bar{\bar{A}} : \bar{\bar{I}})\bar{\bar{A}}^2 + \tfrac{1}{2}(\bar{\bar{A}} \times\!\!\times \bar{\bar{A}} : \bar{\bar{I}})\bar{\bar{A}} - \tfrac{1}{6}(\bar{\bar{A}} \times\!\!\times \bar{\bar{A}} : \bar{\bar{A}})\bar{\bar{I}} = 0. \tag{A14}$$

For all bases $\bar{e}_i$ we have

$$\bar{\bar{A}} : \bar{\bar{I}} = \mathrm{tr}[\alpha_{ij}], \tag{A15}$$

$$\tfrac{1}{2}\bar{\bar{A}} \times\!\!\times \bar{\bar{A}} : \bar{\bar{I}} = \text{sum of principal minors of } [\alpha_{ij}], \tag{A16}$$

$$\tfrac{1}{6}\bar{\bar{A}} \times\!\!\times \bar{\bar{A}} : \bar{\bar{A}} = \det[\alpha_{ij}]. \tag{A17}$$

These functions of $\bar{\bar{A}}$ may be denoted by $\mathrm{tr}\bar{\bar{A}}$, $\mathrm{spm}\bar{\bar{A}}$ and $\det\bar{\bar{A}}$, respectively. The double cross product can be expressed using single and double dot products:

$$\begin{aligned}\bar{\bar{A}} \times\!\!\times \bar{\bar{B}} &= (\bar{\bar{A}} : \bar{\bar{I}})(\bar{\bar{B}} : \bar{\bar{I}})\bar{\bar{I}} - (\bar{\bar{A}} : \bar{\bar{B}}_T)\bar{\bar{I}} - (\bar{\bar{B}} : \bar{\bar{I}})\bar{\bar{A}}_T \\ &\quad - (\bar{\bar{A}} : \bar{\bar{I}})\bar{\bar{B}}_T + (\bar{\bar{A}} \cdot \bar{\bar{B}} + \bar{\bar{B}} \cdot \bar{\bar{A}})_T \\ &= [\mathrm{tr}\bar{\bar{A}}\ \mathrm{tr}\bar{\bar{B}} - \mathrm{tr}(\bar{\bar{A}} \cdot \bar{\bar{B}})]\bar{\bar{I}} - (\mathrm{tr}\bar{\bar{B}})\bar{\bar{A}}_T - (\mathrm{tr}\bar{\bar{A}})\bar{\bar{B}}_T \\ &\quad + (\bar{\bar{A}} \cdot \bar{\bar{B}} + \bar{\bar{B}} \cdot \bar{\bar{A}})_T.\end{aligned} \tag{A18}$$

A dyadic $\bar{\bar{A}}$ is complete if $\det\bar{\bar{A}} \neq 0$, otherwise it is planar. A dyadic $\bar{\bar{A}}$ is linear if $\bar{\bar{A}} \times\!\!\times \bar{\bar{A}} = 0$. For a planar dyadic $\bar{\bar{A}}$, there exists a vector $\bar{a}$ such that $\bar{\bar{A}} \cdot \bar{a} = 0$, and $\bar{\bar{A}}$ can be expressed as a sum of two dyads $\bar{\bar{A}} = \bar{b}\bar{c} + \bar{d}\bar{e}$. For a linear dyadic $\bar{\bar{A}}$ there exists a vector $\bar{a}$ such that $\bar{\bar{A}} \times \bar{a} = 0$, and $\bar{\bar{A}}$ can be expressed as a single dyad $\bar{\bar{A}} = \bar{b}\bar{a}$.

For a complete dyadic $\bar{\bar{A}}$ there exists a unique solution for the equation $\bar{\bar{A}} \cdot \bar{\bar{X}} = \bar{\bar{B}}$, namely $\bar{\bar{X}} = \bar{\bar{A}}^{-1} \cdot \bar{\bar{B}}$ with

$$\bar{\bar{A}}^{-1} = (3\bar{\bar{A}}_T \times\!\!\times \bar{\bar{A}}_T)/(\bar{\bar{A}} \times\!\!\times \bar{\bar{A}} : \bar{\bar{A}}). \tag{A19}$$

Also, for a complete dyadic $\bar{\bar{A}}$ there exists a unique solution for the equation $\bar{\bar{A}} \times\!\!\times \bar{\bar{X}} = \bar{\bar{B}}$, namely

$$\bar{\bar{X}} = \bar{\bar{A}}_T^{-1} \times\!\!\times \bar{\bar{B}} - [3(\bar{\bar{A}} : \bar{\bar{B}})/(\bar{\bar{A}} \times\!\!\times \bar{\bar{A}} : \bar{\bar{A}})]\bar{\bar{A}}. \tag{A20}$$

A dyadic is positive definite (PD) if $\bar{\bar{A}} : \bar{a}\bar{a} > 0$ for every vector $\bar{a} \neq 0$. For a PD dyadic $\bar{\bar{A}}$, the following properties can be proven to be valid:

(i) $\bar{\bar{A}}$ is complete;

(ii) $\bar{\bar{A}}^{-1}$ exists and is PD;

(iii) $\bar{\bar{A}} \times\!\!\times \bar{\bar{A}}$ is PD;

(iv) $\mathrm{tr}\bar{\bar{A}} > 0$, $\mathrm{spm}\bar{\bar{A}} > 0$, $\det\bar{\bar{A}} > 0$;

(v) $\frac{1}{2}(\bar{\bar{A}} + \bar{\bar{A}}_T)$ = symmetric part of $\bar{\bar{A}}$ is PD;

(vi) all eigenvalues of $\frac{1}{2}(\bar{\bar{A}} + \bar{\bar{A}}_T)$ are positive.

---

*Address for the academic year 1972-73: Electrical Engineering Department, University of Illinois, Urbana, Illinois 61801.

[1] J. W. Gibbs, E. B. Wilson, *Vector Analysis* (Dover, New York, 1960).

[2] M. Kline, I. W. Kay, *Electromagnetic Theory and Geometrical Optics* (Interscience, New York, 1965), Chap. III.

[3] See, e.g., A. Sommerfeld, *Elektrodynamik* (Verlag Akademie Geest & Portig, Leipzig, 1967), pp. 27-28.

[4] The differential operations contained in $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ give rise to dispersion; temporal if differentiation with respect to time is involved and spatial if spatial differentiations are involved. Here we follow the lines of Ref. 2 (p. 67, footnote[13]) and assume a mathematical medium with no temporal dispersion, because a physical medium with any polarization inertia would act as the vacuum to the propagating discontinuity. Also, we assume no spatial dispersion, because, although equations up to (17) would be valid, the following conclusions only apply for media with no spatial dispersion. Hence, $\bar{\bar{\epsilon}}$ and $\bar{\bar{\mu}}$ are algebraic quantities, and in Eqs. (7) and (8) the field vectors can be replaced by the discontinuities (starred vectors).

[5] H. Gelman, J. Math. Phys. (N.Y.) 11, 3053 (1970).

[6] T. W. Johnston, Radio Sci. 4, 729 (1969).

[7] I. V. Lindell, Rept. S41, Helsinki Univ. Tech. Radio Lab. (1971).

Perturbation theory for nonderivative nonpolynomial Lagrangians

J. G. Taylor†
Department of Physics, University of Southampton, Southampton, England
(Received 17 May 1971)

We present a regularization scheme for deriving finite, unitary, and causal amplitudes to each order in the major coupling constant for nonderivative nonpolynomial Lagrangians for the self-interaction of massless neutral scalar particles. The results are shown to contain no arbitrary constants, though these may be introduced consistently into the superpropagator, the second-order contribution.

1. INTRODUCTION

The suggestion has been gaining momentum[1] that the traditional ultraviolet divergences of quantum field theories with polynomial interactions can be removed by means of suitable nonpolynomial additions. These may naturally arise when general relativistic effects are included or nonlinear symmetry realizations are used. This leads to the attractive possibility of computation of self-mass and self-charge effects, as well as giving important modifications to high energy behavior. It may even lead to a completely new picture of elementary particles as "black holes," the microscopic analogs of the collapsars expected in great abundance in the heavens.

The main theoretical problem at the basis of such a programme is that of ensuring that the nonpolynomial Lagrangians being encountered do really give finite results for scattering amplitudes and contain no divergent quantities whatsoever. We will attempt to solve that problem in this and succeeding papers. This paper will be limited purely to Lagrangians which have nonderivative coupling in the interaction term. Up to the present the only method for calculation for nonpolynomial Lagrangians has been to expand S-matrix elements in powers of the interaction $L_{\rm int}$, with avoidance of any expansion of matrix elements of $L_{\rm int}$ in powers of any minor coupling constants. That will be the technique followed here, though evidently a discussion of the effect of summation over the different powers of $L_{\rm int}$ is necessary at a later stage. So we will discuss the Nth order S-matrix element in the expansion of

$$S = T\{e^{i\int L_{\rm int}(x)d^4x}\}, \tag{1}$$

in other words, the term

$$(i^N/N!)\int d^4x_1\cdots d^4x_N T\{L_{\rm int}(x_1)\cdots L_{\rm int}(x_N)\}. \tag{2}$$

Our problem is to show that a suitable prescription may be formulated to calculate each term (2) so as to give a finite S-matrix which is, in addition, unitary and causal to each order in N. We will finally be interested in the high energy behavior of the resulting amplitudes, in particular to determine if the localizability criterion given for the "superpropagator" with $N = 2$ extends naturally to all orders. Even for this lowest order case, $N = 2$, there is considerable ambiguity due effectively to the arbitrary high powers of Δ^n which arise on expansion of the superpropagator in certain minor coupling constants. While it is certainly possible to formulate perturbation theory in *both* major and minor coupling constants so that the finite amplitudes satisfy unitarity and causality to each order, the resulting expansion contains an infinity of arbitrary constants. This is unsatisfactory; we wish to show that it is possible to obtain finite, unitary, and causal amplitudes which contain *no* arbitrary constants beyond those introduced in the original Lagrangian. It is worthwhile to obtain such a result so as to avoid the enormous loss of predictive power which an infinity of arbitrary constants engenders.

We start our discussion in the next section by giving the general formulas for the regularized Nth-order term derived from (2). The problem of removal of the regularization is briefly discussed, and then solved completely for the case $N = 3$. We then show how this may be effected for each higher N in Sec. 4, and prove that the resulting amplitudes are unitary and causal, to each order in N, in succeeding sections. The relation to other regularization schemes is considered in Sec. 7, while the previous discussions are extended in Sec. 8 from the pure exponential Lagrangian for zero mass particles considered so far to a class of Lagrangians including nonlocalizable ones. We end the paper with a discussion of further problems raised by the work.

2. THE REGULARIZED NTH-ORDER TERM

We consider initially the simplest Lagrangian, the pure exponential one, for a massless neutral scalar field $\phi(x)$, so that

$$L_{\rm int} = G{:}e^{\lambda\phi}{:}, \tag{3}$$

G being the major and λ the minor coupling constants and the double dots : : denoting normal ordering. Then in (2) the Nth-order term is, after replacing the time-ordered product by the normal-ordered one,[2]

$$M^{(N)}(x_1\cdots x_N) = \frac{(iG)^N}{N!}\int dx_1\cdots dx_N {:}\prod_{i=1}^{N} e^{\lambda\phi(x_i)}{:}\prod_{i<j} e^{\lambda^2\Delta_{ij}}, \tag{4}$$

where $\Delta_{ij} = \Delta_\Gamma(x_i - x_j)$. In order to regularize $M^{(N)}$ of (4), we evidently have to start with the single superpropagator $e^{\lambda^2\Delta}$, the case for $N = 2$. We will see that a suitable regularization of this will allow a regularization of (4) to be given for all higher N. The regularization of $e^{\lambda^2\Delta}$ will be given by means of the Sommerfeld–Watson representation

$$e^{\lambda^2\Delta} = \sum_{n\geq 0}\frac{i}{n!}\lambda^{2n}\Delta^n = \frac{i}{2}\int_\Gamma \frac{\lambda^{2z}\Delta^z dz}{\Gamma(1+z)\tan\pi z}, \tag{5}$$

where Γ is a contour encircling the positive real axis.

The contour Γ is then opened out to be parallel to the imaginary axis and cross the real axis between 0 and -1; we denote this contour as C_0. We wish to obtain the regularized form for $M^{(N)}$ in momentum space, so displace C_0 to the right by one unit to C_1. Fourier transformation may then be performed to give

$$\int d^4x\, \exp[\lambda^2\Delta_F(x) + ikx] = \widetilde{(\exp(\lambda^2\Delta)|^{C_1} + 1)}$$
$$= (2\pi)^4\delta^4(k) + \frac{i}{2}\int_{C_1} dz\, \frac{\lambda^{2z}(-p^2)^{z-2}\pi(4\pi)^{2-2z}}{\tan(\pi z)\sin(\pi z)\Gamma(z-1)\Gamma(z)\Gamma(z+1)}. \tag{6}$$

We may use the fact that $(-p^2)^{z-2}$ has a pole at $z = 0$ with residue proportional[3] to $\delta(k)$ to rewrite (6) as

$$\widetilde{e^{\lambda^2\Delta}|^{C_0}} = \frac{i}{2}\int_{C_0} dz\, \frac{\lambda^{2z}(-p^2)^{z-2}\pi(4\pi)^{2-2z}}{\tan(\pi z)\sin(\pi z)\Gamma(z-1)\Gamma(z)\Gamma(z+1)}. \tag{7}$$

The representation (7) is still not satisfactory since the integral along C_0 does not converge; the final step of the regularization is to replace the factor $\sin\pi z$ in the denominator of (7) by $\sin(1+\delta)\pi z$, with $\delta > \frac{3}{2}$, so that convergence now occurs, both for spacelike or timelike values of p, to give

$$e^{\lambda^2\Delta}\Big|_\delta^{C_1} = \frac{i}{2}\int_{C_0} dz\, \frac{\lambda^{2z}(-p^2)^{z-2}\pi(4\pi)^{2-2z}}{\tan(\pi z)\sin[(1+\delta)\pi z]\Gamma(z-1)\Gamma(z)\Gamma(z+1)}. \tag{8}$$

We now wish to use this regularization to construct the Fourier transform of the regularized version of (4), that is, of

$$M_\delta^{(N)}(x_1\cdots x_N) = \frac{(iG)^N}{N!}\int dx_1\cdots dx_N = \prod_{i=1}^N e^{\lambda\phi(x_i)}$$
$$= \prod_{i<j} e^{\lambda^2\Delta_{ij}}\Big|_\delta^{C_0}. \tag{9}$$

We first exhibit the possible four-dimensional delta functions which arise in the Fourier transform of the c-number amplitude in (9). To achieve this, we use the representation

$$\prod_{i<j} e^{\lambda^2\Delta_{ij}}\Big|_\delta^{C_0} = \sum_{I,J\subset[1,N]}{}' \prod_{\substack{i\in I, j\in J\\ i<j}}{}' e^{\lambda^2\Delta_{ij}}\Big|_\delta^{C_1} (1+\delta)^{\lambda'-N(N-1)/2},$$

where the product Π' involves the use of each $i \in I$ and each $j \in J$ at least once, the summation $\sum'$ is over all such products, and λ' is the number of factors in the Π'. Then we may write

$$S^{(N)}(p_1\cdots p_N) = \sum_{I,J\subset[1,N]}{}' \prod_{\substack{i\in I, j\in J\\ i<j}}{}' \widetilde{\int d^4k_{ij}\, e^{\lambda^2\Delta_{ij}}\Big|_\delta^{C_1}}(k_{ij})$$
$$\times \prod_{j\in I\cup J}\delta^4\Big(p_j + \sum_{i\in I} k_{ji}\Big)\prod_{k\notin I\cup J}\delta^4(p_k)(1+\delta)^{\lambda'-N(N-1)/2} \tag{10}$$

We discuss each term separately in (10), phrasing our arguments in such a fashion that they apply to all of the terms there and are independent of the particular term chosen. We drop the superfluous $\delta^{(4)}$ functions in (10) and also all superpropagator factors in (10) for which the internal momenta k_{ij} are fixed by the external momenta. This is because the problem associated with discussing them is straightforward (involving a product of generalized functions depending on different variables). We thus take the superpropagators in (10) in which the momentum k_{ij} depends on one or more loop momenta which have to be integrated over.

We wish to evaluate the Fourier transform

$$S_\delta^{(N)}(p_1\cdots p_N) = \prod_{l=1}^L \int d^4k_l\, \widetilde{e^{\lambda^2\Delta}\Big|_\delta^{C_1}}(k_l)\prod_{j=1}^N \delta^4\Big(p_j + \sum_{l=1}^L k_l e_{lj}\Big), \tag{11}$$

where $e_{lj} = \pm 1$ if the line l is incident at the vertex j and the momentum is directed towards (away from) that vertex and is zero otherwise. The index runs over the L lines of an Nth order supergraph with $L \leq \frac{1}{2}N(N-1)$ [the normal ordering in (3) cutting out loops at a single vertex]. When $L = \frac{1}{2}N(N-1)$, then we have the complete graph with N vertices.

In order to evaluate the loop momentum integrations in (11), we reduce these integrations to that of standard perturbation theory by the representation

$$(-p^2)^{z-2} = -e^{-i\pi z/2}[\Gamma(2-z)]^{-1}\int_0^\infty d\alpha\alpha^{1-z}e^{i\alpha(p^2+i\epsilon)} \tag{12}$$

so that

$$S^{(N)}(p_1\cdots p_N) = \prod_{l=1}^L\left(\frac{i}{2}\int_{C_1} dz_l\, \frac{\lambda^{2z_l}(4\pi)^{2-2z_l}\cos(\pi z_l)e^{-i\pi z_l/2}}{\sin(1+\delta)\pi z_l\Gamma(z_l)\Gamma(1+z_l)}\int_0^\infty d\alpha_l\,\alpha_l^{1-z_l}\right)$$
$$\times \prod_{l=1}^L\int d^4k_l\, e^{i\alpha_l(p^2+i\epsilon)}\prod_{j=1}^N\delta^4\Big(p_j + \sum_{l=1}^L k_l e_{lj}\Big). \tag{13}$$

The last factor in (13) is the usual Feynman amplitude in momentum space for the appropriate Nth-order graph, before the Feynman parameters have been integrated over, and has the value

$$\delta^4\Big(\sum_{j=1}^N p_j\Big)(\det\theta)^{-2}\exp\left[i\left(\frac{\det\chi}{\det\theta} + i\epsilon\sum_{l=1}^L\alpha_l\right)\right],$$

where

$$\theta_{i,j} = \sum_{l=1}^L \alpha_l a_{li}a_{lj},$$
$$\chi_{i,j} = \theta_{ij} \quad (i,j \leq N),$$
$$\chi_{i,N+1} = \sum_{l=1}^L\sum_{j=1}^N \alpha_l a_{li}b_{lj}p_j = \chi_{N+1,i}, \tag{14}$$
$$\chi_{N+1,N+1} = \sum_{l=1}^L\sum_{j,k=1}^N \alpha_l b_{lj}b_{lk}p_j p_k,$$

and in (14) there are N independent loop momenta $q_1,\ldots,q_N$ chosen in some specified manner, so that

$$k_l = \sum_{i=1}^N a_{li}q_i + \sum_{j=1}^N b_{lj}p_j \tag{15}$$

the a_{li}, b_{lj} having values ± 1 or 0. We thus have to evaluate

$$S_\delta^{(N)}(p_1 \cdots p_N)$$

$$= \prod_{l=1}^{L} \left(\frac{i}{2} \int_{C_1} \frac{dz_l \lambda^{2z_l} (4\pi)^{2-2z_l} \cos(\pi z_l) e^{-i\pi z_l/2}}{\sin[(1+\delta)\pi z_l]\Gamma(z_l)\Gamma(1+z_l)} \int_0^\infty d\alpha_l \alpha_l^{1-z_l} \right)$$

$$\times \delta^4\left(\sum_{j=1}^{N} p_j\right) (\det\theta)^{-2} \exp\left[i\left(\frac{\det\chi}{\det\theta} + i\epsilon \sum_{l=1}^{L} \alpha_l \right) \right]. \tag{16}$$

In order to discuss the convergence of the integrals over α_l and z_l we choose the variables $t_1, \ldots, t_L$ of Speer,[4] defined for the region

$$0 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_L \tag{17}$$

by

$$\alpha_l = \prod_{l'=l}^{L} t_{l'}. \tag{18}$$

Then[4]

$$\det\theta = \prod_{l=1}^{L} t_l^{N_l} E(t_1 \ldots t_{L-1}),$$

$$\det\chi = t_L \prod_{l=1}^{L} t_l^{N_l} F(t_1 \ldots t_{L-1}), \tag{19}$$

where E and F are polynomials in $t_1, \ldots, t_{L-1}$, E does not vanish in the region of integration, and N_l is the number of loops of the subgraph consisting of lines 1 to l. The region of t_l integration now becomes

$$0 \leq t_L \leq \infty, \quad 0 \leq t_l \leq 1 \quad (l = 0, 1, \ldots, L-1). \tag{20}$$

To deal with the whole region of integration over the α variables in (16), we divide it into $L!$ similar terms; if

$$0 \leq \alpha_{\pi^{-1}(1)} \leq \cdots \leq \alpha_{\pi^{-1}(L)},$$

where π is a permutation of $(1, \ldots, L)$, then the transformed variables will be

$$\alpha_l = \prod_{l'=\pi(l)}^{L} t_{l'} \tag{21}$$

with the same region of integration (20) in the t_l variables. Thus the α_l integration now becomes

$$\sum_{\pi \in P_L} \int_0^\infty dt_L t_L^{\gamma_L} \prod_{l=1}^{l-1} \int_0^1 dt_l t_l^{\gamma_{\pi(l)}} E_\pi^{-2}$$

$$\times \exp\left\{ it_L \left[\frac{F_\pi}{E_\pi} + i\epsilon \left(1 + \sum_{l=1}^{L-1} \prod_{l'=\pi(l)}^{L-1} t_{l'} \right) \right] \right\} \tag{22}$$

with

$$\gamma_l = 2(l - N_l) - 1 - \sum_{l'=1}^{l} z_{l'},$$

where P_L is the permutation group on L integers and E_π, F_π are the polynomials constructed from $\det\theta$ and $\det\chi$ but now by means of the change of variables (21) instead of (18). The integration over t_L may be performed exactly to give

$$\sum_{\pi \in P_L} \prod_{l=1}^{L-1} \int_0^1 dt_l t_l^{\gamma_{\pi(l)}} F_\pi^{-1-\gamma_L} E_\pi^{\gamma_L - 1} \Gamma(1+\gamma_L) e^{i\pi(1+\gamma_L)/2}, \tag{23}$$

where

$$1 + \gamma_L = 2(N-1) - \sum_{l=1}^{L} z_l.$$

Thus we may write (16), dropping the overall energy-momentum conserving δ function as

$$S_\delta^{(N)}(p_1 \cdots p_N)$$

$$= \pi \prod_{l=1}^{L} \left[\frac{i}{2} \int_{C_1} \frac{dz_l \lambda^{2z_l} (4\pi)^{2-2z_l} \cos\pi z_l}{\sin[(1+\delta)\pi z_l]\Gamma(z_l)\Gamma(1+z_l)} \right]$$

$$\times \frac{e^{iNt}}{\sin(\pi \sum_{l=1}^{L} z_l)\Gamma(\sum_{l=1}^{L} z_l - 2N + 3)}$$

$$\times \sum_{\pi \in P_L} \prod_{l=1}^{L-1} \int_0^1 dt_l t_l^{\gamma_{\pi(l)}} (-F_\pi)^{-2(N-1)+\sum_{l=1}^{L} z_l}$$

$$\times (E_\pi)^{2N - \sum_{l=1}^{L} z_l}. \tag{24}$$

We can now discuss the convergence of the various integrals in this expression. Firstly the t_l integrals are all convergent at $t_l = 0$ since each z_l on C_1 has $0 < \mathrm{Re} z_l < 1$. Since F_π may vanish in the region of t integration, it is necessary to take explicit account of the $i\epsilon$ term in the exponent of (22), so replacing the quantity F_π/E_π in (24) by

$$E_\pi^{-2}[F_\pi/E_\pi + i\epsilon(1 + t_{L-1} + \cdots)]^{\sum_1^L z_l - 2(N-1)}.$$

This quantity is well behaved in the invariants $(p_i \cdot p_j)$ and the variables t_i and has no poles in $\sum_1^L z_l$ if $\epsilon > 0$; it will have exponential growth of order π as $\sum_1^L z_l$ goes to infinity parallel to the imaginary axis. We will keep $\epsilon > 0$ throughout our calculations, as is usual; we will give a careful discussion of the existence of the boundary values as $\epsilon \to 0$ at the end of Sec. 4, though we immediately expect the boundary value to be a generalized function of a standard kind in the invariants $(p_i p_j)$. Finally the z_l integrations will each converge for arbitrary external momenta if $\delta > \frac{3}{2}$, taking account of the last factor in (24). Thus the regularization for the single superpropagator succeeds for all the higher orders.

In order to remove the regularization it is necessary to bend the contours C_1 back to the positive real axis. To achieve this, it will be necessary to make explicit the singularities of the integrand of (24) in the z_l variables. In other words the t_i integrals will have to be performed explicitly to remove the obvious poles at $\gamma_{\pi(l)}$ taking negative integer values. To achieve this, we can expand the function

$$\mathfrak{F}(t_1 \cdots t_{L-1})$$

$$= E_\pi^{-2}[F_\pi/E_\pi + i\epsilon(1 + t_{L-1} + \cdots)]^{\sum_1^L z_l - 2(N-1)} \tag{25}$$

in a suitable set of bases functions of $t_1 \cdots t_{L-1}$. We take for these products of functions of single variables. If these functions are $\{\phi_n(t)\}$, not necessarily assumed orthonormal on the unit interval $(0, 1)$, then

$$\mathfrak{F}(t_1 \cdots t_{L-1}) = \sum_{\mathbf{m}} a\left(\mathbf{m}, \sum_1^L z_l\right) \phi_{\mathbf{m}}(t), \tag{26}$$

where we use the notation $\mathbf{m} = (m_1, \ldots, m_{L-1})$, $\phi_{\mathbf{m}}(t) =$

$\prod_{l=1}^{L-1}\phi_{m_l}(t_l)$. Then the t_l integrations on the right of (24) become

$$\sum_{\pi\in P_L}\prod_{l=1}^{L-1}\int_0^1 dt_l\, t_l^{\gamma_{\pi(l)}}\mathfrak{F}\left(\mathbf{t},\sum_1^L z_l\right)=\sum_{\mathbf{m}} a\left(\mathbf{m},\sum_1^L z_l\right)\psi(\mathbf{m},\boldsymbol{\gamma}_\pi), \tag{27}$$

where

$$\psi(\mathbf{m},\boldsymbol{\gamma}_\pi)=\prod_{l=1}^{L-1}\int_0^1 dt_l\, t_l^{\gamma_{\pi(l)}}\phi_{m_l}(t_l)=\prod_{l=1}^{L-1}\psi_{m_l}(\gamma_{\pi(l)}). \tag{28}$$

It is necessary to investigate the singularity of $\psi_m(\gamma)$ in its variable γ. Let us consider two simple cases for the choice of functions $\{\phi_m\}$:

(i) $\phi_m(x)=x^m$, $\quad\psi_m(y)=(y+m+1)^{-1}$

so that ψ_m has a simple pole in y at $y=-m-1$.

(ii) $\phi_m(x)=(1-x)^m$, $\quad\psi_m(y)+B(m+1,y+1)$

so that ψ_m has simple poles at $y=-1,-2,\ldots,-m-1$.

In both of these choices the singularities of the integrand remaining in (23), after the t_l integrations have been performed, have been explicitly displayed. This should then allow the contours C_1 to be bent back to the positive real ones. However, it is necessary to prove convergence of the resulting summation over $\mathbf{m}$. In the case (i) above, the coefficients $a(\mathbf{m},\sum_1^L z_l)$ are given by

$$a\left(\mathbf{m},\sum_1^L z_l\right)=\mathbf{D}^{\mathbf{m}}\mathfrak{F}(\mathbf{t})\,|_{\mathbf{t}=\mathbf{0}} \tag{29}$$

while for case (ii)

$$a\left(\mathbf{m},\sum_1^L z_l\right)=\mathbf{D}^{\mathbf{m}}\mathfrak{F}(t)|_{\mathbf{t}=\mathbf{1}}, \tag{30}$$

where $\mathbf{D}^{\mathbf{m}}=\prod_{l=1}^{L-1}/\partial^{m_l}/\partial t_l^{m_l}$. Thus, in order to have convergence of the resulting expressions, it will be necessary that the Taylor expansions about $\mathbf{0}$ or $\mathbf{1}$ in the cases (i) and (ii), respectively, converge in some polydisk about the relevant point. And since the integration over the t_l variables is from 0 to 1, we expect that there will only be convergence of the resulting summation over $\mathbf{m}$ if the Taylor expansions converge in polydisks about $\mathbf{0}$ or $\mathbf{1}$ with at least radii 1, so in the regions

(i) $\Lambda_0=\{\mathbf{t}:|t_l|<r_l,\ r_l>1,\ 1\leq l\leq L-1\}$,

(ii) $\Lambda_1=\{\mathbf{t}:|t_l-1|<r_l^1,\ r_l^1>1,\ 1\leq l\leq L-1\}$.

We will investigate in the next section if this is possible for the case $N=3$, when the contours C_1 are modified to Γ.

3. FINITENESS FOR $N=3$

In the case $N=3$, we have the explicit expressions and values

$$L=N=3,\ E=1+t_2+t_1t_2,\ F=p_3^2t_2^2t_1+p_2^2t_2t_1+p_1^2t_2,$$

$\delta_1=1-z_1$, $\gamma_2=3-z_1-z_2$, $\gamma_3=3-(z_1+z_2+z_3)$, so giving the following, from (23), via the choice (i) of the previous section for the functions $\{\phi_m\}$:

$$\begin{aligned}M_\delta^{(3)}(p_1p_2p_3)&=-\pi\prod_{l=1}^{3}\left(\frac{i}{2}\int_{C_1}\frac{dz_l\,\lambda^{2z_l}(4\pi)^{2-2z_l}\cos\pi z_l}{\sin[(1+\delta)\pi z_l]\Gamma(z_l)\Gamma(1+z_l)}\right)\\&\times\frac{1}{\sin(\pi\sum_1^3 z_l)\Gamma(\sum_1^3 z_l-3)}\\&\times\sum_{\pi\in P_3}\sum_{\mathbf{m}}\prod_{l=1}^{2}\frac{1}{(\gamma_{\pi(l)}+m_l+1)}\\&\times\mathbf{D}^{\mathbf{m}}\{(-F_\pi)^{\sum_1^3 z_l-4}(E_\pi)^{6-\sum_1^3 z_l}\}_{\mathbf{t}=\mathbf{0}}.\end{aligned} \tag{31}$$

On closing the contours C_1 back to Γ we see that the summation over $\mathbf{m}$ can be performed by replacing, for example, $(\gamma_{\pi(l)}+m_l+1)^{-1}$ by $\int_0^\infty e^{-i(\gamma_{\pi(l)}+m_l+1)\alpha}d\alpha$ for $\mathrm{Im}\gamma_{\pi(l)}>0$; this corresponds, after summation over $\mathbf{m}$ in (31), to evaluation of $(-F_\pi)^{-4+\sum_1^3 z_l}(E_\pi)^{6-\sum_1^3 z_l}$ at values $t_l=e^{-i\gamma_{\pi(l)}\alpha}$, so to points on the unit circle. However, E_π can then vanish at such points, for example, at the point

$$t_1=t_2=-\tfrac{1}{2}+i\sqrt{3}/2$$

So we do not expect the summation over $\mathbf{m}$ in (31) to converge after the contours C_1 have been distorted back to the real axis, when the values of $\mathrm{Re}z_l$ can become arbitrarily large and positive.

It is possible to discuss this case by means of the expansion about $\mathbf{t}=\mathbf{1}$, case (ii) mentioned in the previous section. For E_π is nonzero in the region Λ_1, as can be seen by direct inspection. Taking

$$t_i=1+z_i\quad(i=1,2),$$

then

$$E=3+2z_2+z_1+z_1z_2.$$

Thus E has a zero at

$$z_1=-(3+2z_2)/(1+z_2)$$

and for $|z_1|\leq|$ it is necessary that

$$f(r,\theta)=8+3r^2+10r\cos\theta\leq 1$$

where $z=re^{i\theta}$. But for $\cos\theta>0$, the minimum value of $f(r,\theta)$ is 8, while for $\cos\theta<0$ the minimum value of f is for $r=1$ if $|\cos\theta|>\frac{6}{10}$, with value $(11+10\cos\theta)$, and for $r=-(10\cos\theta)/6$ if $|\cos\theta|<\frac{6}{10}$, with value $(8-\frac{100}{12}\cos^2\theta)$. In all of these cases f is positive, so that E does not take the value zero inside Λ_1.

For the choice (ii) of the previous section, we have

$$M_\delta^{(3)}(p_1p_2p_3)=-\pi\prod_{l=1}^{3}\left(\frac{i}{2}\int_{C_1}dz_l\frac{\lambda^{2z_l}(4\pi)^{2-2z_l}\cos\pi z_l}{\sin(1+\delta)\pi z_l\Gamma(z_l)\Gamma(1+z_l)}\right)\times\frac{1}{\sin\pi\sum_1^3 z_l\Gamma(\sum_1^3 z_l-3)}$$

$$\times\sum_{\pi\in P_3}\sum_{\mathbf{m}}\prod_{l=1}^{2}\frac{\Gamma(m_l+1)\Gamma(1+\gamma_{\pi(l)})}{\Gamma(2+m_l+\gamma_{\pi(l)})}\,\mathbf{D}^{\mathbf{m}}\left\{\frac{(-p_{\pi(3)}^2t_2^2t_1-p_{\pi(2)}^2t_2t_1-p_{\pi(1)}^2t_2)^{-4+\sum_1^3 z_l}}{(1+b_2+t_1b_2)^{\sum_1^3 z_l-6}}\right\}_{\mathbf{t}=\mathbf{1}}. \tag{32}$$

We may close the z_3 contour round the real axis, picking up residues from the poles at $z_3 = n_3/(1+\delta)$, $n_3 = 1, 2, \cdots$, or at $z_3 = n_3' - (z_1 + z_2), n_3' = 4, 5, \cdots$, though not from those in the term in curly brackets due to the unwritten $i\epsilon$ term, to give

$$M_\delta^{(3)}(p_1p_2p_3) = \frac{1}{4\pi(1+\delta)} \prod_{l=1}^{2}\left(\frac{i}{2}\int_{C_1} dz_l \frac{\lambda^{2z_l}(4\pi)^{2-2z_l}\cos\pi z_l}{\sin(1+\delta)\pi z_l\Gamma(z_l)\Gamma(1+z_l)}\right) \times \sum_{\pi\in P_3}\sum_{\mathbf{m}} \prod_{l=1}^{2}\frac{\Gamma(m_l+1)\Gamma(1+\gamma_{\pi(l)})}{\Gamma(2+m_l+\gamma_{\pi(l)})}$$
$$\times\Bigg\{\sum_{n_3}\frac{\cos\pi n_3(1+\delta)a_\pi(\mathbf{m}, z_1+z_2+n_3/(1+\delta))\lambda^{2n_3}(4\pi)^{2-2n_3}}{\Gamma(n_3/(1+\delta))\Gamma(1+n_3/(1+\delta))\sin\{\pi[z_1+z_2+n_3/(1+\delta)]\}\Gamma(z_1+z_2+n_3/(1+\delta)-3)}$$
$$+\sum_{n_3'}\frac{(-1)^{1+n_3'}\cos\pi(z_1+z_2)\lambda^{2(n_3'-z_1-z_2)}(4\pi)^{2-2(z_1+z_2)}a_\pi(\mathbf{m},n_3')}{\sin[(1+\delta)\pi(n_3'-z_1-z_2)]\Gamma(n_3'-z_1-z_2)\Gamma(1+n_3'-z_1-z_2)\Gamma(n_3'-3)}\Bigg\}. \tag{33}$$

We may now let $\delta \to 0$ in the last bracket of (33), in other words introducing a different δ for each z variable. We then close the z_2 contour around the positive real axis, picking up the residues from the poles at $z_2 = n_2/(1+\delta_2)$, $n_2 = 1, 2, \cdots$, or at $z_2 = n_2' - z_1$. This pole arises both from the factor $\sin\pi(z_1+z_2)$ in the denominator of the last bracket of (33) as well as in the factor $\Gamma(4-z_1-z_2)$, though only for $5 + m_l > n_2'$, when it is thus a double pole. The total contribution in (33) is thus (at the same time letting $\delta_2 \to 0$ for simplicity):

$$M_\delta^{(3)}(p_1p_2p_3) = \frac{i}{2}\int_{C_1} dz_1 \frac{\lambda^{2z_1}(4\pi)^{2-2z_1}\cos\pi z_1}{\sin(1+\delta_1)\pi z_1\Gamma(z_1)\Gamma(1+z_1)}\sum_{\mathbf{m}}\Bigg\{\sum_{n_2}\frac{\lambda^{2n_2}(4\pi)^{2-2n_2}\Gamma(1+m_2)\Gamma(4-z_1-n_2)\Gamma(1+m_1)\Gamma(2-z_1)}{\Gamma(n_2)\Gamma(1+n_2)\Gamma(\delta+m_2-z_1-n_2)\Gamma(3+m_1-z_1)}$$
$$\times\Bigg[\sum_{n_3}\frac{\lambda^{2n_3}(4\pi)^{2-2n_3}(-1)^{n_2}a(\mathbf{m}, z_1+n_2+n_3)}{\Gamma(n_3)\Gamma(1+n_3)\Gamma(z_1+n_2+n_3-3)\sin\pi z_1}$$
$$+\sum_{n_3'}\frac{(-1)^{n_2}\cos\pi z_1\lambda^{2(n_3'-n_2-z_1)}(4\pi)^{2-2z_1-2n_2}a(\mathbf{m},n_3')}{\sin\pi z_1\Gamma(n_3'-z_1-n_2)\Gamma(1+n_3'-z_1-n_2)\Gamma(n_3'-3)}\Bigg]$$
$$+\sum_{n_2'}\frac{\partial}{\partial z_2}\frac{\Gamma(1+m_2)\Gamma(1+m_1)\Gamma(z-z_1)\lambda^{2z_2}(4\pi)^{2-2z_2}}{\Gamma(\delta+m_2-z_1-z_2)\Gamma(3+m_1-z_1)\Gamma(z_1+z_2-3)\Gamma(z_2)\Gamma(1+z_2)}$$
$$\times\Bigg[\sum_{n_3}\frac{(-1)^{n_2'}\lambda^{2n_3}(4\pi)^{2-2n_3}a(\mathbf{m},z_1+z_2+n_3)}{\Gamma(n_3)\Gamma(1+n_3)\Gamma(z_1+z_2+n_3-3)}+\sum_{n_3'}\frac{\cos\pi(z_1+z_2)\lambda^{2(n_3'-z_1-z_2)}(4\pi)^{2-2z_1-2z_2}a(\mathbf{m},n_3')}{\Gamma(n_3'-z_1-z_2)\Gamma(1+n_3'-z_1-z_2)\Gamma(n_3'-3)}\Bigg]_{z_2=z_2'-z_1}\Bigg\}. \tag{34}$$

We may now bend the contour C_1 back to Γ without having to calculate the explicit residues at the poles $z_1 = n_1$. We can determine the convergence of the right hand side of (34), with C_1 replaced by Γ, by straightforward analysis.

We commence with bounds for $a(\mathbf{m}, n)$:

$$|a(\mathbf{m},n)| \leq m_1!m_2!\left|\left(\frac{1}{2\pi i}\right)^2\int_{\Gamma_1}dt_1\int_{\Gamma_2}dt_2 \times\frac{|-p_3^2t_2^2-p_2^2t_2t_1-p_1^2t_2|^{n-4}}{(1+t_2+t_1t_2)^{n-6}(t_1-1)^{m_1+1}(t_2-1)^{m_2+1}}\right|, \tag{35}$$

where Γ_1 and Γ_2 are the contours $|t-1| = r > 1$, $n \geq 4$, and r is chosen small enough to have no zeros of E inside or on $\Gamma_1 \times \Gamma_2$. We may estimate the right-hand side of (35) by

$$|a(\mathbf{m},n)| \leq m_1!\, m_2!\,|f(p_i^2)|^{n-4}e^{b-n}r^{-m_1-m_2-2},$$

where

$$f = \sup_{t_i\in\Gamma_i}|p_3^2t_2^2 + p_2^2t_2t_1 + p_1^2t_2|,$$
$$e = \inf_{t_i\in\Gamma_i}|1+b_2+t_1b_2|, \tag{36}$$

and from the earlier discussion we know that $e > 0$. We let the final contour Γ close on the positive integers and $\delta_1 \to 0$. Then we see that the various terms on the right-hand side of (34) are bounded, to within logarithmically increasing terms, by

$$(\text{const}) \times \sum_{n_1,n_2,n_3}\prod_{i=1}^{3}\frac{1}{\Gamma(n_i)\Gamma(1+n_i)}\frac{|f(p_i^2)|^{\Sigma n_i-4}}{e^{\Sigma n_i-6}}\times\frac{1}{\Gamma(n_1+n_2+n_3)}\times\sum_{\mathbf{m}}r^{-m_1-m_2-2}$$
$$\times\frac{\Gamma(m_1+1)\Gamma(m_2+1)}{\Gamma(\delta+m_2-n_1-n_2)\Gamma(n_1+n_2-3)\Gamma(3+m_1-n_1)\Gamma(n_1-1)}. \tag{37}$$

The various terms in (34) can be reduced to those of (37) by suitable substitution of the summation variables, such as $n_3^1 = n_3 + n_1 + n_2$, etc. We note that the last summation on the right-hand side of (37), that over m_1 and m_2, is equal to

$$(-1)^{2n_1+n_2-6}[\Gamma(n_1-1)\Gamma(n_1+n_2-3)]^{-1}$$
$$\times [(d/dr)^{n_1-2}(r-1)^{-1}]$$
$$\times [(d/dr)^{n_1+n_2-4}(r-1)^{-1}] = (r-1)^{4-2n_1-n_2}$$

and so is finite if $r > 1$ for each value of n_1 and n_2. Thus

$$|M_0^{(3)}(p_1p_2p_3)|$$
$$\leq (\text{const}) \sum_{\mathbf{n}} \frac{|f(p_i^2)^{\sum n_i - 4}}{e^{\sum n_i 6}\Gamma(\sum n_i)} \prod_{i=1}^{3} \frac{1}{\Gamma(n_i)\Gamma(1+n_i)}$$
$$\times (r-1)^{4-2n_1-n_2}. \tag{38}$$

This is evidently finite, and we can even read off the high energy behavior of (38): If any subset of the $p_i^2 \to +\infty$, then

$$|f(p_i^2)| \leq \text{const} \max_i p_i^2$$

so that

$$|M_0^{(3)}(p_1p_2p_3)| \leq \text{const}\, e^{\text{const}'(\max_i p_i^2)^{1/3}}, \tag{39}$$

which is precisely the high energy behavior for the case $N = 2$. There is thus no change of the degree of localizability on going from $N = 2$ to $N = 3$.

If we try to extend this approach to the case $N = 4$, we find that the problem of convergence of the summation over $\mathbf{m}$ becomes extremely difficult. Thus in this case

$$E = [1 + t_3 + t_3t_2 + t_3t_2t_1 + t_4t_3t_2 + t_5t_4t_3$$
$$+ t_4t_3t_2t_1 + t_5t_4t_3t_2 + t_4t_3^2t_2 + t_4t_3^2t_2b_1$$
$$+ t_5t_4t_3^2t_2t_1 + t_5t_4t_3^2t_2^2t_1 + t_5t_4^2t_3^2t_2t_1$$
$$+ t_5t_4^1t_3^2t_2 + t_5t_4^2t_3^2t_2^2t_1 + t_5t_4^2t_3^3t_2^2t_1]. \tag{40}$$

It is a very hard problem to prove that E is not zero inside the polydisk Λ_1 of the previous section for suitable $r_l^1 > 1$. This problem appears to have been very little considered in the mathematical literature, unlike the case of the zeros of monomials. The problem is complicated by the fact that the graphs being considered are the complete graphs, in the language of graph theory. These are the most nonplanar graphs of any order which can be written down, and little is known of their algebraic analysis. In order to develop a useful convergence scheme we use a different class of functions $\{\phi_m\}$ for our expansion in the next section.

4. FINITENESS FOR HIGHER N

There is one obvious choice for the class of expansion functions $\{\phi_m\}$, and that is the Legendre polynomials. For the function E will be nonzero in a suitably small ellipse around the interval (0, 1) in each of the t_i, and so we take

$$\phi_m(t) = P_m(2t-1).$$

Then in (27)

$$a_\pi(\mathbf{m},\gamma) = \prod_{l=1}^{L-1}\left(\frac{2n_l+1}{2}\right)\int_0^1 dt_l P_{m_l}$$
$$\times (2t_l - 1)E^{-2}(E_\pi/E_\pi + i\epsilon)^{\gamma-2(N-1)}. \tag{41}$$

We can put a useful bound on $a(\mathbf{m},\gamma)$ as follows. For a function f on the interval $(-1, +1)$ which is differentiable to 1st order, we can obtain

$$\left|\int_{-1}^{+1} P_n(t)f(t)dt\right| = \frac{1}{(2n+1)}\left|\int_{-1}^{+1} dtf(t)(P'_{n+1} - P'_{n-1})\right|$$
$$= \frac{1}{(2n+1)}\left|\int_{-1}^{+1} dtf'(P_{n+1} - P_{n-1})\right|$$
$$\leq \frac{1}{(2n+1)} - \frac{1}{\sqrt{n}} - \|f'\|, \tag{42}$$

where $\|f\|^2 = \int_{-1}^{+1} |f|^2\,dt$ and $\|P_n\| \leq 1/\sqrt{n}$. Repeating the steps leading to (42) a total of r times gives the bound

$$(2n+1)\left|\int_{-1}^{+1} P_n f dt\right| \leq \|f^{(r)}\|\Gamma(n-r)/\Gamma(n) \quad (r<n). \tag{43}$$

Extending this to $(L-1)$ variables, we have the bound

$$|a_\pi(\mathbf{m},\gamma)| \leq \prod_{l=1}^{L-1}\frac{\Gamma(m_l - r_l)}{\Gamma(m_l)}$$
$$\times \|\mathbf{D}^{\mathbf{r}}[E_\pi^{-2}(F_\pi/E_\pi + i\epsilon)^{\gamma-2(N-1)}]\| \quad (r_l < m_l), \tag{44}$$

where now the norm on the right hand side of (44) is the natural extension of the L_2 norm from functions of one to $(L-1)$ variables. We may bound this norm by means of Cauchy's multiple integral formula:

$$\mathbf{D}^{\mathbf{r}} \cdot E_\pi^{-2}\left(\frac{F_\pi}{E_\pi} + i\epsilon(1 + t_{L-1} + \cdots)\right)^{\gamma-2(N-1)}$$
$$= \prod_{l=1}^{L-1}\frac{r_l!}{2\pi i}\int \frac{du_l\{E^{-2}[F_\pi/E_\pi + i\epsilon]^{\gamma-2(N-1)}\}_{(u)}}{C(t_l - u_l)^{r_l+1}}, \tag{45}$$

where C is chosen as the small ellipse round $(0, 1)$ inside and on which E_π is nonzero; C may evidently be chosen to be independent of π. If the distance of this ellipse from $(0, 1)$ is d, while upper and lower bounds of E_π and E_π on C are e and $f(p_ip_j)$, respectively, again chosen as independent of π, then (45) allows us to deduce for $\gamma > 2N$ that

$$\|\mathbf{D}^{\mathbf{r}}[E_\pi^{2N-\gamma}(-F_\pi)^{\gamma-2(N-1)}]\|$$
$$\leq \mu f^{\gamma-2(N-1)}e^{2N-\gamma}\prod_{l=1}^{L-1} r_l!\, d^{-r_l-1},$$

where μ is a suitable constant independent of $\mathbf{r}$. Thus in (44), for $r_l < m_l$,

$$|a_\pi(\mathbf{m},\gamma)| \leq \mu \prod_{l=1}^{L-1}\frac{\Gamma(m_l - r_l)\Gamma(r_l)}{\Gamma(m_l)d^{r_l+1}}\frac{f^{\gamma-2(N)}}{e^{\gamma-2N}}. \tag{46}$$

It is this inequality which will allow convergence to be proved for $M_\delta^{(N)}$ when the contours C_1 are distorted to the positive real axis and the regularizing parameter δ is set equal to zero. To show this, let us first evaluate the functions $\psi_m(\gamma)$ of (28):

$$\psi_m(\gamma) = \int_0^1 dt\, t^\gamma P_m(2l-1).$$

This can be evaluated by integration by parts, using Rodrigues' formula, for $\mathrm{Re}\gamma > m$, and then continued to the complete γ plane, to give

$$\psi_m(\gamma) = \Gamma(m-\gamma)/\sin\pi\gamma\,[\Gamma(-\gamma)]^2\Gamma(2+m+\gamma) \quad (47)$$

with poles at $\gamma = -1, -2, \ldots, -m-1$, as expected. Inserting (47) into (23), we have

$$S_\delta^{(N)}(p_1 \cdots p_N) = \pi \prod_{l=1}^{L-1}\left(\frac{i}{2}\int dz_l\, \frac{\lambda^{2z_l}(4\pi)^{2-2z_l}\cos\pi z_l}{\sin[(1+\delta)\pi z_l]\Gamma(z_l)\Gamma(1+z_l)}\right) \times \sum_{\pi\in P_L}\sum_{\mathbf{m}} \frac{a_\pi(\mathbf{m}, \sum_1^L z_l)}{\sin(\pi\sum_1^L z_l)\Gamma(\sum_1^L z_l - 2N + 3)}$$

$$\times \prod_{l=1}^{L-1} \frac{\Gamma(m_l + 2 + \sum^{\pi(l)} z_{l'} - 2N_l + 2l)}{\sin\pi\sum_1^{\pi(l)} z_{l'}\Gamma(1 + m_{\pi(l)} + 2N_{\pi(l)} - \sum_1^{\pi(l)} z_{l'} - \pi(l))\Gamma(1 + \sum_1^{\pi(l)} z_{l'} - 2N_{\pi(l)} + 2\pi(l))} \quad (48)$$

We now have to close the contours C_1 in (48), and then let $\delta \to 0$. We can repeat this again, step by step, where now we have to deal with an expression of the form, dropping the convergence factor for simplicity

$$\prod_{l=1}^{L-1}\int_{C_1} \frac{dz_l}{\sin\pi z_l}\left(\prod_{l=1}^{L-1}\frac{1}{\sin\pi\sum_1^l z_{l'}}\right)\frac{f(z_1\cdots z_L)}{\sin\pi\sum_1^l z_l}, \quad (49)$$

where $f(z_1 \cdots z_L)$ has no poles in the right half-planes of its variables. Let us introduce the substitution operator $S(n,z)$, which acts on a function of z to substitute the value n: $S(n,z)f(z) = f(n)$. The effect of distortion of the z_L contour C_1 to the positive real axis is achieved by removing the factor $[\sin\pi z_L)\sin(\pi\sum_1^L z_l)]^{-1}$ in the integrand of (48) neglecting the various permutations π, and replacing it by the factor

$$\left(\sin\pi\sum_1^{L-1} z_l\right)^{-1}\left(\sum_{n_L} S(n_L, z_L) - \sum_{n_L'} S\left(n_L' - \sum_1^{L-1} z_l, z_L\right)\right).$$

Closing the z_{L-1} contour round the positive real axis replaces the factor

$$\left[\sin\pi z_{L-1}\left(\sin\pi\sum_1^{L-1} z_l\right)^2\right]^{-1}$$

in the new integrand by the factor

$$\left[\left(\sin\pi\sum_1^{L-2} z_l\right)^{-2}\sum_{n_{L-1}} S(n_{L-1}, z_{L-1})(-1)^{n_{L-1}}\right.$$

$$\left. + \left(\sin\pi\sum_1^{L-2} z_l\right)^{-1}\sum_{n_{L-1}'} S\left(n_{L-1}' - \sum^{L-2} z_l, z_{L-1}\right)\frac{\partial}{\partial} z_{L-1}\right].$$

The same step applied to z_{L-2} replaces the factors $[\sin\pi z_{L-2}(\sin\pi\sum_1^{L-2} z_l)]^{-1}$ in the new integrand by the factor

$$\left[\left(\sin\pi\sum^{L-3} z_l\right)^{-r}\sum_{n_{L-2}} S(n_{L-2}, z_{L-2})(-1)^{n_{L-2}}\right.$$

$$\left. + \left(\sin\pi\sum_1^{L-3} z_l\right)^{-1}\sum_{n_{L-2}'} S\left(n_{L-2}' - \sum^{L-3} z_l, z_{L-2}\right)\frac{\partial^{r-1}}{\partial z_{L-2}^{r-1}}\right].$$

There will be a total of 2^{L-1} such terms, the typical one involving a product of various $S(n_i, z_L')$ and various $S(n_j' - \sum^{j-1} z_n, z_j)\partial^l/\partial z_j^l$ substitution operators. The net result of these substitution operators is to produce a set of $(L-1)$-fold series whose typical term is bounded, again to within unimportant terms, and again neglecting similar contributions from the various permutations π, by

$$(\text{const}) \times \sum_{n_1,\ldots,n_{L-1}}\left(\prod_{i=1}^{L-1}\frac{1}{\Gamma(n_i)\Gamma(1+n_i)}\right)\frac{f(p_i p_j)^{\sum n_i}}{\Gamma(\sum n_i)}$$

$$\times \sum_{\mathbf{m}}\prod_{l=1}^{L-1}\frac{\Gamma(m_l + \sum^l n_{l'})\Gamma(m_l - r_l)\Gamma(r_l)}{\Gamma(m_l - \sum^l n_{l'})[\Gamma(\sum^l n_{l'})]^2\Gamma(m_l)d^{r_l+1}}. \quad (50)$$

For each $n_1 \cdots n_{L-1}$, in (50) we choose $r_l = 2\sum^l n_{l'} + 2$, so that, for very large m_l,

$$\frac{\Gamma(m_l + \sum^l n_{l'})\Gamma(m_l - 2\sum^l n_{l'} - 2)}{\Gamma(m_l - \sum^l n_{l'})\Gamma(m_l)} \leqslant \frac{(\text{const})\,2^{2\sum^l n_{l'}}}{m_l^2}, \quad (51)$$

where the constant is independent of m_l and $\sum^l n_{l'}$. Then the summation over $\mathbf{m}$ will certainly be convergent, for each $\mathbf{n}$.

We note that we can only use the bound (46) provided r_l is chosen less than m_l. Thus the summation over each m_l has actually to be split into two parts, one for values greater than $2\sum^l n_{l'}$ and the other for values less than that. We bound this latter contribution by the choice $r_l = 1$, and so we have in detail

$$\sum_{\mathbf{n}}\prod_{l=1}^{L-1}\left[\frac{\Gamma(m_l + \sum^l n_{l'})\Gamma(m_l - 2\sum^l n_{l'} - 2)\Gamma(2\sum^l n_{l'} + 2)\theta(m_l - 2\sum^l n_{l'} - 2)}{\Gamma(m_l - \sum^l n_{l'})\Gamma(m_l)d^{2\sum^l n_{l'}}}\right.$$

$$\left. + \frac{\Gamma(m_l + \sum^l n_{l'})}{\Gamma(m_l - \sum^l n_{l'})}\theta\left(2\sum^l n_l + 2 - m_l\right)\theta\left(m_l - \sum^l n_{l'}\right)\Gamma\left(m_l + \sum^l n_{l'}\right)\Gamma\left(\sum^l n_{l'} - m_l + 1\right)\theta\left(\sum^l n_{l'} + 1 - m_l\right)\right], \quad (52)$$

where $\theta(n) = 1 \quad \text{for } n \geq 0, \quad \theta(n) = 0 \quad \text{for } n < 0.$

Since, for all n,

$$\sqrt{2\pi}n^{n+1/2}e^{-n} < \Gamma(n+1) < 2\sqrt{2\pi}n^{n+1/2}e^{-n}$$

then

$$\sum_{m=0}^{n} \Gamma(m+n)\Gamma(n-m+1) < [\Gamma(n)]^2 \cdot 2^{2n}n^2(\text{cons}),$$
$$\sum_{m=n+1}^{2n+2} \frac{\Gamma(n+m)}{\Gamma(m-n)} < (\text{const})[\Gamma(n)]^2(9e^2)^n n \tag{53}$$

for all m and n, where the constants are independent of m and n, the contributions from these finite ranges of m_l will also be convergent, and give the same behavior for large n as that arising from (50) and (51). By putting together (50), (51), (52), and (53), the summation over n will be bounded, to unimportant factors, by

$$\sum_{n_1 \cdots n_{L-1}} \left(\prod_{l=1}^{L-1} \frac{\Gamma(2\sum^l n_{l'})}{\Gamma(n_l)\Gamma(1+n_l)[\Gamma(\sum^l n_{l'})]^2} \right) \frac{f(p_i p_j)^{\Sigma n_i}}{\Gamma(\sum n_j) d^{N \Sigma n_j}} \tag{54}$$

which is certainly finite. Again we may read off the behavior of (54) for some subset of the $(p_i p_j)$ becoming large; its value is

$$\sum \frac{(f(p_i p_j))^{rn}}{n^{3rn}} \sim e^{[\max_{i,j}(p_i p_j)]^{1/3}} \tag{55}$$

if r of the integers n_i are summed to infinity in (51), the remaining $(L-1-r)$ being held finite.

We now have to prove the existence of the limit as $\epsilon \to 0$ of the quantity $S_0{}^{(N)}(p_1 \cdots p_N)$ obtained by the previous discussion. In the process we will justify the use of the inequality $\gamma > 2N$ in the discussion after Eq. (45). To obtain a limit as $\epsilon \to 0$, it is evident that we must separate $S_0{}^{(N)}(p_1 \cdots p_N)$ into two contributions, one involving contributions from poles in the variables z_l for which $\gamma = \sum_1^l z_l < 2(N-1)$ and the other with $\sum_1^l z_l \geq 2(N-1)$. The proof of the convergence of the Legendre polynomial expansion breaks down for the first of these contributions, since for it the right-hand side of (14) becomes infinite as $\epsilon \to 0$, at least for certain timelike values of the external momenta. The previous discussion certainly applies, even at $\epsilon = 0$, in the Euclidean region of external momenta. However, in order to obtain the physical amplitudes directly, without continuation from the Euclidean region, it is necessary to perform the above-mentioned separation explicitly. This we do as follows.

We firstly perform a suitable number of integrations by parts with respect to the t_l-variables in (24). If there are m_l integrations by parts for t_l then the factor $\int_0' dt_l t_l^{\gamma_{\pi(l)}}$ in (24) will be replaced by

$$\left[\sum_{r=1}^{m_l} \prod_{s=1}^{r} \frac{1}{(\gamma_{\pi(l)}+s)} \delta^{(r-1)}(t_l - 1) + \prod_{s=1}^{m_l} \frac{1}{(\gamma_{\pi(l)}+s)} (-1)^{m_l} dt_l^{\gamma_{\pi(l)}+m_l} \left(\frac{\partial}{\partial t_l}\right)^{m_l} \right]. \tag{56}$$

We take $m_l = -2(N^l - 1) + 2NL$, where N^l is the number of vertices contained in the first l lines. Then the last term in this new factor (56) will have poles in $\sum_{N=1}^{\pi(l)} z_{l'}$ which have not been made explicit, in other words are not in the factor $\prod_{s=1}^{m_l}[\gamma_{\pi(l)}+s]^{-1}$, at $\sum_{l'=1}^{\pi(l)} z_{l'} \geq 2LN$.

We may thus rewrite (24) as a sum of terms, arising when the expression (56) is substituted for $\int_0' dt_l t_l^{\gamma_l}$ in (24), the characteristic one being (dropping the permutation π

$$\pi \prod_{l=1}^{L} \left[\frac{i}{2} \int_{C_1} dz_l f_\delta(z_l) \right] \frac{l^{iN\pi}}{\sin(\pi \sum_1^l z_l)\Gamma(\sum_1^L z_l - 2N + 3)}$$
$$\times \prod_{l=1}^{L'} \int_0^1 db_l \prod_{s=1}^{\gamma_l} \frac{1}{(\gamma_l + s)} \delta^{(\gamma_l - 1)} \prod_{l=L'+1}^{L-1} \prod_{s=1}^{2NL-2(N^l-1)} \frac{1}{(\gamma_l + s)}$$
$$\times \int_0^1 dt_l t_l^{\gamma_l + 2NL - 2(N-1)} \left(\frac{\partial}{\partial t_l}\right)^{2LN-2(N^l-1)}$$
$$\times \left(\frac{F}{E} + i\epsilon\right)^{\Sigma z_l - 2(N-1)} E^{-2} \tag{57}$$

where

$$f_\delta(z) = \frac{\lambda^{2z}(4\pi)^{2-2z}\cos\pi z}{\sin[(1+\delta)\pi z]\Gamma(z)\Gamma(1+z)}$$

We may now shift the contours C_1 for each z_l variable to the right by an amount $(2N-1)$, to the contour which we denote by C_{2N}; this contour will be parallel to the imaginary axis and have real part between $2N-1$ and $2N$. Following the discussions already given in this and the preceding sections, the resulting contribution will be a sum of terms, each one involving some of the z_l's being integrated over. In all these cases the summation is only over a finite set of integers, and the new components of the variables t_i will all be greater than minus one. Furthermore, in each of those terms with at least *one* z_l variable integrated over C_{2N}, the exponent of $(F/E + i\epsilon)$ in (57) will be greater than zero, so that the preceding method of expansion in Legendre polynomials in the remaining variables t_j can be applied, even for $\epsilon = 0$, thus allowing the value of the regularizing parameter to be taken to zero. The only singular term will be that arising when there are no variables z_l integrated along C_{2N}. But then this contribution is regular in δ as $\delta \to 0$, as can be seen by inspection, and for $\delta = 0$ will be a sum of a finite number of terms, each one involving an integration over a subset of the variables $(t_1 \ldots t_{L-1})$ of a finite derivative of a finite inverse power of $(F/E + i\epsilon)$, multiplied by E^{-2}, and possibly multiplied by a finite power of $\log(F/E + i\epsilon)$ (from residues at multiple poles in the z-variables). But the arguments of Hepp[5] then apply to this case, provided they are extended to include the distributions $[\log(x+i0)]^\gamma (x+i0)^{-n}$. Since these products can always be written as finite sums of well-defined distributions[3] this extension is immediate.

Thus the existence of the boundary value of $S_0{}^{(N)}(p_1 \cdots p_N)$ as an element of a suitable space of generalized functions has been proved. The estimate (55) is valid for the contribution arising from those terms of (57) which have at least one z_l variable integrated over C_{2N}, and so this space must contain at least functions with the corresponding high energy increase in each of the invariants. There are only a finite number of terms on the remainder, each being a distribution in the space $\mathfrak{D}^1(\mathbf{R}^{4(N-1)})$. Thus the sum of all of these contributions will be an element of the generalized function space S'_α in each of the components of the momenta, with $\alpha < \frac{3}{2}$.

5. UNITARITY

We prove unitarity using the notation of Bogoliubov and Shirkov.[6] This is discussed in terms of the S-matrix functional

$$S[f] = \sum_N G^N \int S_N(x_1 \cdots x_N) f(x_1) \cdots f(x_N) dx_1 \cdots dx_N$$

and the unitarity condition for the operators S_N then reads

$$\sum_{k=0}^{N} p(k \mid N-k) S_k S_{N-k}^{+} = 0, \tag{58}$$

where $P(k \mid N-k)$ is the operation of summing over all possible divisions of the vertices $X_1, \ldots, X_N$ into two subsets with k and $(N-k)$ members, respectively. By using that

$$S_N(x_1 \cdots x_N) = : \prod_{i=1}^{N} e^{\lambda \phi_i} : S_\delta^{(N)}(x_1 \cdots x_N)$$

where $S_\delta^{(N)}$ is given by (11) as

$$S_\delta^{(N)}(x_1 \cdots x_N) = i^N \prod_{i<j} e^{\lambda^2 \Delta_{ij}} \Big|_\delta^{C_0} \tag{59}$$

the unitarity condition (58) becomes

$$\sum_{k=0}^{N-k} P(k \mid N-k) S_0^k(x_1 \cdots x_k) S_0^{N-k*}(x_{k+1} \cdots x_N) \times \prod_{\substack{i \in k \\ j \in N-k}} e^{\lambda^2 \Delta_{ij}^{(-)}} : \prod_{i=1}^{N} e^{\lambda \varphi(x_i)} = 0, \tag{60}$$

where $\Pi_{i \in k, j \in N-k}$ denotes the product over all choices of the index i taken from the set $(1, 2, \ldots, k)$ and the index j from the set $((k+1), k+2, \ldots, N)$.

Hence a sufficient condition for unitarity is

$$\sum_{k=0}^{N} P(k \mid N-k) S_0^{(k)}(x_1 \cdots x_k) S_0^{(N-k)*}(x_{k+1} \cdots x_N) \times \prod_{i \in k, j \in N-k} e^{\lambda^2 \Delta_{ij}^{(-)}} = 0. \tag{61}$$

In order to prove (61), we will first prove a modified form of unitarity for the regularized generalized functions $S_\delta^{(N)}(x_1 \cdots x_N)$ for $\delta > \frac{3}{2}$. To obtain this, we will consider a regularized version, denoted by $e^{\lambda^2 \Delta^{(-)}} \Big|_\delta^{C_0}$ of the generalized function $e^{\lambda^2 \Delta^{(-)}}$. This will be defined, by analogy with (8), to have the Fourier transform

$$e^{\lambda^2 \Delta^{(-)}} \Big|_\delta^{C_0} = \frac{i}{2} \int_{C_1} \frac{dz (\lambda^2)^z (4\pi)^{2-2z} (p_+^2)^{z-z} e^{-i\pi z}}{\tan(\pi z)\, \delta m[(1+\delta)\pi z] \Gamma(z-1) \Gamma(z) \Gamma(z+1)} + \frac{(2\pi)^4 \delta^4(p),}{(1+\delta)}, \tag{62}$$

where p_+^2, the function defined by

$$p_+^2 = p^2 \quad (\text{for } p^2 > 0, p_0 > 0)$$
$$= 0 \quad (\text{otherwise}),$$

This is again finite if $\delta > \frac{3}{2}$.

Since for all z

$$[\Delta_F(x)]^z = \theta(x_0)[\Delta^{(+)}(x)]^z + \theta(-x_0)[\Delta^{(-)}(x)]^z,$$

then

$$e^{\lambda^2 \Delta} \Big|_\delta^{C_0} = \theta(x_0) e^{\lambda^2 \Delta^{(+)}} \Big|_\delta^{C_0} + \theta(-x_0) e^{\lambda^2 \Delta^{(-)}} \Big|_\delta^{C_0}. \tag{63}$$

Hence the "cutting formula" of Veltman[7] will apply to graphs built up with lines $e^{\lambda^2 \Delta} \Big|_\delta^{C_0}$, provided the sums over intermediate states are given by $e^{\lambda^2 \Delta^{(-)}} \Big|_\delta^{C_0}$. In other words the "regularized unitarity" condition

$$\sum_{k=0}^{N} P(k \mid N-k) S_\delta^{(k)}(x_1 \cdots x_k) S_\delta^{(N-k)*}(x_{k+1} \cdots x_N) \times \prod_{i \in k, j \in N-k} e^{\lambda^2 \Delta_{ij}^{(-)}} \Big|_\delta^{C_0} = 0 \tag{64}$$

will be valid. We may now extend this to $\delta = 0$, since in momentum space the generalized function $\widetilde{e^{\lambda^2 \Delta^{(-)}}}$ is a multiplier of generalized functions of the invariant k^2 which are smooth at $k^2 = 0$ and are in the space S_α^1 for $\alpha < \frac{3}{2}$. Thus on the left-hand side of (64) we may first distort all the contours C_1 back to the positive real axis, in all three factors of the kth term, and then let $\delta \to 0$, with assurance that the left-hand side of (64) will remain finite. The resulting expression is then analytic in δ, for δ along the real axis down to the value zero, as may be seen by direct inspection of formulas similar to (34) but now with $\delta \neq 0$. Hence the right-hand side of (64) remains zero, and the unitarity condition (61) for the physical amplitudes $S_0^{(N)}(x_1 \cdots x_N)$ has been proved.

6. CAUSALITY

With the same notation as the preceding section the causality condition for the amplitudes $S_0^{(N)}$ is[6]

$$\sum_{k=0}^{N} P(k \mid N-k) S_0^{(k+1)}(x_0, x_1, \ldots, x_k) S_0^{(N-k)*}(x_{k+1} \cdots x_N) \times \prod_{\substack{i \in k+1 \\ j \in N-k}} e^{\lambda^2 \Delta_{ij}^{(-)}} = 0 \tag{65}$$

if at least one of $x_1 \cdots x_N$ is spacelike or in the backward light cone to x_0, where the operator $P(k, N-k)$ only divides the vertices $(x_1 \cdots x_N)$ into two subsets of k and $(N-k)$ elements and leaves the x_0 dependence always in the first factor $S_0^{(k+1)}$. We may formulate a "regularized causality" in the same fashion as the regularized unitarity of the previous section. This will read

$$\sum_{k=0}^{N} P(k \mid N-k) S_\delta^{(k+1)}(x_0 \cdots x_k) S_\delta^{(N-k)*}(x_{k+1} \cdots x_N) \times \prod_{\substack{i \in k+1 \\ j \in N-k}} e^{\lambda^2 \Delta_{ij}^{(-)}} \Big|_\delta^{C_1} = 0. \tag{66}$$

Again (66) is valid for $\delta > \frac{3}{2}$ from Veltman's cutting formula,[7] and again it can be continued down to $\delta = 0$, the regularized quantities in the left-hand side having the contours C_1 distorted back to the positive real axis and then δ being allowed to go to zero in the resulting expressions to give (65).

Of course, it is only possible to write (65) as a relation between generalized functions provided that these have test functions of compact support. That this is the case for the exponential interaction was shown by (55), in other words that

$$S_0^{(N)}(p_1 \cdots p_N) \in S_{\frac{1}{2}}(\mathbf{R}^{4(N-1)}) \tag{67}$$

for any $\alpha < \frac{3}{2}$. The corresponding space of coordinate space test functions is then $S^{\alpha}(\mathbf{R}^{4(N-1)})$, which is known to have a dense set of functions with compact support provided $\alpha > 1$. This we can certainly choose, so that the causality condition (65) can be given a generalized function formulation without the need for added complications.

7. RELATION TO OTHER REGULARIZATION SCHEMES

There have been various regularization schemes developed recently to discuss the superpropagator, $N = 2$. All of these methods lead to the same parameter-free result with the branch cut in λ^2. We wish to discuss how the extension of these various schemes to the case of $N > 2$ produces the same results as those obtained in this paper.

We start by remarking that there is an ambiguity allowed in our regularization through the introduction of the parameter δ which is, for $N = 2$, that arising by replacing the factor $\Gamma(1 + z)^{-1}$ by $[\Gamma(1 + z)^{-1} + f(z)\sin\pi z]$. The function $f(z)$ is to a certain extent arbitrary, though assumed to be analytic and to have finite order of growth less than or equal to unity. Evidently an identical substitution in the higher-order terms $S_{\delta}^{(N)}$ can be performed and finite results obtained for the physical quantities $S_0^{(N)}$ under further suitable conditions on f provided the modifications of all superpropagators are identical, and a similar modification made in $e^{\lambda^2\Delta^{(-)}}|_{\delta}^{c}$, then unitarity and causality will be satisfied by these physical quantities. In other words unitarity and causality in higher orders gives no restriction on the ambiguity present in the definition of the superpropagator.

Let us now turn to discuss in detail other regularization schemes than that used in this paper. One of these, proceeding through a sequence of L_2 functions[8], can be shown, by a little manipulation, to be the same as the introduction of the convergence factor $\exp[-(\sum^{L}\alpha_l^{-1})c]$ into (16), where c is a positive constant. This evidently introduces convergence of the α_l integrals in (16) at $\alpha_l = 0$, even if the z_l have $\mathrm{Re}z_l > 0$. In order to be able to let $c \to 0$, it appears necessary first to go through the procedure of performing the t_l integrals and deforming the C_1 contours to the positive real axis, as described in Sec. 4. However, the function $\exp[-c\sum^{L}\alpha_l^{-1}]$ cannot be expanded following the method of Sec. 4, though it can by method (ii) of Sec. 3. Since that approach has not been completed, it is not clear that such a regularization will succeed, or if it does that it will agree with the results of Sec. 4.

Another approach to regularization of the superpropagator has recently been proposed[9] along the lines of analytic renormalization.[4] This consists in replacing the usual Feynman propagator Δ by $\Delta^{1-\epsilon}$ and keeping ϵ in a suitable region in order to perform the necessary manipulations without catastrophes. If this is done in the Nth-order term $S^{(N)}$, we obtain

$$S_{\epsilon}^{(N)}(p_1\cdots p_N) = \pi e^{i\pi N}\prod_{l=1}\int_{\Gamma}\frac{i}{2}\times\frac{dz_l\lambda^{2z_l}(4\pi)^{2-2z_l}}{\tan(\pi z_l)\Gamma(1+z_l)\Gamma((1-\epsilon)z_l)}$$

$$\times\frac{1}{\sin\pi[2N-3-(1-\epsilon)\sum^{L}z_l]\Gamma[3-2N+\sum^{L}z_l(1-\epsilon)]}$$

$$\times\sum_{\pi\in P_L}\left(\prod_{l=1}^{L-1}S_{\delta}^{1}dt_l t_l^{\lambda_{\pi}^{(\epsilon)}(l)}\right)\frac{(-F_{\pi})^{-2(N-1)+(1-\epsilon)\Sigma^{L}z_l}}{(E_{\pi})^{-2N+(1-\epsilon)\Sigma^{L}z_l}} \tag{68}$$

where $\gamma_l^{(\epsilon)} = 2(l - N_l) - 1 - \sum^{l}z_{l'}(1-\epsilon)$ and Γ is a contour encircling the positive real axis counter clockwise. The poles of the integrand of (67) at $z_l = n_l/(1-\epsilon)$ pinch the contours Γ at the poles $z_l = n_l$ as $\epsilon \to 0$ with $\mathrm{Im}\epsilon > 0$. If these former pole contributions are dropped on expansion of the contours Γ and then $\epsilon \to 0$, precisely the residues in the z_l variable at the nonnegative integers will have to be calculated. They are precisely those calculated in Secs. 3 and 4, though without the need to justify the rejection of the residues at $n_l(1-\epsilon)^{-1}$. In any case the analytic renormalization approach is seen to lead to precisely the same results as the method used in this paper. Another approach to renormalization has been given by Lehman and Pohlmeyer.[10] It is to be expected that this regularization scheme produces the same finite amplitudes as the one of this paper; we hope to discuss this elsewhere.

8. MORE GENERAL INTERACTIONS

Let us consider the class of interaction Lagrangians for massless particles:

$$L_{\mathrm{int}} = \int_0^{\infty} l(t) : e^{t\phi} : dt.$$

The rules for writing down the Nth-order contribution in L_{int} to S-matrix elements have been given elsewhere.[2] Each of the N vertices acquires a factor $l(t_j)t_j^{m_j}$ if there are m_j external particles at the jth vertex, while for the line joining vertices i and j the constant λ^2 is replaced by $t_i t_j$. Finally the n variables t_j are integrated over. This is equivalent to introducing a factor $a(m_j + \sum_{l\in j} z_l)$ for the jth vertex into the Eq. (12), where $\sum_{l\in j}$ is summation over the lines l which meet at this vertex, and

$$a(n) = \int_0^{\infty} t^n l(t)\, dt.$$

We assume that $a(n)$ has an analytic continuation into the n plane and is of exponential growth of order $\frac{1}{2}\alpha$ at infinity; the superpropagator is localizable for $\alpha < 1$ and definitely nonlocalizable if $\alpha > 1$. The z_l integrations along C_1 are then convergent in (24) provided that $\delta > (\frac{3}{2} - \alpha)$. We see that only in the nonlocalizable case $\alpha > \frac{3}{2}$ will these integrals converge without regularization. Since that situation has the difficulty that the resulting series for the superpropagator has zero radius of convergence in $\Delta(x)$ in coordinate space, and its definition is fraught with perils,[11] we will not discuss this in any detail here. We can certainly determine the resulting high energy behavior, following along the argument given at the end of Sec. 4, to be

$$\exp[(\max_{i,j}|p_i p_j|)^{1/(3-\alpha)}].$$

Such behavior agrees precisely with that for the superpropagator, $N + 2$, and is evidently nonlocalizable if $\alpha > 1$. It is also evident that the proofs of unitarity and causality given in Secs. 7 and 8 go through exactly as for the pure exponential interaction, though evidently breaking down at causality for the nonlocalizable case of $\alpha > 1$.

9. DISCUSSION

We have discussed a prescription for obtaining finite, unitary and causal amplitudes to each order in the major coupling constant G for a class of nonderivative non-polynomial Langrangians for massless fields. There are numerous questions which now need to be solved:

(1) Are all other properties of physical interest, particularly that of positive metric, satisfied by the prescription for localizable theories?

(2) Is it possible to sum over the major coupling constant so that on-mass-shell S-matrix elements are polynomially bounded, at least for localizable interactions, as has been shown to be necessary?[12]

(3) Can this approach be extended to massive particles, for example, along the lines suggested by Karowski?[13]

(4) Can derivative interactions be included, at least in the quadratic case of chiral couplings and general relativity?

We hope to give answers to these questions elsewhere.

ACKNOWLEDGMENTS

I would like to thank B. W. Keck for helpful discussions and A. Salam for much stimulation and inspiration.

---

†Now at: Department of Mathematics, Kings College, London, England.

[1]A. Salam "Computation of Renormalization Constants," *Proceedings of the Coral Gables Conference on Fundamental Interactions at High Energy, 1971* (Gordon and Breach, New York, 1971).
[2]See, for example, J. G. Taylor, "Non-Polynomial Lagrangians," in Ref. 1.
[3]I. M. Gel'fand and F. E. Shilov, *Generalized Functions, Vol. 1* (Academic, New York, 1964).
[4]E. R. Speer, J. Math. Phys. (N.Y.) **9**, 1404 (1968).
[5]K. Heep, Commun. Math. Phys. **2**, 301 (1961).
[6]N. N. Boguliubov and D. Shirkov, *Theory of Quantized Fields* (Interscience, New York, 1969).
[7]M. Veltman, Physica (Utrecht) **29**, 186 (1963).
[8]R. Blomer and F. Constantinescu, "On the Zero Mass Superpropagator," Univ. of Munich preprint (1970).
[9]P. K. Mitter, "On the Analytic Approach to the Regularization of Weak Interaction Singularities," Oxford University preprint (1970).
[10]H. Lehman and K. Pohlmeyer, Commun. Math. Phys. **20**, 101 (1971).
[11]S. Fels, Phys. Rev. D **1**, 2370 (1970).
[12]H. Epstein, V. Glaser, and A. Martin, Commun. Math. Phys. **13**, 257 (1969).
[13]M. Karowski, Commun. Math. Phys. **19**, 289 (1970).

Method of calculating quasiaverages

N. N. Bogolubov Jr.[*†]
The Institute for Theoretical Physics, State University of New York, Stony Brook, New York 11790
(Received 3 July 1972)

Some features of quasiaverages for model systems with four-fermion interaction are considered. A new method of defining quasiaverages for the systems under consideration is proposed.

Up to the present time, the class of exactly soluble dynamical model systems has consisted mainly of one- and two-dimensional systems.[1,2] In this paper we shall concentrate on the treatment of certain three-dimensional model systems which can be solved exactly.

In a number of works[3,4] we have dealt with so-called model problems of statistical physics, which allow asymptotically exact solution (for $V \to \infty$, where V is the volume of the system). Our investigations have yielded asymptotically exact expressions not only for the free energy, but also for the Green's functions and many-time correlation functions. It is worthwhile to mention that the investigations have been carried out with mathematical rigor, and a special majorization technique has been devised to establish the fact of asymptotic accuracy.

While investigating the Green's functions and many-time correlation functions, we had to make use of the notion of quasiaverages and to introduce into the Hamiltonian under review the so-called source terms, which tended to zero after the limiting transition $V \to \infty$ was performed.

Now we explain the above procedure by considering one of the simplest examples of model systems studied by us, namely, the system characterized by the BCS Hamiltonian

$$H = T - V(L^{\dagger} \cdot L)/2$$

$$T = \sum_{(f)} T_f a_f^{\dagger} a_f, \tag{1}$$

$$L = (1/V)\sum \lambda_f a_{-f} a_f .$$

Here $a_f, a_f^{\dagger}$ are the Fermi operators, V is the volume of the system, $f = (p, \sigma)$, the union of momentum p and spin σ, the momentum p takes the usual quasidiscrete values, and $T_f = p^2/2m - \mu$, μ being the chemical potential. Some general conditions are imposed on λ_f, which must vanish fast enough when $|p| \to \infty$. If one sets up the following equation of motion for the Hamiltonian,

$$i\frac{da_f}{dt} = T_f a_f - a_{-f}^{\dagger} L \cdot \lambda_f, \tag{2}$$

then by the definition of L it is clear that the operators $L, L^{\dagger}$ commute with our operators a_f, a_f with an accuracy up to the values of the order $1/V$. Therefore it would seem natural to assume that the operator L is almost a C-number. But then

$$\langle L \rangle_H = C,$$

where the usual average of Hamiltonian H is denoted by $\langle \cdots \rangle_H$,

$$\langle \cdots \rangle_H = \mathrm{Sp}(\cdots e^{H/\theta})/\mathrm{Sp}e^{-H/\theta}.$$

As the Hamiltonian H is invariant under the gradient transformations

$$a_f \to e^{i\phi} a_f, \qquad \phi = \text{const},$$

it is easy to see that all

$$\langle a_{-f} a_f \rangle_H = 0,$$

and, consequently,

$$\langle L \rangle_H = 0.$$

Thus it turns out that the operator L itself can not even "approximately" be regarded as a C-number. In order to avoid this difficulty, we made use of the notion of a quasiaverage and introduced source terms into the Hamiltonian. Instead of H we considered the Hamiltonian

$$\Gamma = H - \zeta \cdot V(L + L^{\dagger}), \qquad \zeta > 0. \tag{3}$$

We defined the quasiaverages over H as the usual averages over the Hamiltonian Γ, in which $\zeta \to 0$ occurs after the limiting transition $V \to \infty$. As a result we could produce a rigorous proof that for any $\zeta > 0$

$$\langle\!\langle (L - C)(L^{\dagger} - C) \rangle\!\rangle \to 0, \qquad \text{for } V \to \infty,$$

where C is some positive value. Together with exact equations of motion (2) we made use of the "approximate equations" of motion

$$i\frac{da_f}{dt} = T_f a_f - a_{-f}^{\dagger} C \cdot \lambda_f, \tag{4}$$

which, by the way, correspond to a Hamiltonian

$$H(C) = T - \tfrac{1}{2} V(CL^{\dagger} + CL - C^2), \tag{5}$$

which we call the "approximative" Hamiltonian. Using a specially devised majorization technique, we could prove that the correlation functions for a product of the Fermi operators constructed from (3) are asymptotically close to the correlation functions constructed from the Hamiltonian (5) in the sense of quasiaverages, i.e., the limiting transition $V \to \infty$ is followed by $\zeta \to 0$. As a result we obtained majorization estimates for the difference of the correlation functions constructed from the Hamiltonians (1), (5).

The corresponding estimates of the approximation for the correlation Green's functions at $V \to \infty$ were not uniform with respect to $\zeta \to 0$; therefore, the order of the limiting transitions, namely, $V \to \infty$ followed by $\zeta \to 0$, was quite important. This fact constrained the definition of quasiaverages (5). The question arises, however, as to how one can define quasiaverages for the given Hamiltonian, without introducing source terms since, physically, the quasiaverages characterize the system under review and the introduction of sources is, in a sense an artificial trick violating the invariance properties of the system. As we see, such a program of defining the quasiaverages is feasible, though the majorization estimates become much more cumbersone. The

point is that in this case we can not prove that the expression $\langle (L - C)(L^\dagger - C)\rangle_H$ is a small value, as this average turns out to be far from being small; instead it is larger than C^2. Nonetheless, we manage to prove in a rigorous way that the value

$$\langle (L^\dagger L - C^2)^2\rangle_H \leq \mu_V \to 0, \quad \text{for } V \to \infty, \tag{6}$$

is small and, besides, that

$$\left\langle \frac{dL^\dagger}{dt} \cdot \frac{dL}{dt} \right\rangle_H \leq \nu_V \to 0, \quad \text{at } V \to \infty. \tag{7}$$

We may understand the special importance of this proof if we consider the Hamiltonian without sources, H [Eq. (1)], and introduce the auxiliary operator constructions

$$\begin{aligned} \alpha_f &= u_f a_f + v_f a^\dagger_{-f}(L/C), \\ \alpha^\dagger_f &= u_f a^\dagger_f + v_f (L^\dagger/C) a_{-f}, \end{aligned} \tag{8}$$

where

$$\begin{aligned} u_f &= 2^{-1/2}(1 + T_f/E_f)^{1/2}, \\ v_f &= [-\epsilon(f)/\sqrt{2}](1 - T_f/E_f)^{1/2}, \\ E_f &= (C^2\lambda_f^2 + T_f^2)^{1/2}. \end{aligned}$$

It may be mentioned that these operator constructions (8) "satisfy only approximately" the commutation relations of the Fermi statistics when $f \neq f'$. If the operator L were a C-number and $L = C$, then these constructions would exactly coincide with the "new" Fermi operators[5] associated with the "old" canonical u–v transformation.

Writing down the equations of motion with respect to the operator constructions (8), taking into account Eqs. (2), we can reduce them to the following form after performing cumbersome manipulations:

$$i \frac{d\alpha^\dagger_f}{dt} + E_f \alpha^\dagger_f = R_f, \tag{9}$$

where the "average", i.e., $\langle R^\dagger_f \cdot R_f\rangle_H$ is just expressed in terms of these two averages (6), (7).

Let us then take up the estimations for the averages (6), (7). To prove inequality (6), we apply the theorem of proximity of free energies.[4] At one time this theorem was proved for the BCS Hamiltonian. However, it became apparent afterwards that the results of this theorem were of a more general nature, and this made it possible to apply the theorem for the asymptotically exact $V \to \infty$ calculations of the simplest binary correlation averages. Further, this theorem was generalized to a broader class of the model systems, its range of applicability extending considerably. We shall show that the results of this theorem are sufficiently strong to provide for the proof of the above-mentioned inequalities (6), (7).

Now we state Theorem 1.

Theorem 1: Let the operators T, L, $L^\dagger$ in the Hamiltonian H satisfy conditions

$$\left.\begin{aligned} &|\lambda_f| \leq Q = \text{const} \\ &(1/V)\sum_{(f)} |\lambda_f \cdot T_f^2| \leq Q_0 = \text{const} \end{aligned}\right\} I',$$

where Q and Q_0 are constants as $V \to \infty$, and let the free energy calculated per unit volume for the Hamiltonian $T = \Sigma_f T_f a^\dagger_f a_f$ be restricted by a constant

$$|(\theta/V)\ln \mathrm{Sp} e^{-T/\theta}| \leq M_0 = \text{const}.$$

The approximative Hamiltonian for system (1) is (5). The the following inequalities are valid:

$$0 \leq \mathrm{abs} \min_{(C)} f_{(\infty)}(H(C)) - f_V(H) \leq \epsilon'(1/V),$$

and here the value $\epsilon'(1/V) \to 0$ uniformly with respect to θ in the interval $0 < \theta \leq \theta_0$, where θ_0 is an arbitrary temperature, and

$$f_{(\infty)}(H(C)) = \tfrac{1}{2}C^2 - \tfrac{1}{2}(2\pi)^{-3}[E_f - T_f - 2\theta \ln(1 + e^{-E_f/\theta}]df.$$

In order to apply this theorem for the proof of inequalities (6), (7), we give reason as follows. Consider the systems which are defined by the Hamiltonian dependent linearly on some parameter τ:

$$H_\tau = \Gamma_0 + \tau\Gamma_1. \tag{10}$$

We define formally the expression

$$f_V(H_\tau) = -(\theta/V) \ln \mathrm{Sp} e^{-H_\tau/\theta},$$

which we call a free energy per unit volume V for the model system H_τ.

One can prove the validity of the inequality

$$\frac{d^2 f_V(H_\tau)}{d\tau^2} \leq 0.$$

Denoting

$$\frac{d}{d\tau} f_V(H_\tau) = \frac{1}{V} \frac{\mathrm{Sp}\Gamma_1 e^{-H_\tau/\theta}}{\mathrm{Sp} e^{-H_\tau/\theta}} = \frac{\langle \Gamma_1\rangle_{H_\tau}}{V}.$$

We arrive at inequalities that hold for any operators Γ_0, Γ_1:

$$(1/V)\langle \Gamma_1\rangle_{\Gamma_0+\Gamma_1} \leq f_V(\Gamma_0 + \Gamma_1) - f_V(\Gamma_0) \leq (1/V)\langle \Gamma_1\rangle_{\Gamma_0} \tag{11}$$

To estimate (6), we set

$$\begin{aligned} H &= \Gamma_0 + \Gamma_1, \quad \Gamma_1 = \rho G = \rho V(L^\dagger L - C^2)^2, \\ \Gamma_0 &= H - \Gamma_1, \end{aligned}$$

where ρ is a fixed positive number. From (11) we get

$$(\rho/V)\langle G\rangle_H \leq f_V(H) - f_V(H - \rho G),$$

and, consequently,

$$\langle (L^\dagger L - C^2)^2\rangle_H \leq (1/\rho)[f_V(H) - f_V(H - V\rho(L^\dagger L - C^2)^2)]. \tag{12}$$

It remains now to apply the above-stated theorem for the right-hand side of (12) and to show that these free energies coincide at the limit $V \to \infty$. We introduce the following "abbreviation." When we say that the system with the Hamiltonian

$$H' = \{H - V\rho(L^\dagger L - C^2)^2\}$$

is approximated by the system

$$H(S) = \{H - V\rho(2S[L^\dagger L - C^2] - S^2)\}, \tag{13}$$

we understand this to mean that the corresponding free energies constructed on their basis are close in the sense of Theorem 1.

Keeping in mind that

$$H = T - \tfrac{1}{2}VL^{\dagger}L,$$

we write (13) as

$$\{T - V(2S\rho + \tfrac{1}{2})L^{\dagger}L + (S^2 + 2SC^2)V\rho\}. \tag{14}$$

Using Theorem 1, we see that (14) is approximated by a form quadratic in the Fermi operators

$$H(S, C') = \{T - V(\tfrac{1}{2} - 2S\rho)(C'L^{\dagger} + \overset{*}{C}'L - \overset{*}{C}'C') + V\rho(S^2 + 2SC^2)\}. \tag{15}$$

We recall that the system (13) is approximated by the form (5). We remark that, in (5), C is a complex number in the general case, but it can be shown that the absolute minimum of the function $f(H(C))$ is realized with a real C. Therefore, C can be regarded as real in the case of the approximative Hamiltonian $H(C)$. Comparing the "approximative forms" of Hamiltonian (15) and (5), we see that in order for the right-hand side inequality (12) to be of the order $\eta(1/V) \to 0$ at $V \to \infty$, one has to choose the solution

$$C' = C \quad \text{and} \quad S = 0.$$

Now it is easy to verify that such a solution really exists and, besides, one can find $\rho = \rho_0 > 0$, such that there exists an unique solution of the problem as absolute minimum of the free energy constructed from the form (15). In other words, proceeding from this theorem, we find that

$$\langle (L^{\dagger}L - C^2)^2 \rangle_H \leq (1/\rho)\epsilon(1/V) \to 0, \quad \text{for } V \to \infty. \tag{16a}$$

Inequality (7) is proved in a similar way. As a result of this reasoning we arrive at the estimation

$$\langle R_f^{\dagger} R_f \rangle_H \leq \bar{\epsilon}(1/V), \tag{16b}$$

where $\bar{\epsilon}(1/V) \to 0$ for $V \to \infty$.

Proceeding from the equations of motion (9) and taking into account the estimation (16a), we can follow reasoning similar to that in Ref. 6, obtaining a corresponding estimation of the difference of the correlation functions constructed from Hamiltonian (1) and the corresponding approximative Hamiltonian (5).

We now turn to an important lemma.

Lemma 1: Let the equations of motion for the operators α_f, where $|\alpha_f| \leq b_1 = \text{const}$, have the form

$$i\frac{d\alpha_f}{dt} = E_f\alpha_f + R_f,$$

where

$$\langle R_f^{\dagger}R_f + R_f R_f^{\dagger} \rangle_H \leq 2\bar{\epsilon}_V \to 0 \quad \text{for } V \to 0,$$

and let B be a bounded operator $|B| \leq b_2 = \text{const}$. Now we set up the difference

$$\mathfrak{D} = \langle \alpha_f^{\dagger} B \rangle_H - e^{-E(f)/\theta}\langle B \cdot \alpha_f^{\dagger} \rangle_H.$$

Then the following estimate holds:

$$|\mathfrak{D}| \leq (2/\theta)b_2(\bar{\epsilon}_V)^{1/2}.$$

Proof: Using the spectral representation for two-time averages,[7,8] we write averages for $\mathfrak{D}$:

$$\langle \alpha_f^{\dagger}(t)B(\tau) \rangle_H = \int_{-\infty}^{+\infty} J_{\alpha^{\dagger}B}(\omega)e^{i\omega(t-\tau)}d\omega,$$

$$\langle B(\tau)\cdot\alpha_f^{\dagger}(t) \rangle_H = \int_{-\infty}^{+\infty} J_{\alpha^{\dagger}B}(\omega)e^{\omega/\theta}e^{i\omega(t-\tau)}d\omega.$$

If we put $t = \tau$, then $\mathfrak{D}$ takes the form

$$\mathfrak{D} = \int_{-\infty}^{+\infty} J_{\alpha^{\dagger}B}(\omega)(1 - e^{[\omega - E(f)]/\theta})d\omega. \tag{1-Λ}$$

Notice that

$$|1 - e^{[\omega - E(f)]/\theta}| \leq [|\omega - E(f)|/\theta](1 + e^{\omega/\theta}),$$

we have the estimation

$$|\mathfrak{D}| \leq (1/\theta)\int_{+\infty}^{-\infty} |J_{\alpha^{\dagger}B}(\omega)| \cdot |\omega - E(f)|(1 + e^{\omega/\theta})d\omega,$$

and consequently

$$|\mathfrak{D}| \leq (1/\theta)\left(\int_{-\infty}^{+\infty} J_{\alpha^{\dagger}\alpha}(\omega)(\omega - E(f))^2(1 + e^{\omega/\theta})d\omega\right)^{1/2} \otimes \left(\int_{-\infty}^{+\infty} J_{B^{\dagger}B}(\omega)(1 + e^{\omega/\theta})d\omega\right)^{1/2}. \tag{2-Λ}$$

From the lemma's conditions we have $|B| \leq b_2$, implying

$$\langle B^{\dagger}B + BB^{\dagger} \rangle_H \leq 2b_2^2$$

and

$$\langle B^{\dagger}B + BB^{\dagger} \rangle_H = \int_{-\infty}^{+\infty} J_{B^{\dagger}B}(\omega)(1 + e^{\omega/\theta})d\omega \leq 2b_2^2. \tag{3-Λ}$$

On the other hand, we know that

$$\int_{-\infty}^{+\infty} J_{\alpha^{\dagger}\alpha}(\omega)[\omega - E(f)]^2(1 + e^{\omega/\theta})d\omega = \langle R_f^{\dagger}R_f + R_f R_f^{\dagger} \rangle_H. \tag{4-Λ}$$

Applying the lemma's condition, we have

$$\langle R_f^{\dagger}R_f + R_f R_f^{\dagger} \rangle_H \leq 2\bar{\epsilon}_V,$$

and, substituting (3-Λ), (4-Λ) into (2-Λ), we find finally

$$|\mathfrak{D}| \leq (2/\theta)b_2(\bar{\epsilon}_V)^{1/2},$$

which completes the proof of Lemma 1.

Consider now an example of the calculation of the average

$$\langle a_f^{\dagger}a_f \rangle_H. \tag{17}$$

We mention that, among the averages constructed from Hamiltonian (1) and composed of the product of Fermi operators, only those averages which contain an equal number of creation and annihilation operators are different from zero.

We write down the operators $a_f^{\dagger}, a_f$ expressed in terms of $\alpha_f, \alpha_f^{\dagger}$

$$a_f = u_f\alpha_f - v_f\alpha_{-f}^{\dagger}(L/C) + \hat{\eta}_f, \quad a_f^{\dagger} = u_f\alpha_f^{\dagger} - v_f(L^{\dagger}/C)\alpha_{-f} + \hat{\eta}_f^{\dagger}, \tag{18}$$

$$L = (1/V)\sum_{(f)} \lambda_f a_{-f}a_f, \quad \hat{\eta} = v_f^2(1 - L^{\dagger}L/C^2)a_f. \tag{19}$$

Substituting (18), (19) into (17), we find

$$\langle a_f^\dagger a_f\rangle_H = \langle u_f^2 \alpha_f^\dagger \alpha_f - u_f v_f \alpha_f^\dagger \alpha_{-f}^\dagger (L/C) - u_f \alpha_f^\dagger \eta_f - u_f v_f (L^\dagger/C)\alpha_{-f}\alpha_f + v_f^2 (L/C)\alpha_{-f}\alpha_{-f}^\dagger (L/C) - v_f (L^\dagger/C)\alpha_{-f}\eta_f + \eta_f^\dagger [u_f \alpha_f - v_f \alpha_{-f}^\dagger (L/C) + \eta_f]\rangle_H . \tag{20}$$

To estimate the terms on the right of (20), we take account of (16). Denoting $(LL^\dagger/C^2 - 1) = y$ and taking into account the estimate (16), we see that

$$\langle y^2\rangle_H \le \eta_V^{(1)} \to 0 \quad \text{for } V \to \infty. \tag{21}$$

We note that the terms in (20) which involve the operators $\hat\eta_f$, $\hat\eta_f$, can be estimated by the inequality[9]

$$|\langle \hat\eta_f W\rangle_H| \le \{\langle \hat\eta_f \cdot \hat\eta_f^\dagger\rangle_H \cdot \langle W^\dagger \cdot W\rangle\}^{1/2}$$

and, further, by the estimate (21). Using the asymptotic commutativity, we can commute the operator $\hat\eta_f$ in the terms where it is sandwiched between operators L, α_{fj} and then estimate it following the above-mentioned procedure.

In all cases we construct estimates for these terms, which are majorized by the value $\epsilon'_V \to 0$ for $V \to 0$.

For estimation of the following averages it is convenient to find, with the help of Lemma 1,

$$\langle \alpha_f^\dagger \alpha_{-f}^\dagger L/C\rangle_H, \quad \langle (L^\dagger/C)\alpha_{-f}\alpha_f\rangle_H, \quad \langle \alpha_f^\dagger \alpha_f\rangle_H,$$

$$\langle \alpha_f \alpha_f^\dagger\rangle_H .$$

Now we may estimate the averages $\langle \alpha_f^\dagger \alpha_{-f}^\dagger L/C\rangle_H$. It should be noticed that the operators α_f, $\alpha_f^\dagger$ "satisfy only approximately" the commutation relation of the Fermi statistics when $f \neq f'$. We get

$$|\alpha_f^\dagger \alpha_f^\dagger + \alpha_f^\dagger \alpha_{f'}^\dagger| \le \text{const}/V, \quad |\alpha_{f'}\alpha_f + \alpha_f \alpha_{f'}| \le \text{const}/V.$$

Let us use Lemma 1 and put $B = \alpha_{-f}^\dagger L/C$; then we get the estimate

$$|\langle \alpha_f^\dagger \alpha_{-f}^\dagger (L/C)\rangle_H - \langle \alpha_{-f}^\dagger \alpha_f^\dagger (L/C)\rangle_H e^{-E(f)/\theta}| \le (2/\theta) b_2 (\bar\epsilon)^{1/2} + K_2/V,$$

$$K_2 = \text{const},$$

whence

$$|\langle \alpha_f^\dagger \alpha_{-f}^\dagger L/C\rangle_H| \le (2/\theta) b_2 (\bar\epsilon_V)^{1/2} + K_2/V.$$

In a similar way we get

$$|\langle \alpha_{-f}\alpha_f L^\dagger/C\rangle_H| \le (2/\theta) b_2 (\bar\epsilon_V)^{1/2} + K_2/V.$$

Applying these estimates, we see that the first and fifth terms give the main contribution here, i.e.,

$$u_f^2 \langle \alpha_f^\dagger \alpha_f\rangle_H + V_f^2 \langle (L^\dagger/C)\alpha_{-f}\alpha_f L/C\rangle_H .$$

The second term can be transformed by permuting the operators $L^\dagger$ with α_{-f}, $\alpha_f^\dagger$ and further with the help of (21). Approximately speaking, it can be represented in form

$$V_f^2 \langle \alpha_{-f}\alpha_{-f}^\dagger\rangle_H + \{\text{"small terms" when } V \to \infty\}.$$

Taking into consideration the identity

$$\alpha_f^\dagger \alpha_f + \alpha_f \alpha_f^\dagger - 1 = v_f^2 [(LL^\dagger/C^2) - 1] + \hat G/V,$$

where $\hat G$ is a bounded operator ($|G| \le$ const, as $V \to \infty$), and Lemma 1 ($B = \alpha_f$), we find the estimation for the average $\langle \alpha_f^\dagger \alpha_f\rangle_H$.

At the end of this sequence of steps, we get

$$|\langle a_f^\dagger a_f\rangle_H - [(1 + e^{E_f/\theta})^{-1} + v_f^2 \text{th} E_f/2\theta]| < \xi_V^{(2)},$$

where the expression $\xi_V^{(2)} \to 0$ as $V \to \infty$.

Hence

$$\lim_{V\to\infty} \langle a_f^\dagger a_f\rangle_H = (1 + e^{E_f/\theta})^{-1} + v_f^2 \text{th} E_f/2\theta.$$

Note also that averages involving only the combinations of Fermi operators such as

$$a_{-f}a_f, \ldots, a_n^\dagger a_{-n}^\dagger \tag{22}$$

and constructed from the Hamiltonian $H(1)$ are equal to zero following the selection rules, e.g.,

$$\langle a_{-f}a_f\rangle_H = 0.$$

Therefore, while defining the "quasiaverages" from such products of Fermi operators over the Hamiltonian H, one should complete the operators such as

$$L/C, \; L^\dagger/C. \tag{23}$$

The point is that if we multiply an operator of type (22) by an appropriate operator of type (23), the resulting product is gauge invariant.

We illustrate this by considering the calculation of the quasiaverage

$$\langle a_{-f} a L^\dagger/C\rangle_H \tag{24}$$

Substituting (18) into (24), we obtain

$$\langle a_{-f}a_f L^\dagger/C\rangle_H = \langle u_f^2 \alpha_{-f}\alpha_f L^\dagger/C - u_f v_f \alpha_{-f}\alpha_{-f}^\dagger L L^\dagger/C^2 + u_f \alpha_{-f}\eta_f L^\dagger/C + u_f v_f \alpha_f^\dagger (L/C)\alpha_f L^\dagger/C - v_f^2 \alpha_f^\dagger (L/C)\alpha_{-f}^\dagger L L^\dagger/C^2 + v_f \alpha_f^\dagger (L/C)\hat\eta_f L^\dagger/C + \hat\eta_{-f} u_f \alpha_f L^\dagger/C - \hat\eta_{-f} v_f \alpha_{-f}^\dagger L L^\dagger/C^2 + \hat\eta_{-f}\eta_{+f} L^\dagger/C\rangle_H .$$

Applying the estimates of the preceding example, we see that the second and fourth terms give the main contribution here, i.e.,

$$-u_f v_f \langle \alpha_{-f}\alpha_{-f}^\dagger L L^\dagger/C^2\rangle, \quad u_f v_f \langle \alpha_f^\dagger (L/C)\alpha_f L^\dagger/C\rangle_H .$$

As a result we find

$$|\langle a_{-f}a_f L^\dagger/C\rangle_H - u_f v_f [(1 - e^{E_f/\theta})/(1 + e^{E_f/\theta})]| \le \xi_V^{(1)},$$

where the expression $\xi_V^{(1)} \to 0$ as $V \to \infty$.

Hence the quasiaverage over the Hamiltonian H is defined by

$$\langle a_{-f}a_f\rangle_H = \lim_{(V\to\infty)} \langle a_{-f}a_f L^\dagger/C\rangle_H = u_f v_f [(1 - e^{E_f/\theta})/(1 + e^{E_f/\theta})].$$

Using this procedure, one can calculate not only binary averages, but also averages of more complicated operators.

Thus we see that it is possible to calculate quasiaverages for the Hamiltonian H, without completing the Hamiltonian H with source terms. But here the majorization technique becomes complicated. This circumstance is evinced by the fact that in the given case only the operator $LL^{\dagger}$ turns out to be "approximately" a C-number when $V \to \infty$, but not the operator L. The proposed treatment can also be generalized to more complicated cases of model systems.

ACKNOWLEDGMENTS

The author wishes to thank Professor N. N. Bogolubov, Professor C. N. Yang, and Professor A. S. Goldhaber for valuable discussions. I thank the State University of New York for its support.

*On leave of absence from Steclov Institute of Mathematics, Academy of Sciences of the USSR, Moscow.

†Work supported in part under National Science Foundation Grant No. GP-32998X.

[1]C. N. Yang, Phys. Rev. 168, 1920 (1968); Phys. Rev. Lett. 19, 1312 (1967).
[2]E. H. Lieb and W. Linger, Phys. Rev. 130, 1605 (1963).
[3]N. N. Bogolubov, Jr., Theor. Math. Phys. (Moscow) 4, 3 (1970).
[4]N. N. Bogolubov, Jr., Theor. Math. Phys. (Moscow) 5, 1 (1970).
[5]N. N. Bogolubov, Preprint JINR P-1451, Dubna, 1961.
[6]N. N. Bogolubov, Jr., Preprint JINR P-4184, Dubna, 1968.
[7]V. L. Bonch-Bruevich, Zh. Eksp. Teor. Fiz. 31, 522 (1956) [Sov. Phys. JETP 4, 456 (1957)]; see also V. L. Bonch-Bruevich and S. V. Tyablikov, *The Green's Function Method in Statistical Mechanics* (North-Holland, New York, 1962).
[8]N. N. Bogolubov and S. V. Tyablikov, Dokl. Akad. Nauk SSSR 126, 53 (1959) [Sov. Phys. Doklady 4, 589 (1959)].
[9]The proof of this inequality is in Ref. 5.

A stochastic Gaussian beam*

G. C. Papanicolaou, D. McLaughlin† and R. Burridge
Courant Institute and Department of Mathematics University Heights, New York University, New York, New York 10012
(Received 3 July 1972; revised manuscript received 29 August 1972)

We consider the propagation of a Gaussian beam in a strongly focusing medium with random deviations from uniformity. We compute the intensity and intensity fluctuations on the beam axis and the mean power of the fundamental mode when the random inhomogeneities are weak and the distance between source and observation point is large. We also compute the mean power transferred to each higher mode.

1. INTRODUCTION

Several investigators[1,2] have considered Gaussian beams in homogeneous or inhomogeneous deterministic media. We consider here a stochastic Gaussian beam modeled so that the following conditions hold:

(i) The beam propagates in a strongly focusing axisymmetric medium with random deviations from axial uniformity.
(ii) The random inhomogeneities are weak and the source is far from the observation point.

Our analysis is based on an explicit representation of the field of the beam in terms of a stochastic process which satisfies a stochastic differential equation. We analyze this stochastic process in the limit of weak inhomogeneities and long distances between source and observation point by using the method one of us[3] employed previously to study wave propagation in a slab of random medium.

In Sec. 2 we formulate the problem. In Sec. 3 we analyze the above mentioned stochastic differential equation. Sections 4, 5, 6, and 7 contain the main results which are as follows. The expected value of the beam intensity on its axis remains constant even at large distances from the source, but the fluctuations grow exponentially with distance from the source. The expected value of the power in the fundamental mode, normalized to 1 at the source, decays with distance from the source. Finally we give a general formula for the average power transfered to each higher mode.

The above results can be generalized to nonaxisymmetric (nonorthogonal[1]) beams by using certain group theoretical methods[4] developed by Burridge and Papanicolaou[5] for a random slab problem. We shall present these results elsewhere.

The related problem of a Gaussian beam propagating through a system of lenses with random imperfections has been treated by Steier[6], using the methods of geometrical optics.

2. FORMULATION OF THE PROBLEM

Let $e^{i\omega t}\Psi(x,\tilde{y},\tilde{z})$ be a time harmonic complex-valued scalar field satisfying the reduced wave equation

$$\partial_x^2\Psi + \partial_{\tilde{y}}^2\Psi + \partial_{\tilde{z}}^2\Psi + k^2n^2(x,\tilde{y},\tilde{z})\Psi = 0, \quad i = \sqrt{-1}, \tag{2.1}$$

where $x,\tilde{y},\tilde{z}$ are Cartesian coordinates, $\partial_x, \partial_{\tilde{y}}, \partial_{\tilde{z}}$ denote partial derivatives, k is the free space wavenumber, and n is the index of refraction.

We shall suppose that, for each x, n^2 attains a maximum value of approximately unity on the x axis and we shall restrict attention to wave propagation with large k in the x direction and confined to the neighborhood of the x axis. Under these conditions it is useful to make the so-called parabolic approximation in solving (2.1).

This approximation may be obtained as follows: Write

$$y = k^{1/2}\tilde{y}, \quad z = k^{1/2}\tilde{z}, \quad \Psi(x,\tilde{y},\tilde{z}) = e^{-ikx}\psi(x,y,z). \tag{2.2}$$

Inserting (2.2) into (2.1) yields

$$-2i\partial_x\psi + k^{-1}\partial_x^2\psi + (\partial_y^2 + \partial_z^2)\psi + k[n^2(x,k^{-1/2}y, k^{-1/2}z) - 1]\psi = 0. \tag{2.3}$$

When k is large, we may neglect the term $k^{-1}\partial_x^2\psi$ and consider the initial value problem (with suitable initial condition):

$$-2i\partial_x\psi + \partial_y^2\psi + \partial_z^2\psi + k(n^2 - 1)\psi = 0, \quad x > 0,$$
$$\psi(0,y,z) = \psi_0(y,z) \quad \text{(given)}. \tag{2.4}$$

This is the parabolic approximation. It is also called the forward scattering approximation. Note that (2.4) also governs wave propagation in the negative x direction provided that we multiply Ψ by $e^{-i\omega t}$ instead of $e^{i\omega t}$. We shall use this fact in Sec. 4.

Let us now make some further assumptions about n^2, but first let us expand n^2 as a Taylor series in $\tilde{y},\tilde{z}$ up to quadratic terms:

$$n(x,\tilde{y},\tilde{z}) = a(x) - [b_{11}(x)\tilde{y}^2 + 2b_{12}(x)\tilde{y}\tilde{z} + b_{22}(x)\tilde{z}^2]$$
$$= a(x) - k^{-1}[b_{11}(x)y^2 + 2b_{12}(x)yz + b_{22}z^2], \tag{2.5}$$

where $a(x)$ is approximately 1 and the quadratic form is positive definite. The linear terms in $\tilde{y},\tilde{z}$ have been neglected in (2.5) since we wish n^2 to have a maximum on the x axis.

We shall now restrict (2.5) to be axisymmetric about the x axis and allow $a(x)$ and $b(x)$ to be random functions of x as follows:

$$n^2(x,\tilde{y},\tilde{z}) = a(x) - k^{-1}b(x)(y^2 + z^2), \tag{2.6}$$

with

$$a(x) = 1 + \epsilon\alpha(x), \quad b(x) = b_0 + \epsilon\beta(x). \tag{2.7}$$

Here ϵ is a small parameter and $\alpha(x), \beta(x)$ are stationary stochastic processes with expected value zero.

If the random medium is not strongly focusing, i.e., if the term $-k^{-1}(y^2 + z^2)$ is not present in (2.6), then it is natural to assume $\alpha = \alpha(x,y,z)$, a random field. When the field ψ is confined by the focusing to a narrow cylinder about the x-axis, then our assumption (2.6) is a reasonable one. The problem without focusing has been treated by Klyatskin and Tatarskii.[7] The problem with focusing and $\alpha = \alpha(x,y,z), \beta = \beta(x,y,z)$ has been analyzed by one of the authors,[8] but here we seek more detailed information which is difficult to obtain by the

method presented there. We now proceed with the formulation of the present problem.

Using (2.6) and (2.7), we rewrite (2.4):

$$i\partial_x\psi = \tfrac{1}{2}[(\partial_y^2 + \partial_z^2) - b_0(y^2+z^2)]\psi + \epsilon[\tfrac{1}{2}k\alpha(x) - \tfrac{1}{2}(y^2+z^2)\beta(x)]\psi, \quad \psi(0,y,z) = \psi_0(y,z). \tag{2.8}$$

By redefining dependent and independent variables we may take $b_0 = 1$ as we do in the sequel.

We shall choose ψ_0 to be the fundamental mode of the unperturbed problem and so

$$\psi_0(0,y,z) = (1/\sqrt{\pi})\, e^{-(y^2+z^2)/2}. \tag{2.9}$$

The orthonormal modes $h_{pq}(y,z)$, $p,q = 0,1,2,\cdots$, satisfy the eigenvalue problem

$$\tfrac{1}{2}[\partial_y^2 + \partial_z^2 - (y^2+z^2)]h_{pq}(y,z),$$
$$h_{pq}(y,z) = [\pi 2^p p!\, 2^q q!]^{-1/2} H_p(y) H_q(z) e^{-(y^2+z^2)/2},$$
$$\lambda_{pq} = -(p+q+1), \quad p,q = 0,1,2,\cdots. \tag{2.10}$$

Here $H_p(y)$ denotes the p^{th} Hermite polynomial,

$$H_p(y) = (-1)^p e^{y^2}\left(\frac{d^p}{dy^p} e^{-y^2}\right), \tag{2.11}$$

and we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_{pq}(y,z) h_{p'q'}(y,z)\,dy\,dz = \delta_{pp'}\delta_{qq'}. \tag{2.12}$$

Thus the initial field (2.9) equals $h_{00}(y,z)$, the fundamental mode.

From (2.8) and (2.9) it follows that the solution of (2.8) is

$$\psi(x,y,z) = \frac{1}{\sqrt{\pi}} \exp\left(-\int_0^x \mu(\xi)\,d\xi - \frac{i\epsilon k}{2}\int_0^x \alpha(\xi)\,d\xi\right) \times \exp[-\tfrac{1}{2}i\mu(x)(y^2+z^2)], \tag{2.13}$$

where $\mu(x)$ is a complex-valued random function satisfying the equation

$$\frac{d\mu(x)}{dx} + \mu^2(x) + [1+\epsilon\beta(x)] = 0, \quad x \geq 0, \quad \mu(0) = -i. \tag{2.14}$$

Define $u(x)$ as the complex-valued solution of

$$\frac{d^2u(x)}{dx^2} + [1+\epsilon\beta(x)]u(x) = 0, \quad x \geq 0, \quad u(0) = 1, \quad \frac{du(0)}{dx} = -i. \tag{2.15}$$

Then we have that

$$\mu(x) = \frac{1}{u(x)}\frac{du(x)}{dx}, \tag{2.16}$$

$$\psi(x,y,z) = \frac{1}{\sqrt{\pi}u(x)} \exp\left(\frac{-i\epsilon k}{2}\int_0^x \alpha(\xi)\,d\xi - \frac{i}{2}\mu(x)(y^2+z^2)\right). \tag{2.17}$$

Thus the random field ψ is completely determined by the random function $u(x)$.

Our objective is to determine the statistical characteristics of $u(x)$, hence of ψ, given the statistical characteristics of $\alpha(x)$ and $\beta(x)$. We shall do this in the limit of large x and small ϵ with $\epsilon^2 x$ of order one. In particular we shall obtain the first two moments of the beam intensity $J(x)$ on the beam axis

$$J(x) = |\psi(x,0,0)|^2 = 1/(\pi|u(x)|^2), \tag{2.18}$$

and the expected value of $Q_{00}(x)$, the squared modulus of the amplitude of the fundamental mode

$$Q_{00}(x) = \left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(x,y,z)\frac{e^{-(y^2+z^2)/2}}{\sqrt{\pi}}\,dy\,dz\right|^2, \quad Q_{00}(0) = 1. \tag{2.19}$$

We also study the transfer of power to the higher modes.

Note that $J(x)$ and $Q_{00}(x)$ do not depend on $\alpha(x)$ since it only affects the phase of ψ.

3. THE STATISTICAL PROBLEM FOR $u(x)$

In this section we analyze the stochastic differential equation (2.15). Let $A(x)$ and $B(x)$ be complex-valued random functions and define

$$u(x) = A(x)\cos x + B(x)\sin x, \quad \dot{u}(x) = -A(x)\sin x + B(x)\cos x. \tag{3.1}$$

Here the dot stands for d/dx. On using (3.1) in (2.15) and rearranging the result, we obtain the following system of equations for $A(x)$ and $B(x)$

$$\begin{pmatrix}\dot{A}(x)\\ \dot{B}(x)\end{pmatrix} = \epsilon\begin{pmatrix}\cos x & -\sin x\\ \sin x & \cos x\end{pmatrix}\begin{pmatrix}0 & 0\\ -\beta(x) & 0\end{pmatrix}\begin{pmatrix}\cos x & \sin x\\ -\sin x & \cos x\end{pmatrix}\begin{pmatrix}A(x)\\ B(x)\end{pmatrix}, \quad \begin{pmatrix}A(0)\\ B(0)\end{pmatrix} = \begin{pmatrix}1\\ -i\end{pmatrix}. \tag{3.2}$$

Let $M(x,x')$, $x \geq x'$, denote the fundamental solution matrix of (3.2), i.e., the matrix-valued solution of (3.2) with initial condition $M(x',x') = I$, the identity matrix. Since the matrix multiplying $(A(x),B(x))$ on the right side of (3.2) has trace zero we have

$$\det M(x,x') = 1, \quad x \geq x'. \tag{3.3}$$

Thus $M(x,x')$ is a real 2×2 unimodular matrix-valued process; it belongs in $Sl(2,R)$, the group of all such matrices.

Define the matrices b_1, b_2, b_3 by

$$b_1 = \tfrac{1}{2}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}, \quad b_2 = \tfrac{1}{2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \quad b_3 = \tfrac{1}{2}\begin{pmatrix}0 & -1\\ -1 & 0\end{pmatrix}. \tag{3.4}$$

Since $M \in Sl(2,R)$, we can write[5]

$$M(x,0) = e^{b_1\chi} e^{b_2\theta} e^{b_1\phi}. \tag{3.5}$$

Here $\chi(x)$, $\theta(x)$, $\phi(x)$ are random functions. The parametrization (3.5) is analogous to the Euler-angle para-

metrization of $SU(2)$, the group of 2×2 unitary unimodular matrices, except that here χ, θ, ϕ vary over the ranges

$$0 \le \chi < 4\pi, \quad 0 \le \phi < 2\pi, \quad 0 \le \theta < \infty. \tag{3.6}$$

Now we insert (3.5) into (3.2) and, after some computation, obtain the following system of equations:

$$\begin{aligned}
&\dot{\chi} = \epsilon\beta(x)[1 + \cos(2x + \chi)\coth\theta],\\
&\dot{\theta} = \epsilon\beta(x)\sin(2x + \chi),\\
&\dot{\phi} = -\epsilon\beta(x)\csc\theta\cos(2x + \chi),\\
&\theta(0) = 0, \quad \chi(0) + \phi(0) = 0, \quad \chi(0) - \phi(0) \text{ arbitrary.}
\end{aligned} \tag{3.7}$$

Let us denote expected values by $E\{\ \}$. As in Sec. 2 we assume that

$$E\{\beta(x)\} = 0, \tag{3.8}$$

$$E\{\beta(x)\beta(x')\} = R(x - x'). \tag{3.9}$$

$R(x)$ is the covariance of the process $\beta(x)$. Let us also assume that

$$|\beta(x)| \le 1 \tag{3.10}$$

almost surely. Under these circumstances and a few other additional assumptions on $\beta(x)$ we may use formally a result of R. Z. Hashminskii[9] to obtain the asymptotic behavior of the processes $\chi(x), \theta(x), \phi(x)$ defined by (3.7). We say formally because one condition in that theorem cannot be satisfied namely, the right sides of (3.7) are not bounded as a function of χ, θ, ϕ in the range (3.6). By making additional restrictions on $\beta(x)$ however, the following analysis can be made rigorous.[10]

The above mentioned result is as follows: Let

$$\tau = \epsilon^2 x, \quad \chi^{(\epsilon)}(\tau) = \chi(\tau/\epsilon^2), \quad \theta^{(\epsilon)}(\tau) = \theta(\tau/\epsilon^2), \quad \phi^{(\epsilon)}(\tau) = \phi(\tau/\epsilon^2). \tag{3.11}$$

Then if $f(\chi, \theta, \phi)$ is any bounded smooth function of its arguments, the conditional expectation

$$P^{(\epsilon)}(\tau; \chi, \theta, \phi) = E\{f(\chi^{(\epsilon)}(\tau), \theta^{(\epsilon)}(\tau), \phi^{(\epsilon)}(\tau))\}, \tag{3.12}$$

given $\chi^{(\epsilon)}(0) = \chi, \theta^{(\epsilon)}(0) = \theta, \phi^{(\epsilon)}(0) = \phi$ converges as $\epsilon \to 0$ to $P^{(0)}(\tau; \chi, \theta, \phi)$, which satisfies the diffusion equation

$$\begin{aligned}
\frac{\partial P^{(0)}}{\partial \tau} &= \gamma\left(\frac{\partial^2}{\partial\theta^2} + \coth\theta\,\frac{\partial}{\partial\theta} + \frac{1}{\sinh^2\theta}\,\frac{\partial^2}{\partial\phi^2}\right)P^{(0)}\\
&\quad + \left(\int_0^\infty R(s)\,ds + \gamma\coth^2\theta\right)\frac{\partial^2 P^{(0)}}{\partial\chi^2}\\
&\quad - \tfrac{1}{2}\int_0^\infty R(s)\sin 2s\,ds\ (1 - \csc^2\theta)\,\frac{\partial P^{(0)}}{\partial\chi}.
\end{aligned}$$

$$P^{(0)}(0; \chi, \theta, \phi) = f(\chi, \theta, \phi), \quad \gamma = \tfrac{1}{2}\int_0^\infty R(s)\cos 2s\,ds. \tag{3.13}$$

In the next section we shall see that the quantities of principal interest to us do not depend on χ. This leads to a very significant simplification in (3.13) since the terms involving the χ derivatives can be ignored on the right side of (3.13). The differential operator in the θ, ϕ variables in (3.13) is the Laplace–Beltrami operator in the Lobachevski plane with θ, ϕ as polar coordinates.[11]

This could have been predicted heuristically from the second and third equations in (3.7) since the third may be written in the form $\sinh\theta\dot{\phi} = -\epsilon\beta(x)\cos(2x + \chi)$. Thus, the "radial velocity" $\dot{\theta}$ and the "transverse velocity" $\sinh\theta\dot{\phi}$ are proportional to $\sin(2x + \chi)$ and $-\cos(2x + \chi)$, respectively. Since x is a fast varying variable in the above limit while χ is slowly varying, we would expect these equations to approximate a Brownian motion which is characterized by the fact that all directions of infinitesimal motion are equally likely. The first and second equations of (3.13), on the other hand, would lead us not to expect Brownian motion in the (θ, χ) plane.

The Laplace–Beltrami operator is self-adjoint relative to the volume element

$$\sinh\theta\,d\theta d\phi \tag{3.14}$$

and hence, because $\theta(0) = 0$, it suffices to have the fundamental solution of (3.13) (with the χ derivatives absent) which is initially concentrated at $\theta = 0$. This is given by[3]

$$P(\tau, \theta, \phi) = \frac{e^{-\gamma\tau/4}}{4\sqrt{2}(\pi\gamma\tau)^{3/2}}\int_\theta^\infty \frac{\rho e^{-\rho^2/4\gamma\tau}\,d\rho}{\sqrt{\cosh\rho - \cosh\theta}}, \quad \theta \ge 0, \quad 0 \le \phi < 2\pi. \tag{3.15}$$

This function is the transition probability density, relative to the volume element (3.14), of $\theta^{(\epsilon)}(\tau), \phi^{(\epsilon)}(\tau)$ given $\theta(0) = 0, \phi(0)$ arbitrary, in the limit $\epsilon \to 0, \tau$ fixed. Thus if $f(\theta, \phi)$ is a bounded smooth function of θ and ϕ, we have

$$\lim_{\epsilon\to 0} E\{f(\theta^{(\epsilon)}(\tau), \phi^{(\epsilon)}(\tau))\} = \int_0^\infty \left(\int_0^{2\pi} f(\theta, \phi)\,d\phi\right)P(\tau, \theta)\sinh\theta\,d\theta. \tag{3.16}$$

Here and hereafter $E\{\ \}$ denotes expectation conditional on $\theta(0) = 0$. In (3.16) we have also used the fact that P of (3.16) is independent of ϕ, i.e., $\phi^{(\epsilon)}(\tau)$ is uniformly distributed in $(0, 2\pi)$ in the limit $\epsilon \to 0$, given $\theta(0) = 0$.

We now proceed with the application of (3.15), (3.16) to the beam problem.

4. MEAN INTENSITY OF THE BEAM ON ITS AXIS

First we alter the original formulation (2.8) of the problem for the field $\psi(x, y, z)$ in the following way. According to (2.4) the source of the beam is located on the plane $x = 0$, and $J(x)$ in (2.18) is the intensity of the beam at $(x, 0, 0)$. Thus $J(x)$ is considered a function of the observation point. From (3.1) and the definition of the matrix $M(x, 0)$ we have

$$\begin{pmatrix} u^+(x) \\ \dot{u}^+(x) \end{pmatrix} = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix}\begin{pmatrix} M_{11}(x,0) & M_{12}(x,0) \\ M_{21}(x,0) & M_{22}(x,0) \end{pmatrix}\begin{pmatrix} 1 \\ -i \end{pmatrix}, \tag{4.1}$$

where now we have written u^+ for u to indicate that it corresponds to waves propagating in the positive x direction. We shall later use superscript $^-$ to indicate propagation in the negative x direction.

If we locate the source at the point x and replace $e^{i\omega t}$ by $e^{-i\omega t}$, then the beam points in the negative x direction. We may seek the intensity at $x = 0$ as a function of x, the location of the source. With this point of view we have

$$\begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix} \begin{pmatrix} M_{11}(x,0) & M_{12}(x,0) \\ M_{21}(x,0) & M_{22}(x,0) \end{pmatrix} \begin{pmatrix} u^- & (x) \\ \dot{u}^- & (x) \end{pmatrix}, \tag{4.2}$$

where $u^-(x)$ and $\dot{u}^-(x)$ are defined by (4.2). Thus, $J^-(x) = (1/\pi)|u^-(x)|^{-2}$ is the beam intensity at $(0,0,0)$ considered as a function of the location of the source. Similar remarks hold for $Q_{00}^-(x)$. Hereafter we shall adopt this point of view. While $u^-(x)$, $J^-(x)$, and $Q_{00}^-(x)$ are physical quantities distinct from those denoted by $u(x)$, $J(x)$, and $Q_{00}(x)$ in Sec. 2, $J(x)$ and $Q_{00}(x)$ have the same moments in both cases.

Let us now express $J^-(x)$ in terms of the functions χ, θ, ϕ of (3.5). From (4.2), (3.4), and (3.5) it follows that

$$\begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos(x+\chi/2) & \sin(x+\chi/2) \\ -\sin(x+\chi/2) & \cos(x+\chi/2) \end{pmatrix} \begin{pmatrix} e^{\theta/2} & 0 \\ 0 & e^{-\theta/2} \end{pmatrix} \times \begin{pmatrix} \cos\phi/2 & \sin\phi/2 \\ -\sin\phi/2 & \cos\phi/2 \end{pmatrix} \begin{pmatrix} u^-(x) \\ \dot{u}^-(x) \end{pmatrix}, \tag{4.3}$$

and hence

$$u^-(x) = e^{i[(\chi(x)/2)+x]} \{\cos[\phi(x)/2]\, e^{-\theta(x)/2} + i \sin[\phi(x)/2]\, e^{\theta(x)/2}\}, \tag{4.4}$$

$$\dot{u}^-(x) = e^{i[(\chi(x)/2)+x]} \{\sin[\phi(x)/2]\, e^{-\theta(x)/2} - i \cos[\phi(x)/2]\, e^{\theta(x)/2}\}, \tag{4.5}$$

$$J^-(x) = (1/\pi)\{\cos^2[\phi(x)/2]e^{-\theta(x)} + \sin^2[\phi(x)/2]e^{\theta(x)}\}^{-1}. \tag{4.6}$$

For comparison we note the corresponding form for $J^+(x)$:

$$J^+(x) = (1/\pi)[\cos^2(x+\chi/2)\, e^{\theta(x)} + \sin^2(x+\chi/2)e^{-\theta(x)}]^{-1}.$$

To find the expected value of $J^-(x)$ in the limit $x \to \infty$, $\epsilon \to 0$, $\epsilon^2 x = \tau$ fixed, we use (4.6) in (3.16) and obtain

$$\lim_{\epsilon\to 0} E\left\{J^-\left(\frac{\tau}{\epsilon^2}\right)\right\} = \int_0^\infty \left[\int_0^{2\pi} \frac{1}{\pi} \left(\cos^2\frac{\phi}{2}\, e^{-\theta} + \sin^2\frac{\phi}{2}\, e^{\theta}\right)^{-1} d\phi\right] P(\tau,\theta)\, \sinh\theta\, d\theta. \tag{4.7}$$

Here $P(\tau,\theta)$ is given by (3.15). The angular integral in (4.7) is elementary. We have

$$\int_0^{2\pi} \left(\cos^2\frac{\phi}{2}\, e^{-\theta} + \sin^2\frac{\phi}{2}\, e^{\theta}\right)^{-1} d\phi = 2\pi, \quad \theta \geq 0. \tag{4.8}$$

Thus, since $P(\tau,\theta)$ is a probability density, (4.7) and (4.8) yield

$$\lim_{\epsilon\to 0} E\{J^-(\tau/\epsilon^2)\} = 1/\pi = J(0). \tag{4.9}$$

This is the main result of this section.

The result (4.9) is somewhat surprising. Because the operator on the right side of (2.8) is self-adjoint we have, for any $x \geq 0$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\psi(x,y,z)|^2\, dy\, dz = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{e^{-(y^2+z^2)}}{\pi}\, dy\, dz = 1. \tag{4.10}$$

Thus, in addition to (4.10) which is true for all $x \geq 0$ without taking expected values, the expected value of the intensity at a point on the beam axis tends to the constant $1/\pi$ when the fluctuations are weak and the generator is far away. This is a consequence of the special form (2.6), (2.7) of the strongly focusing index of refraction.

We could, in principle, obtain (4.9) by using the methods of Secs. 3 and 4 in a previous work of one of the authors.[8] This, however, is not a simple matter. Our analysis bypasses these difficulties because it exploits the explicit representation (2.13) of the field ψ. This representation is in turn a consequence of the form (2.6), (2.7) for $n^2(x,y,z)$.

5. FLUCTUATIONS OF THE INTENSITY

In this section we compute the expected value of the square of the difference of the beam intensity $J(x)$ from its expected value

$$\lim_{\epsilon\to 0} E(\{[J^-(\tau/\epsilon^2)] - E[J^-(\tau/\epsilon^2]\}) = \lim_{\epsilon\to 0} E\{[J^-(\tau/\epsilon^2)]^2\} - \pi^{-2}. \tag{5.1}$$

From (4.6), (3.15) and (3.16) we have

$$\lim_{\epsilon\to 0} E\left\{\left[J^-\left(\frac{\tau}{\epsilon^2}\right)\right]^2\right\} = \int_0^\infty \left\{\pi^2 \int_0^{2\pi} \left(\cos^2\frac{\phi}{2}\, e^{-\theta} + \sin^2\frac{\phi}{2}\, e^{\theta}\right)^{-2} d\phi\right\} P(\tau,\theta)\sinh\theta\, d\theta. \tag{5.2}$$

The angular integral is again elementary:

$$\int_0^{2\pi} \left(\cos^2\frac{\phi}{2}\, e^{-\theta} + \sin^2\frac{\phi}{2}\, e^{\theta}\right)^{-2} d\phi = 2\pi\, \cosh\theta, \quad \theta \geq 0. \tag{5.3}$$

Thus, using (5.3) in (5.2) we obtain

$$\begin{aligned}\lim_{\epsilon\to 0} E\left\{\left[J^-\left(\frac{\tau}{\epsilon^2}\right)\right]^2\right\} &= \frac{e^{-\gamma\tau/4}\pi^{-2}}{2\sqrt{2\pi}\,(\gamma\tau)^{3/2}} \times \int_0^\infty \cosh\theta\, \sinh\theta \int_\theta^\infty \frac{\rho e^{-\rho^2/4\gamma\tau}\, d\rho}{\sqrt{\cosh\rho - \cosh\theta}}\, d\theta \\ &= \frac{e^{-\gamma\tau/4}\pi^{-2}}{2\sqrt{2\pi}\,(\gamma\tau)^{3/2}} \int_0^\infty \rho e^{-\rho^2/4\gamma\tau} \int_0^\rho \frac{\cosh\theta\, \sinh\theta\, d\theta}{\sqrt{\cosh\rho - \cosh\theta}}\, d\rho \\ &= \frac{e^{-\gamma\tau/4}\pi^{-2}}{2\sqrt{2\pi}\,(\gamma\tau)^{3/2}} \times \int_0^\infty \rho e^{-\rho^2/4\gamma\tau} \frac{2\sqrt{2}}{3} (1 + 2\cosh\rho) \sinh\frac{\rho}{2}\, d\rho \\ &= \pi^{-2} e^{2\gamma\tau}.\end{aligned} \tag{5.4}$$

We have therefore

$$\lim_{\epsilon\to 0} E\{(J(\tau/\epsilon^2) - E[J^-(\tau/\epsilon^2)])^2\} = \pi^{-2}(e^{2\gamma\tau} - 1). \quad (5.5)$$

This is the main result of this section. Its physical meaning should be clear. Even though the expected value of the intensity $E\{J(x)\}$ is approximately constant when ϵ is small, x is large, and $\tau = \epsilon^2 x$ is of order one, the expected value of its fluctuation grows exponentially in the parameter τ at a rate equal to 2γ where γ is given by (3.13). The fluctuations in the intensity are, of course, zero on the plane of the source and (5.5) is indeed zero when $\tau = 0$.

6. MEAN POWER OF THE FUNDAMENTAL MODE

In this section we compute the expected value of the power $Q_{00}(x)$ in the fundamental mode, given by (2.19), in the familiar limit.

By using (2.17), (2.19) and (4.4), (4.5), we find that

$$Q^-_{00}(x) = 2[\cosh\theta(x) + 1]^{-1}. \quad (6.1)$$

On using (6.1) and (3.15) in (3.16) it follows that

$$\lim_{\epsilon\to 0} E\left\{Q^-_{00}\left(\frac{\tau}{\epsilon^2}\right)\right\} = \frac{e^{-\gamma\tau/4}}{\sqrt{2\pi}\,(\gamma\tau)^{3/2}} \int_0^\infty \sinh\theta[\cosh\theta + 1]^{-1}$$

$$\times \int_\theta^\infty \frac{\rho e^{-\rho^2/4\gamma\tau}\,d\rho}{\sqrt{\cosh\rho - \cosh\theta}}\,d\theta$$

$$= \frac{e^{-\gamma\tau/4}}{\sqrt{2\pi}\,(\gamma\tau)^{3/2}}$$

$$\times \int_0^\infty \rho e^{-\rho^2/4\gamma\tau} \int_0^\rho \frac{\sinh\theta\,d\theta}{(\cosh\theta + 1)\sqrt{\cosh\rho - \cosh\theta}}\,d\rho$$

$$= \frac{4e^{-\gamma\tau/4}}{\sqrt{\pi}} \int_0^\infty \frac{\rho^2 e^{-\rho^2}d\rho}{\cosh(\rho\sqrt{\gamma\tau})}. \quad (6.2)$$

Formula (6.2) is the desired result. In the limit under consideration the expected value of the power in the fundamental mode decays exponentially with τ. For $\tau = 0$ the expected value equals 1 of course. Numerical integration of (6.2) yields the graph shown in Fig. 1. The function (6.2) arises also in a different context[3,12] and its graph was obtained there.

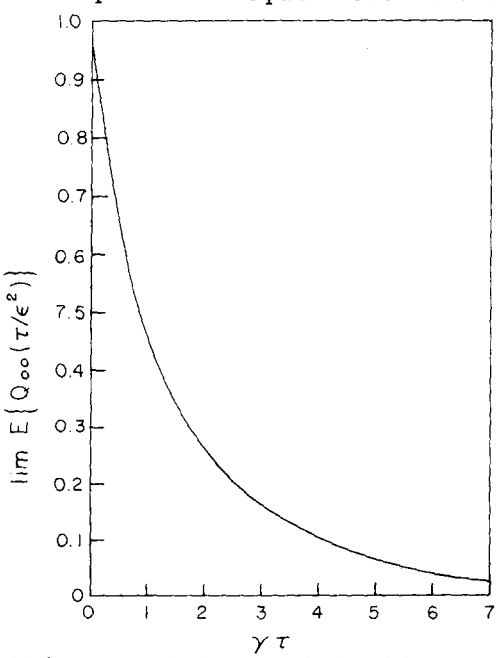

FIG. 1. Here we plot the expected value of the power remaining in the fundamental mode at the point $x = \tau/\epsilon^2$, in the limit $\epsilon \to 0$ versus $\gamma\tau$. See formula (6.2).

Finally, we note the result (6.2) and the graph shows in a very explicit manner how power leaks out of the excited (fundamental) mode into the other (higher) modes.

7. EXPECTED VALUE OF THE POWER IN HIGHER MODES

In previous sections we have been concerned only with the total field on the beam axis and with the power in the fundamental mode. However, one may also consider the higher modes of propagation. We calculate here the expected value of the power in each higher mode in the familiar asymptotic limit. We also derive an expression for the generating function for the modal energy distribution.

We define the $(2p, 2q)$ amplitude of the field by

$$I^-_{2p,2q}(x) = \int_{R^2} h_{2p,2q}(y,z)\psi^-(x,y,z)\,dy\,dz,$$

$$p, q = 0, 1, 2, \ldots, \quad (7.1)$$

where the basis functions $\{h_{2p,2q}\}$ are given by (2.10)–(2.11). The power in the $(2p, 2q)$th mode is then defined as

$$Q^-_{2p,2q}(x) = |I^-_{2p,2q}(x)|^2, \quad p, q = 0, 1, 2, \ldots. \quad (7.2)$$

We need only consider these modes since all others vanish. Utilizing the integral identity for Hermite polynomials,[13]

$$\int_{-\infty}^{\infty} e^{-t^2} H_{2p}(xt)\,dt = \sqrt{\pi}\,\frac{(2p)!}{p!}(x^2 - 1), \quad (7.3)$$

we calculate $Q^-_{2p,2q}(x)$ in terms of u^- and $\dot{u}^-$:

$$Q^-_{2p,2q}(x) = 4^{-p-q+1}\binom{2p}{p}\binom{2q}{q}$$

$$\times\left[\left(|u^-|^2 + |\dot{u}^-|^2 + i(\dot{u}^-(u^-)^* - (\dot{u}^-)^* u^-)\right)^{-1}\right.$$

$$\left.\times\left(\frac{|u^-|^2 + |\dot{u}^-|^2 - i(\dot{u}^-(u^-)^* - (\dot{u}^-)^* u^-)}{|u^-|^2 + |\dot{u}^-|^2 + i(\dot{u}^-(u^-)^* - (\dot{u}^-)^* u^-)}\right)^{p+q}\right]. \quad (7.4)$$

Here we have used (2.16), (2.17) for ψ^- and we have expressed the factorials in terms of binomial coefficients. * stands for complex conjugate. When we use (4.4) and (4.5) in (7.4), we obtain the following expression:

$$Q^-_{2p,2q}(x) = 2^{2(p+q)+1}\binom{2p}{p}\binom{2q}{q}(\cosh\theta + 1)^{-1}\tanh^{2(p+q)}\frac{\theta}{2}. \quad (7.5)$$

While we could calculate the expected value of $Q_{2p,2q}$, we prefer first to sum over the degenerate modes corresponding to the same $r \equiv p + q$. Thus, we fix r and define $\tilde{Q}_r(x)$ by

$$\tilde{Q}_r(x) = \sum_{p+q=r} Q_{2p,2q}(x), \quad r = 0, 1, 2, \cdots. \quad (7.6)$$

When we employ the identities

$$\sum_{p+q=r} \frac{1}{4^r}\binom{2p}{p}\binom{2q}{q} = \sum_{p+q=r} (-1)^r \binom{-1/2}{p}\binom{-1/2}{p} = (-1)^r \binom{-1}{r}, \quad (7.7)$$

we find that

$$\tilde{Q}_r(x) = 2(-1)^r \binom{-1}{r} (\cosh\theta + 1)^{-1} \tanh^{2r}\frac{\theta}{2}. \quad (7.8)$$

Next we use (7.8) and (3.15) in (3.16) and obtain

$$\lim_{\epsilon\to 0} E\left\{\tilde{Q}_r\left(\frac{\tau}{\epsilon^2}\right)\right\} = \frac{1}{2\sqrt{2\pi}}\frac{e^{-\gamma\tau/4}}{(\gamma\tau)^{3/2}} \times \int_\theta^\infty \left[\frac{2(-1)^r \binom{-1}{r} \tanh^{2r}(\theta/2)}{\cosh\theta + 1}\right] \times \int_0^\infty \frac{\rho e^{-\rho^2/4\gamma\tau} d\rho}{\sqrt{\cosh\rho - \cosh\theta}} \sinh\theta\, d\theta = \frac{e^{-\gamma\tau/4}}{2\sqrt{2\pi}(\gamma\tau)^{3/2}} \int_0^\infty \rho e^{-\rho^2/4\gamma\tau} F_r(\rho) d\rho, \quad (7.9)$$

where $F_r(\rho)$ is given by

$$F_r(\rho) = 2(-1)^r \binom{-1}{r} \int_0^\rho \frac{\tanh^{2r}(\theta/2)\sinh\theta\, d\theta}{(\cosh\theta + 1)\sqrt{\cosh\rho - \cosh\theta}}. \quad (7.10)$$

The qualitative behavior of the expected value of the power in the rth mode may be obtained from (7.9). For fixed $r > 0$, the expected value is zero initially ($\tau = 0$) and increases with increasing τ until a maximum is reached, after which it decays. For fixed τ, the expected value of the power in the rth mode decreases with increasing r. (7.9) and (7.10) are the main results of this section. The function $F_r(\rho)$ of (7.10) may be simplified through the change of variable $T = \tanh^2(\theta/2)$,

$$F_r(\rho) = \sqrt{2}\,(-1)^r \binom{-1}{r} \int_0^{\tanh^2(\rho/2)} \times \frac{T^r dT}{\sqrt{1-T}\sqrt{(1-T)\cosh^2(\rho/2) - 1}}. \quad (7.11)$$

In this form the integral can be evaluated by repeated integration by parts using formula (231, 7a).[14]

It is also possible to define a modal generating function for which the integrations are elementary. Specifically we define G by

$$G(x;z) = \sum_{r=0}^\infty \tilde{Q}_r(x) z^r. \quad (7.12)$$

On using (7.9) we find that

$$\lim_{\epsilon\to 0} E\left\{G\left(\frac{\tau}{\epsilon^2};z\right)\right\} = \frac{e^{-\gamma\tau/4}}{2\sqrt{2\pi}(\gamma\tau)^{3/2}} \int_0^\infty \rho e^{-\rho^2/4\gamma\tau} F(\rho;z) d\rho, \quad (7.13)$$

where $F(\rho;z)$ is given by

$$F(\rho;z) = \sum_{r=0}^\infty F_r(\rho) z^r. \quad (7.14)$$

We now insert (7.11) in (7.14) and obtain

$$F(\rho;z) = \sqrt{2}\int_0^{\tanh^2(\rho/2)} \frac{1}{\sqrt{1-T}\sqrt{(1-T)\cosh^2(\rho/2)-1}} \times \left\{\sum_{r=0}^\infty \binom{-1}{r}(-1)^r (zT)^r\, dT\right\} = \sqrt{2}\int_0^{\tanh^2(\rho/2)} \frac{1}{\sqrt{1-T}\sqrt{(1-T)\cosh^2(\rho/2)-1}} \times \frac{dT}{1-zT}. \quad (7.15)$$

This last integral is known; formula (221, 7b) yields[14]

$$F(\rho;z) = \frac{\sqrt{2}}{\sqrt{(1-z)[(1-z)S+1]}} \log\left[1 + 2(1-z)S - 2 \times \sqrt{(1-z)S[(1-z)S+1]}\right], \quad S = \sinh^2(\rho/2). \quad (7.16)$$

Thus, we have obtained a representation of the expected value of the modal generating function as a single integral over ρ, (7.13). Clearly the expected value of the power in the rth mode may be obtained by differentiation as follows:

$$\lim_{\epsilon\to 0} E\left\{\tilde{Q}_r\left(\frac{\tau}{\epsilon^2}\right)\right\} = \frac{e^{-\gamma\tau/4}}{2\sqrt{2\pi}(\gamma\tau)^{3/2}} \int_0^\infty \rho e^{-\rho^2/4\gamma\tau} \frac{1}{r!} \times \left[\frac{\partial^r F(\rho;z)}{\partial z^r}\right]_{z=0} d\rho. \quad (7.17)$$

ACKNOWLEDGMENTS

The authors wish to thank J. B. Keller for reading the manuscript and suggesting several improvements and the referee for bringing to our attention Ref. 6 and for suggesting improvements.

*Research supported by the Office of Naval Research under Contract No. N00014-67-A-0467-0015.

†Present address: Department of Mathematics, Iowa State University, Ames, Iowa 50010.

[1] J. A. Arnaud, Bell Syst. Tech. J. **49**, 2311 (1970).
[2] G. A. Deschamps, "Beam Optics and Complex Rays," URSI Symposium on Electromagnetic Waves, Stresa (1968).
[3] G. C. Papanicolaou, SIAM J. Appl. Math. (Soc. Ind. Appl. Math.) **21**, 13 (1971).
[4] R. Burridge, D. McLaughlin, and G. C. Papanicolaou, (to appear).
[5] R. Burridge and G. C. Papanicolaou, Commun. Pure Appl. Math., (to appear).
[6] W. H. Steier, Bell Syst. Tech. J. **45**, 451 (1966).
[7] B. N. Klyatskin and B. N. Tatarskii, Radiofizika **13**, 1061 (1970).
[8] G. C. Papanicolaou, J. Math. Phys. (submitted).
[9] R. Z. Hashminskii, Theor. Probability Appl. **11**, 390 (1966).
[10] G. C. Papanicolaou and R. Hersh, Indiana Univ. Math. J. **21**, 815 (1972).
[11] C. Caratheodory, *Conformal representation* (Cambridge U.P., Cambridge, 1958).
[12] J. A. Morrison, G. C. Papanicolaou, and J. B. Keller, Commun. Pure Appl. Math. **24**, 473 (1971).
[13] W. Magnus, F. Oberhetinger, and R. P. Soni, *Formulas and theorems for the special function of mathematical physics* (Springer, New York, 1966).
[14] W. Gröbner and N. Hofreiter, *Integraltafel* (Springer, Vienna, 1949).

U and V sectors in the Bronzan–Lee model

A. L. Choudhury
Department of Mathematics, Elizabeth City State University, Elizabeth City, North Carolina 27909
(Received 10 December 1971; revised manuscript received 28 July 1972)

The Lehmann–Symanzik–Zimmermann (LSZ) formalism is used in the Bronzan–Lee model to investigate the scattering processes in the lowest sectors. In this model, there are three heavy recoilless particles U, V, and N interacting with a θ particle. The processes allowed are $U \rightleftarrows V\theta$ and $V \rightleftarrows N\theta$. In U and V sectors of this model all τ functions have been calculated, and relevant scattering and production amplitudes are evaluated. The mass, wavefunction, and vertex function renormalization constants have been found and they agree with the results previously obtained by Bronzan and later also by Liossatos.

I. INTRODUCTION

The Lehmann–Symanzik–Zimmermann[1] approach has been often used as the calculation technique in the Lee model[2] to determine the scattering amplitudes in various sectors. When the Lee model was first proposed, and later on when Pauli and Källén gave a sound mathematical foundation to the model, the usual technique was to solve eigenvalue equations for the corresponding sectors of the model, and the solutions were used to get the scattering amplitudes of the relevant processes. The dispersion-theoretical technique based on the Lehmann–Symanzik–Zimmermann formalism was first used by Treiman and Goldberger[3] to calculate the $N\theta$ scattering. The technique has been used by Amado[4] to calculate the $V\theta$-scattering matrix in the heavy mass approximation for recoilless V and N particles. In these papers the calculations were restricted to one particular process only.

Maxon and Curtis[5] were first to show that the LSZ technique can be used to compute all relevant processes in a particular sector simultaneously. Their method was based on the fact that the S matrix can be expressed in terms of the τ functions, which are the vacuum expectation values of the time-ordered Heisenberg operators corresponding to a particular process. These τ functions satisfy coupled Matthews–Salam equations[6] connecting different processes. In the Lee model, since each state vector can be expanded in the Tamm–Dancoff sense and each sector is a closed subspace of Hilbert space, these equations involve τ functions from a particular sector only. In the lower sectors, the solutions of such equations are not too complicated. Maxon and Curtis[5] used such techniques to obtain all τ functions in the V sector. Scarfone[7] extended the work in the VN sector, where he showed that the VN bound state problem can also be treated along the same lines. In the $V\theta$ sector Maxon[8] reduced the coupled linear differential equations to singular integral equations of the Muskhelishivili-type[9] and obtained an exact solution to get the τ functions. The problem in the $V\theta$ sector was first discussed by Pauli and Källén[2] and later explicitly solved by Amado,[4] using the dispersion relation technique under the assumption that V and N are heavy recoilless particles. Pagnamenta[10] obtained the solution valid off the mass shell, which also yielded the $N\theta\theta'$ production process amplitude. Maxon and Curtis' method of solution gives simultaneously all the τ amplitudes corresponding to all physically relevant processes in the sector. Scarfone[11] used the same technique in the VV sector and obtained the relevant scattering amplitudes. The technique involved requires the solution of similar singular integral equations. Fortunately this can be accomplished.

Bronzan[12] introduced a modification of the Lee model, which subsequently became known as the Bronzan–Lee model. The modification consists of including a heavy recoilless U particle along with the usual V, N, and θ particles. In the Bronzan–Lee model, the processes allowed are $U \rightleftarrows V\theta$ and $V \rightleftarrows N\theta$. This is quite interesting because, in contrast to the usual Lee model, this model has a nontrivial vertex function renormalization associated with it. The problem of renormalization has been solved by Bronzan by using a Wigner–Brillouin perturbation technique. Chen-Cheung[13] also independently used a similar model to study the renormalization problem. The Bronzan–Lee model has also been used by Liossatos[14] to study the compositeness conditions, by establishing the fact that in the limit $Z_V \to 0$, the model becomes equivalent to a model containing only U, N, θ particles interacting via some four-point interactions.

In this paper, we apply the LSZ technique to the Bronzan–Lee model to calculate all the so-called "physical processes" in the lowest sectors of the model. The renormalization constants have been obtained by using the standard techniques. We have calculated all τ functions in the two lowest sectors of the extended model. In Sec. II, we introduce the model and obtain the selection rules for reduction of the Hilbert space into sectors. We also obtain equations of motion of the fields. In Sec. III, we define the τ functions for the V sector, obtain the Matthews–Salam equations and solve them. Finally, we use the τ functions to get the amplitude of the scattering process in that sector. In Sec. IV, the U sector is treated extensively. The Matthews–Salam equations of the corresponding sector are deduced. Mathematically the problem reduces to solving similar equations in the $V\theta$ and VV sectors of the usual Lee model already treated by Maxon[8] and Scarfone.[11] The coupled linear differential equations have been reduced to two singular integral equations. We have solved them following Maxon[8] and Scarfone.[11] The renormalization constants have been deduced by standard techniques. In Sec. V, the τ functions are used to obtain the scattering amplitudes.

II. THE BRONZAN-LEE MODEL

The renormalized Hamiltonian of the Bronzan–Lee model[12,14] for three heavy recoilless particles U, V, and N interacting through a θ particle is assumed to have the following form

$$H = H_0 + H_{\text{int}}, \tag{1}$$

where

$$H_0 = m_U Z_U \psi_U^\dagger \psi_U + m_V Z_V \psi_V^\dagger \psi_V + m_N \psi_N^\dagger \psi_N + \sum_{\mathbf{k}} \omega a_{\mathbf{k}}^\dagger a_{\mathbf{k}} \tag{2}$$

and

$$H_{\text{int}} = g_1 Z_1(\psi_U^+ \psi_V A + A^+ \psi_V^+ \psi_U) + g_2(\psi_V^+ \psi_N A + A^+ \psi_N^+ \psi_V) - \delta m_U Z_U \psi_U^+ \psi_U - \delta m_V Z_V \psi_V^+ \psi_V. \quad (3)$$

The operators ψ_U, ψ_V stand for the renormalized annihilation operators of the U, V particles. ψ_N and $a_{\mathbf{k}}$ are the annihilation operators of the N and θ particles, respectively. g_1 and g_2 are the renormalized coupling constants. δm_U and δm_V are the mass renormalization constants. A is given by the following expression:

$$A = \sum_{\mathbf{k}} X(\omega) a_{\mathbf{k}}, \quad X(\omega) = \frac{f(\omega)}{\sqrt{2\omega}}, \quad \omega = (k^2 + \mu^2)^{1/2}, \quad (4)$$

where $f(\omega)$ is the usual cut off to assure the convergence of integrals. The commutation relations are given by

$$[\psi_U, \psi_U^+] = 1/Z_U, \quad [\psi_V, \psi_V^+] = 1/Z_V, \quad [\psi_N, \psi_N^+] = 1, \quad [a_{\mathbf{k}}, a_{\mathbf{k}}^+] = \delta_{\mathbf{k}\mathbf{k}'}. \quad (5)$$

All other commutators vanish. If we take anticommutation relations for U, V, and N particles, the results of the calculation do not change. If we define

$$Q_1 = Z_U \psi_U^+ \psi_U + Z_V \psi_V^+ \psi_V + \psi_N^+ \psi_N \quad (6a)$$

and

$$Q_2 = Z_U \psi_U^+ \psi_U - \psi_N^+ \psi_N + \sum_{\mathbf{k}} a_{\mathbf{k}}^+ a_{\mathbf{k}}, \quad (6b)$$

we notice that

$$[H, Q_i] = 0. \quad (7)$$

Thus the sectors of this extended model are specified by the eigenvalues of Q_i, i.e., q_i.

The equations of motion turn out to be

$$Z_U\left(i\frac{d}{dt} - m_U^0\right)\psi_U(t) = g_1 Z_1 \sum_{\mathbf{k}} X(\omega)\psi_V(t)a_{\mathbf{k}}(t), \quad (8a)$$

$$Z_V\left(i\frac{d}{dt} - m_V^0\right)\psi_V(t) = g_1 Z_1 \sum_{\mathbf{k}} X(\omega)a_{\mathbf{k}}^+(t)\psi_U(t) + g_2 \sum_{\mathbf{k}} X(\omega)\psi_N(t)a_{\mathbf{k}}(t). \quad (8b)$$

$$\left(i\frac{d}{dt} - m_N\right)\psi_N(t) = g_2 \sum_{\mathbf{k}} X(\omega)a_{\mathbf{k}}^+(t)\psi_V(t), \quad (8c)$$

$$\left(i\frac{d}{dt} - \omega\right)a_{\mathbf{k}}(t) = g_1 Z_1 X(\omega)\psi_V^+(t)\psi_U(t) + g_2 X(\omega)\psi_N^+(t)\psi_V(t), \quad (8d)$$

where we have put $m_U^0 = m_U - \delta m_U$ and $m_V^0 = m_V - \delta m_V$.

III. V SECTOR ($q_1 = 1$, $q_2 = 0$)

Following Maxon and Curtis,[5] we can use the LSZ formalism to calculate the τ functions of this sector. We label τ_α^A as the functions of sector A. In this sector we define

$$\tau_1^V(s) = \langle 0|T(\psi_V(s)\psi_V^+)|0\rangle, \quad (9a)$$

$$\tau_2^V(s,\omega) = X^{-1}(\omega)\langle 0|T(\psi_N(s)a_{\mathbf{k}}(s)\psi_V^+)|0\rangle, \quad (9b)$$

$$\tau_{2R}^V(s,\omega) = X^{-1}(\omega)\langle 0|T(\psi_V(s)a_{\mathbf{k}}^+\psi_N^+)|0\rangle, \quad (9c)$$

$$\tau_3^V(s,\omega,\omega') = X^{-1}(\omega)X^{-1}(\omega')\langle 0|T(\psi_N(s)a_{\mathbf{k}'}(s)a_{\mathbf{k}}^+\psi_N^+)|0\rangle, \quad (9d)$$

where T is the usual time ordering operator and $\psi_{V,N}(s)$ and $a_{\mathbf{k}}(s)$ are the Heisenberg operators.

The resulting Matthews–Salam equations are

$$Z_V\left(i\frac{d}{ds} - m_V^0\right)\tau_1^V(s) = i\delta(s) + g_2\sum_{\mathbf{k}} X^2(\omega)\tau_2^V(s,\omega), \quad (10a)$$

$$\left(i\frac{d}{ds} - m_N - \omega\right)\begin{Bmatrix}\tau_2^V(s,\omega)\\ \tau_{2R}^V(s,\omega)\end{Bmatrix} = g_2\tau_1^V(s), \quad (10b)$$

$$\left(i\frac{d}{ds} - m_N - \omega\right)\tau_3^V(s,\omega,\omega') = i\delta(s)\delta_{\mathbf{k}\mathbf{k}'}X^{-2}(\omega) + g_2\tau_{2R}^V(s,\omega'), \quad (10c)$$

where we have again put $m_V^0 = m_V - \delta m_V$.

Since $\tau_2^V(0,\omega) = \tau_{2R}^V(0,\omega) = 0$, we see that $\tau_2^V(s,\omega) = \tau_{2R}^V(s,\omega)$. Now we take the Fourier transform of τ_α^A defined by the relation

$$\tau_\alpha^A(s,\cdots) = \frac{i}{2\pi}\int_{-\infty}^{+\infty} dW\, e^{-iWs}\,\hat{\tau}_\alpha^A(W,\cdots). \quad (11)$$

We get the Eq. (10a)–(10c):

$$Z_V(W - m_V^0)\hat{\tau}_1^V(W) = 1 + g_2\sum_{\mathbf{k}} X^2(\omega)\hat{\tau}_2^V(W,\omega), \quad (12a)$$

$$(W - m_N - \omega)\hat{\tau}_2^V(W,\omega) = g_2\hat{\tau}_1^V(W), \quad (12b)$$

$$(W - m_N - \omega)\hat{\tau}_3^V(W,\omega,\omega') = X^{-2}(\omega)\delta_{\mathbf{k}\mathbf{k}'} + g_2\hat{\tau}_2^V(W,\omega'). \quad (12c)$$

The final forms of $\hat{\tau}_\alpha^V(W,\cdots)$ are given straightforwardly by

$$\hat{\tau}_1^V(W) = 1/h^+(W - m_N), \quad (13a)$$

$$\hat{\tau}_2^V(W,\omega) = \hat{\tau}_{2R}^V(W,\omega) = g_2/h^+(W - m_N)(W - m_N - \omega + i\epsilon), \quad (13b)$$

$$\hat{\tau}_3^V(W,\omega,\omega') = \frac{X^{-2}(\omega)\delta_{\mathbf{k}\mathbf{k}'}}{(W - m_N - \omega + i\epsilon)} + \frac{g_2^2}{h^+(W - m_N)(W - m_N - \omega + i\epsilon)(W - m_N - \omega' + i\epsilon)} \quad (13c)$$

Here we have defined

$$h^\pm(x) = h(x \pm i\epsilon) = (x - m \pm i\epsilon)\alpha^\pm(x), \quad m = m_V - m_N \quad (14)$$

and

$$\alpha^\pm(x) = \alpha(x \pm i\epsilon) = 1 + \frac{x - m}{\pi}\int_\mu^\infty d\omega \frac{\operatorname{Im} h^+(\omega)}{(\omega - m)^2(\omega - x \mp i\epsilon)}, \quad (14a)$$

where

$$\operatorname{Im} h^\pm(\omega) = \pm\,(g_2^2/4\pi)f^2(\omega)(\omega^2 - \mu^2)^{1/2}\theta(\omega - \mu). \quad (14b)$$

The quantities δm_V and Z_V are obtained, as usual, by requiring that $\hat{\tau}_1^V(W)$ must have a pole for $W = m_V$ and the residue of $\hat{\tau}_1^V(W)$ must be one. We get

$$\delta m_V = -\frac{1}{Z_V\pi}\int_\mu^\infty d\omega \frac{\operatorname{Im} h^+(\omega)}{\omega - m} \quad (15)$$

and

$$Z_V = 1 - \frac{1}{\pi}\int_\mu^\infty d\omega \frac{\operatorname{Im} h^+(\omega)}{(\omega - m)^2}. \quad (16)$$

The scattering matrix for $N\theta \to N\theta'$ is given by

$$S_{\mathbf{k}\mathbf{k}'} = \delta_{\mathbf{k}\mathbf{k}'} + 2\pi i\,\delta(\omega-\omega')X(\omega)X(\omega')(\omega-\omega')^2\hat{\tau}_3^V(W,\omega,\omega')\Big|_{W=m_N+\omega}. \quad (17)$$

Hence

$$S_{\mathbf{k}\mathbf{k}'} = \delta_{\mathbf{k}\mathbf{k}'} - 2\pi i\,\delta(\omega-\omega')X(\omega)X(\omega')[g_2^2/h^+(\omega)]. \quad (18)$$

IV. U SECTOR ($q_1 = 1$, $q_2 = 1$)

In this sector we define the τ functions as follows:

$$\tau_1^U(s) = \langle 0|\,T(\psi_U(s)\psi_U^+)\,|0\rangle, \quad (19a)$$

$$\tau_2^U(s,\omega) = X^{-1}(\omega)\langle 0|\,T(\psi_V(s)a_{\mathbf{k}}(s)\psi_U^+)\,|0\rangle, \quad (19b)$$

$$\tau_{2R}^U(s,\omega) = X^{-1}(\omega)\langle 0|\,T(\psi_U(s)a_{\mathbf{k}}^+\psi_V^+)\,|0\rangle, \quad (19c)$$

$$\tau_3^U(s,\omega,\omega') = X^{-1}(\omega)X^{-1}(\omega')\langle 0|\,T(\psi_N(s)a_{\mathbf{k}'}(s)a_{\mathbf{k}}(s)\psi_U^+)\,|0\rangle, \quad (19d)$$

$$\tau_{3R}^U(s,\omega,\omega') = X^{-1}(\omega)X^{-1}(\omega')\langle 0|\,T(\psi_U(s)a_{\mathbf{k}}^+a_{\mathbf{k}'}^+\psi_N^+)\,|0\rangle, \quad (19e)$$

$$\tau_4^U(s,\omega,\omega') = X^{-1}(\omega)X^{-1}(\omega')\langle 0|\,T(\psi_V(s)a_{\mathbf{k}'}(s)a_{\mathbf{k}}^+\psi_V^+)\,|0\rangle \quad (19f)$$

$$\tau_5^U(s,\omega,\omega',\omega'') = X^{-1}(\omega)X^{-1}(\omega')X^{-1}(\omega'')\times\langle 0|\,T(\psi_N(s)a_{\mathbf{k}'}(s)a_{\mathbf{k}''}(s)a_{\mathbf{k}}^+\psi_V^+)\,|0\rangle \quad (19g)$$

$$\tau_{5R}^U(s,\omega,\omega',\omega'') = X^{-1}(\omega)X^{-1}(\omega')X^{-1}(\omega'')\times\langle 0|\,T(\psi_V(s)a_{\mathbf{k}}(s)a_{\mathbf{k}''}^+a_{\mathbf{k}'}^+\psi_N^+)\,|0\rangle, \quad (19h)$$

$$\tau_6^U(s,\omega,\omega',\omega'',\omega''') = X^{-1}(\omega)X^{-1}(\omega')X^{-1}(\omega'')X^{-1}(\omega''')\times\langle 0|\,T(\psi_N(s)a_{\mathbf{k}''}(s)a_{\mathbf{k}'''}(s)a_{\mathbf{k}}^+a_{\mathbf{k}'}^+\psi_N^+)\,|0\rangle. \quad (19i)$$

The corresponding Matthews–Salam equations are

$$Z_U\left(i\frac{d}{ds}-m_U^0\right)\tau_1^U(s) = i\delta(s) + g_1Z_1\sum_{\mathbf{k}}X^2(\omega)\tau_2^U(s,\omega), \quad (20a)$$

$$Z_V\left(i\frac{d}{ds}-m_V^0-\omega\right)\begin{Bmatrix}\tau_2^U(s,\omega)\\ \tau_{2R}^U(s,\omega)\end{Bmatrix} = g_1Z_1\tau_1^U(s) + g_2\sum_{\mathbf{k}'}X^2(\omega')\begin{Bmatrix}\tau_3^U(s,\omega,\omega')\\ \tau_{3R}^U(s,\omega,\omega')\end{Bmatrix}, \quad (20b)$$

$$\left(i\frac{d}{ds}-m_N-\omega-\omega'\right)\begin{Bmatrix}\tau_3^U(s,\omega,\omega')\\ \tau_{3R}^U(s,\omega,\omega')\end{Bmatrix} = g_2\begin{Bmatrix}\tau_2^U(s,\omega)+\tau_2^U(s,\omega')\\ \tau_{2R}^U(s,\omega)+\tau_{2R}^U(s,\omega')\end{Bmatrix}, \quad (20c)$$

$$Z_V\left(i\frac{d}{ds}-m_V^0-\omega'\right)\tau_4^U(s,\omega,\omega') = i\delta(s)\delta_{\mathbf{k}\mathbf{k}'}X^{-2}(\omega') + g_1Z_1\tau_{2R}^U(s,\omega) + \sum_{\mathbf{k}''}g_2X^2(\omega'')\tau_5^U(s,\omega,\omega'',\omega'), \quad (20d)$$

$$\left(i\frac{d}{ds}-m_N-\omega'-\omega''\right)\tau_5^U(s,\omega,\omega',\omega'') = g_2[\tau_4^U(s,\omega,\omega')+\tau_4^U(s,\omega,\omega'')], \quad (20e)$$

$$\left(i\frac{d}{ds}-m_N-\omega'-\omega''\right)\tau_{5R}^U(s,\omega,\omega',\omega'') = g_2[\tau_4^U(s,\omega',\omega)+\tau_4^U(s,\omega'',\omega)], \quad (20f)$$

$$\left(i\frac{d}{ds}-m_N-\omega'-\omega''\right)\tau_6^U(s,\omega,\omega',\omega'',\omega''') = i\delta(s)\frac{\delta_{\mathbf{k}\mathbf{k}'''}\delta_{\mathbf{k}'\mathbf{k}''}+\delta_{\mathbf{k}\mathbf{k}''}\delta_{\mathbf{k}'\mathbf{k}'''}}{X(\omega)X(\omega')X(\omega'')X(\omega''')} + g_2[\tau_{5R}^U(s,\omega'',\omega,\omega')+\tau_{5R}^U(s,\omega''',\omega,\omega')]. \quad (20g)$$

If we go over to the Fourier transform defined by Eq. (11), we find that Eqs. (20a)–(20g) change to

$$Z_U(W-m_V^0)\hat{\tau}_1^U(W) = 1 + g_1Z_1\sum_{\mathbf{k}}X^2(\omega)\hat{\tau}_2^U(W,\omega) \quad (21a)$$

$$Z_V(W-m_V^0-\omega)\begin{Bmatrix}\hat{\tau}_2^U(W,\omega)\\ \hat{\tau}_{2R}^U(W,\omega)\end{Bmatrix} = g_1Z_1\hat{\tau}_1^U(W) + g_2\sum_{\mathbf{k}'}X^2(\omega)\begin{Bmatrix}\tau_3^U(W,\omega,\omega')\\ \tau_{3R}^U(W,\omega,\omega')\end{Bmatrix} \quad (21b)$$

$$(W-m_N-\omega-\omega')\begin{Bmatrix}\hat{\tau}_3^U(W,\omega,\omega')\\ \hat{\tau}_{3R}^U(W,\omega,\omega')\end{Bmatrix} = g_2\begin{Bmatrix}\tau_2^U(W,\omega)+\hat{\tau}_2^U(W,\omega')\\ \tau_{2R}^U(W,\omega)+\tau_{2R}^U(W,\omega')\end{Bmatrix} \quad (21c)$$

$$Z_V(W-m_V^0-\omega)\hat{\tau}_4^U(W,\omega,\omega') = \delta_{\mathbf{k}\mathbf{k}'}X^{-2}(\omega) + g_1Z_1\tau_{2R}^U(W,\omega) + \sum_{\mathbf{k}''}g_2X^2(\omega'')\hat{\tau}_2^U(W,\omega,\omega'',\omega') \quad (21d)$$

$$(W-m_N-\omega'-\omega'')\hat{\tau}_5^U(W,\omega,\omega',\omega'') = g_2[\hat{\tau}_4^U(W,\omega,\omega')+\hat{\tau}_4^U(W,\omega,\omega'')] \quad (21e)$$

$$(W-m_N-\omega'-\omega'')\hat{\tau}_{5R}^U(W,\omega,\omega',\omega'') = g_2[\hat{\tau}_4^U(W,\omega',\omega)+\hat{\tau}_4^U(W,\omega'',\omega)] \quad (21f)$$

$$(W-m_N-\omega''-\omega''')\hat{\tau}_6^U(W,\omega,\omega',\omega'',\omega''') = \frac{\delta_{\mathbf{k}\mathbf{k}''}\delta_{\mathbf{k}'\mathbf{k}'''}+\delta_{\mathbf{k}\mathbf{k}'''}\delta_{\mathbf{k}'\mathbf{k}''}}{X(\omega)X(\omega')X(\omega'')X(\omega''')} + g_2[\hat{\tau}_{5R}^U(W,\omega'',\omega,\omega')+\hat{\tau}_{5R}^U(W,\omega''',\omega,\omega')]. \quad (21g)$$

From the Eqs. (21b) and (21c) and by defining

$$\Phi^-(x,\omega) = \Phi(x,\omega-i\epsilon) = \hat{\tau}_2^U(x+m_N,\omega)/g_1Z_1\hat{\tau}_1^U(x+m_N). \quad (22)$$

where we have put $W-m_N = x$, we get an integral equation, analytically continued in the complex z plane:

$$h(x-z)\Phi(x,z) = 1 - \frac{1}{\pi}\int_\mu^\infty d\omega\,\frac{\operatorname{Im}h^+(\omega)}{\omega+z-x}\Phi^-(x,\omega). \quad (23)$$

This equation is equivalent to Liossatos'[14] integral equation if we substitute

$$\Phi(x,z) = \psi(x,z)/h(x-z). \tag{23'}$$

Scarfone also obtained a similar equation while determining the τ functions for the VV sector. The solution can easily be obtained either following the prescription of Trueman[15] or that of Maxon.[8] It is

$$\Phi(x,z) = \frac{1}{C(x,m)} \times \left(\frac{1}{x-z-m} + \frac{h^+(x-m)}{z-m}[I_x(x-z) - I_x^+(x-m)]\right), \tag{24a}$$

where

$$C(x,m) = Z_V[1 + h^+(x-m)I_x^+(x-m)]. \tag{24b}$$

The expression $I_x(z)$ is given by

$$I_x(z) = \frac{1}{\pi}\int_\mu^\infty d\omega\, \mathrm{Im}\left(\frac{1}{h^+(\omega)}\right)\frac{1}{\alpha^+(x-\omega)}\frac{1}{(\omega - z)} \tag{24c}$$

and satisfies the following identity:

$$\frac{I_x(z) - I_z^+(x-m)}{x-z-m} = \frac{1}{h(x-z)h(z)} - \frac{x-2m}{h^+(x-m)(x-z-m)(z-m)} - \frac{I_x(x-z) - I_z^+(x-m)}{z-m}. \tag{24d}$$

We can easily verify that this solution is identical with the solution of Liossatos,[14] provided we note that if $m = 0$, $A(z,x)$ and $A(x)$ introduced by him are related to our $I_x(z)$ by the following relations

$$A(z,x) = [(x-z)/z]\, I_x(x-z) - (x/z)\, I_x^+(x) \tag{24e}$$

and

$$A(x) = A(x,x) = -I_x^+(x). \tag{24f}$$

From relation (22), we find

$$\hat\tau_2^U(x+m_N,\omega) = \frac{g_1\hat\tau_1^U(x+m_N)}{h^+(x-\omega)}\Gamma(\omega,x), \tag{25}$$

where we have put, following Liossatos,[14]

$$\Gamma(\omega,x) = Z_1 h^+(x-\omega)\Phi^-(x,\omega). \tag{25a}$$

$\Gamma(\omega,x)$ is the $UV\theta$ off-mass-shell vertex function.

Substituting (25) in (21a) and carrying out the summation over the momentum by converting it into integration, we find

$$\hat\tau_1^U(x+m_N) = [G(x)]^{-1}, \tag{26}$$

where

$$G(x) = Z_U(x + m_N - m_U^0) - g_1^2 Z_1^2 J(x) \tag{27a}$$

and

$$g_2^2 J(x) = \delta m_V + \tfrac{1}{2}x - m - h^+(x-m)/2Z_V C(x,m). \tag{27b}$$

Following Maxon and Curtis,[5] we note that $\hat\tau_1(x+m_N)$ must have a pole at $x = x_0 = m_U - m_N$, where m_U is chosen to be the physical mass of the U particle. This is satisfied when $G(x_0) = 0$. This condition gives us the mass renormalization

$$Z_U\delta m_U = g_1^2 Z_1^2 J(x_0) = \tfrac{1}{2}\gamma^2[x_0 - 2(m - \delta m_V) - h^+(x_0 - m)/Z_V C(x_0,m)], \tag{28a}$$

where we have set

$$\gamma^2 = g_1^2 Z_1^2/g_2^2 \tag{28b}$$

To get the wavefunction renormalization constant Z_U we set $\langle 0|\psi_U|U\rangle = 1$, where ψ_U is the renormalized annihilation operator. This leads to the result that the residue of $\hat\tau_1^U(z+m_N)$ at $x = x_0$ becomes unity, which is again equivalent to the condition $G'(x_0) = dG(x)/dx|_{x=x_0} = 1$. We thus get

$$Z_U = 1 + \frac{\gamma^2}{2} - \frac{\gamma^2}{2Z_V^2}\frac{h^{+\prime}(x_0-m) - [h^+(x_0-m)]^2 I_{x_0}^{+\prime}(x_0-m)}{[1 + h^+(x_0-m)I_{z_0}^+(x_0-m)]^2}. \tag{29}$$

The vertex function renormalization constant Z_1 is determined by requiring the on-mass shell condition

$$\Gamma(x_0 - m, x_0) = 1. \tag{29a}$$

We thus obtain

$$Z_1 = C(x_0,m) = Z_V[1 + h^+(x_0-m)I_{x_0}^+(x_0-m)]. \tag{29b}$$

We can rewrite $G(x)$ in the following form:

$$G(x) = Z_U(x - x_0) - \gamma^2 g_2^2[J(x) - J(x_0)] = D(x)/2Z_V C(x,m), \tag{27a'}$$

where

$$D(x) = Z_V(2Z_U - \gamma^2)(x-x_0)C(x,m) + \gamma^2(h^+(x-m) - h^+(x_0-m)C(x,m)/C(x_0,m)). \tag{27a''}$$

In deducing (27b) we carried out several contour integrations. Thus, summarizing the results, we find

$$\hat\tau_1^U(W) = 1/G(W - m_N) = 2Z_V C(W - m_N, m)/D(W - m_N), \tag{30a}$$

$$\hat\tau_2^U(W,\omega) = [g_1/G(W-m_N)h^+(W - m_N - \omega)]\Gamma(\omega, W - m_N). \tag{30b}$$

$$\hat\tau_3^U(W,\omega,\omega') = g_2[\hat\tau_2^U(W,\omega) + \hat\tau_2^U(W,\omega')]/(W - m_N - \omega - \omega' + i\epsilon). \tag{30c}$$

The other two $\hat\tau^U$ functions $\hat\tau_{2R}^U$ and $\hat\tau_{3R}^U$ become exactly equal to $\hat\tau_2^U$ and $\hat\tau_3^U$, respectively.

From Eqs. (21d), (21e), and remembering $\hat\tau_2^U(W,\omega) = \hat\tau_{2R}^U(W,\omega)$, we get

$$h^+(z-\omega')\hat\tau_4^U(x+m_N,\omega,\omega') = \gamma_{\mathbf{kk}'}X^{-2}(\omega') + g_1 Z_1\hat\tau_2^U(x+m_N,\omega) - g_2^2\sum_{\mathbf{k}''} \times [X^2(\omega'')\hat\tau_4^U(x+m_N,\omega,\omega'')/(\omega'' + \omega' - x - i\epsilon)]. \tag{31}$$

Substituting

$$T(x, \omega' - i\epsilon, \omega) = [1/g_2^2 h^+(x - \omega')]\{h^+(x - \omega)[h^+(x - \omega') \times \hat{\tau}_4^U(x + m_N, \omega, \omega') - \delta_{\mathbf{kk}'} X^{-2}(\omega')]\} \quad (32)$$

and defining

$$L(x, \omega) = (\gamma/g_2) h^+(x - \omega)\hat{\tau}_2^U(x + m_N, \omega), \quad (33)$$

and

$$h(x - z)T(x, z, \omega) = L(x, \omega) + \frac{1}{x - z - \omega} - \frac{1}{\pi}\int_\mu^\infty d\omega'' \frac{\mathrm{Im}\, h^+(\omega'')}{\omega'' + z - x} T(x, \omega'' - i\epsilon, \omega). \quad (34)$$

As usual in arriving at (34), we have converted the summation to an integration and analytically continued the equation into the z plane. We return to our problem, when we put $z = \omega' - i\epsilon$. This is an inhomogeneous Muskhelishivili-type of integral equation. It is perhaps worth mentioning that this integral equation, so far as the mathematical structure is concerned, is similar to the integral equation Scarfone[11] solved in connection with the VV sector of Lee model. Although the structure is the same, the inhomogeneity function $L(x, \omega)$ is different from that of Scarfone. The reciprocal of the complex function $h(z)$ on the real axis takes the role of a V particle Feynman propagator, whereas in the VV case the function is replaced by the VN propagator. Fortunately in both cases the analytical structure is similar: The cut is from μ to infinity, and the behavior of the function at infinity is the same. We can, therefore, solve it using the standard technique adopted by Maxon.[8] We find

$$T(x, z, \omega) = \frac{x - z - m}{h(x - z)(\omega - m)(x - z - \omega)} + \Phi(x, z)\left[L(x, \omega) - \left(\frac{Z_V}{h^+(\omega)}\right) + Z_V h^+(x - \omega)\left(\frac{I_x^+(\omega)}{x - \omega - m} - \frac{I_x^+(x - \omega)}{\omega - m}\right)\right] + \frac{1}{h^+(\omega)(x - z - \omega)} - \frac{1}{(\omega - m)\alpha(x - z)(x - z - \omega)}$$

$$+ \frac{\alpha^+(x - \omega)}{\omega - m}\left(\frac{(x - 2\omega)(x - z - m)}{(x - z - \omega)(z - \omega)} I_x(x - z) - \frac{x - \omega - m}{z - \omega} I_x^+(x - \omega) - \frac{(\omega - m)}{(x - z - \omega)} I_x^+(\omega)\right). \quad (35)$$

The final form of $\hat{\tau}_6^U(x + m_N, \omega, \omega')$ after some algebraic manipulations turns out to be

$$\hat{\tau}_4^U(x + m_N, \omega, \omega') = \frac{\delta_{\mathbf{kk}'} X^{-2}(\omega)}{h^+(x - \omega')} + \frac{g_2^2}{(x - \omega - m)(x - \omega' - m)} \times \left[\frac{1}{x - \omega' - \omega + i\epsilon}\left(\frac{1}{\alpha^+(\omega)\alpha^+(x - \omega)} - (\omega - m)I_x^+(\omega)\right) - \frac{(x - \omega - m)^2}{(\omega - m)(\omega' - \omega)} I_x^+(x - \omega) + \frac{(x - 2\omega)(x - \omega' - m)^2 I_x^+(x - \omega')}{(\omega - m)(x - \omega - \omega')(\omega' - \omega)} + \frac{h^+(x - m)}{(\omega - m)(\omega' - m)} \times \left((\omega - m)I_x^+(\omega) - (x - \omega - m)I_x^+(x - \omega) - \frac{1}{\alpha^+(\omega)\alpha^+(z - \omega)}\right) \times \frac{[(x - \omega' - m)I_x^+(x - \omega') - (x - 2m)I_x^+(x - m)]}{[1 + h^+(x - m)I_x^+(x - m)]}\right] + \frac{g_1^2 \Gamma(\omega, x)}{Z_V G(x) h^+(x - \omega)(x - \omega - m)(x - \omega' - m)} \times (x - \omega - m)\left(1 + \frac{h^+(x - m)}{(\omega' - m)} \times \frac{[(x - \omega' - m)I_x^+(x - \omega') - (x - 2m)I_x^+(x - m)]}{[1 + h^+(x - m)I_x^+(x - m)]}\right). \quad (36)$$

The other τ functions are expressed in terms of $\hat{\tau}_4^U(W, \omega, \omega')$ as follows:

$$\hat{\tau}_5^U(W, \omega, \omega', \omega'') = \frac{g_2[\hat{\tau}_4^U(W, \omega, \omega') + \hat{\tau}_4^U(W, \omega, \omega'')]}{W - m_N - \omega' - \omega'' + i\epsilon}, \quad (37)$$

$$\hat{\tau}_{5R}^U(W, \omega, \omega', \omega'') = \frac{g_2[\hat{\tau}_4^U(W, \omega'', \omega) + \hat{\tau}_4^U(W, \omega', \omega)]}{W - m_N - \omega' - \omega'' + i\epsilon}, \quad (38)$$

and

$$\hat{\tau}_6^U(W, \omega, \omega', \omega'', \omega''') = \frac{\delta_{\mathbf{kk}''}\delta_{\mathbf{k}'\mathbf{k}'''} + \delta_{\mathbf{kk}'''}\delta_{\mathbf{k}'\mathbf{k}''}}{X(\omega)X(\omega')X(\omega'')X(\omega''')(W - m_N - \omega'' - \omega''' - i\epsilon)} + \frac{g_2^2[\hat{\tau}_4^U(W, \omega, \omega'') + \hat{\tau}_4^U(W, \omega, \omega''') + \hat{\tau}_4^U(W, \omega', \omega''') + \hat{\tau}_4^U(W, \omega', \omega''')]}{(W - m_N - \omega - \omega' + i\epsilon)(W - m_N - \omega'' - \omega''' + i\epsilon)}. \quad (39)$$

We have thus completely determined the τ functions of the U sector and, at least in principle, solved all relevant physical problems of this sector. In the next section the above τ functions will be used to find the scattering and production amplitudes of this sector for the Bronzan–Lee model.

V. PRODUCTION AND SCATTERING AMPLITUDES

The $V\theta$ elastic scattering matrix is given by[5]

$$\delta_{\mathbf{kk}'} = \delta_{\mathbf{kk}'} + 2\pi i \delta(\omega - \omega')X(\omega)X(\omega')(\omega - \omega')^2 \hat{\tau}_4^U \times (W, \omega, \omega')\big|_{W = m_V + \omega}. \quad (40)$$

From (36) we can evaluate $S_{\mathbf{kk}'}$:

$$S_{\mathbf{kk}'} = \delta_{\mathbf{kk}'} + 2\pi i \delta(\omega - \omega')X(\omega)X(\omega') \times g_2^2\left[\frac{1 + h^+(\omega)A(\omega)}{h^+(\omega)(1 - h^+(\omega)A(\omega))} - \frac{2\gamma^2}{D(\omega + m)[1 - h^+(\omega)A(\omega)]}\right], \quad (41a)$$

where $D(x)$ is given by (27b) and

$$A(\omega) = -I^{+}_{\omega+m}(\omega). \tag{41b}$$

For the production process $V\theta \to N\theta'\theta''$, the production amplitude is given by

$$P_{\mathbf{k},\mathbf{k}'\mathbf{k}''} = 2\pi i\,\delta(m_V + \omega - m_N - \omega' - \omega'')\frac{X(\omega)X(\omega')X(\omega'')}{\sqrt{2}} \times (m_V + \omega - m_N - \omega' - \omega'')^2\hat{\tau}^U_5 \times (W, \omega, \omega', \omega'')\Big|_{\substack{W = m_V+\omega \\ = m_N+\omega'+\omega''}} \tag{42}$$

Substitution of $\hat{\tau}^U_5(W, \omega, \omega', \omega'')$ from (38) yields

$$P_{\mathbf{k},\,k'k''} = 2\pi i\,\delta(m_V + \omega - m_N - \omega' - \omega'')\frac{X(\omega)X(\omega')X(\omega'')}{\sqrt{2}} \times g_2^3\left(\frac{2}{h^+(\omega')h^+(\omega'')[1 - h^+(\omega)A(\omega)]} - \frac{2\gamma^2 h^+(\omega)}{h^+(\omega')h^+(\omega'')D(\omega+m)[1 - h^+(\omega)A(\omega)]}\right). \tag{43}$$

The scattering amplitude of the reverse process $N\theta\theta' \to V\theta$ can also be easily computed.

The scattering amplitude corresponding to the $N\theta\theta' \to N\theta''\theta'''$ process is given by

$$S_{\mathbf{k}\mathbf{k}',\mathbf{k}''\mathbf{k}'''} = \tfrac{1}{2}(\delta_{\mathbf{k}\mathbf{k}''}\delta_{\mathbf{k}'\mathbf{k}'''} + \delta_{\mathbf{k}\mathbf{k}'''}\delta_{\mathbf{k}'\mathbf{k}''}) + 2\pi i\,\delta(\omega + \omega' - \omega'' - \omega''')X(\omega)X(\omega')X(\omega'')X(\omega''') \times \tfrac{1}{2}(\omega + \omega' - \omega'' - \omega''')^2\hat{\tau}^U_6(W, \omega, \omega', \omega'', \omega''')\Big|_{\substack{W = m_N+\omega+\omega' \\ = m_N+\omega''+\omega'''}} \tag{44}$$

Substituting the value $\hat{\tau}^U_6(W, \omega, \omega', \omega'', \omega''')$ from (39), we get finally

$$\begin{aligned} S_{\mathbf{k}\mathbf{k}',\mathbf{k}''\mathbf{k}'''} &= \tfrac{1}{2}(\delta_{\mathbf{k}\mathbf{k}''}\delta_{\mathbf{k}'\mathbf{k}'''} + \delta_{\mathbf{k}\mathbf{k}'''}\delta_{\mathbf{k}'\mathbf{k}''}) - \pi i\,\delta(\omega + \omega' - \omega'' - \omega''') \\ &\times g_2^2 X(\omega)X(\omega')X(\omega'')X(\omega''') \\ &\left(\frac{\delta_{\mathbf{k}\mathbf{k}''} + \delta_{\mathbf{k}\mathbf{k}'''}}{h^+(\omega')}X^{-2}(\omega) + \frac{\delta_{\mathbf{k}'\mathbf{k}''} + \delta_{\mathbf{k}'\mathbf{k}'''}}{h^+(\omega)}X^{-2}(\omega')\right) \\ &- 2\pi i\,\delta(\omega + \omega' - \omega'' - \omega''')X(\omega)X(\omega')X(\omega'')X(\omega''') \\ &\times \frac{g_2^4 h^+(\omega + \omega' - m)}{h^+(\omega)h^+(\omega')h^+(\omega'')h^+(\omega''')[1 - h^+(\omega + \omega' - m)A(\omega + \omega' - m)]} \\ &\times \left(\frac{\gamma^2 h^+(\omega + \omega' - m) - D(\omega + \omega')}{D(\omega + \omega')}\right). \end{aligned} \tag{45}$$

Thus we see that the first term of expression (45), involving the Kronecker deltas, gives no scattering. The second term, with the first bracket, corresponds to scattering where one θ in the process is unscattered and the other θ particle is scattered in the N source. The last term, involving g_2^4, corresponds to $N\theta\theta'$ scattering where both θ particles interact with N particles.

VI. CONCLUDING REMARKS

We have seen that the LSZ technique can be used straightforwardly to calculate all the τ function in the two lowest sectors of the Bronzan–Lee model. The case is similar to the $V\theta$ and VV sectors of the usual Lee model so far as the mathematical complexity is concerned. The most important difference between the conventional Lee model and the Bronzan–Lee model is the fact that in the latter, vertex function renormalization has to be carried out at every point of the Feynman diagram where a U particle is absorbed or gives rise to a V and θ pair. The comparison between the $V\theta$ and VV sectors of the Lee model and the U sector of the Bronzan–Lee model can be best made if we compare the Feynman diagrams of the physical processes of the sectors in question. We choose the $N\theta\theta' \to N\theta''\theta'''$ process of the U sector of the Bronzan–Lee model and $N\theta\theta'$ scattering of the Lee model. Mathematically identical problems also arise from the $2N\theta\theta' \to 2N\theta''\theta'''$ process of the VV sector. If we look at the processes as shown in Figs. 1–3, we note that for the $N\theta\theta' \to N\theta''\theta'''$ or $2N\theta\theta' \to 2N\theta''\theta'''$ process the second-order diagram is essentially the same in all three cases. But if we go over to the fourth-order diagram all processes are distinctly different. All these processes have one diagram in common, which is equivalent to $N\theta\theta'$ scattering in $V\theta$ sector of ordinary Lee model. Fortunately, although the processes are different, they give rise to the same kind of integral equation and allow us to solve the problem straightforwardly. Looking at Fig. 1, we see that in the Bronzan–Lee model we must have terms proportional to $g_1^2 \cdot g_2^2$. Indeed, such terms exist in Eq. (45). Also we note that the function $h^+(\omega)$ is the reciprocal of the V-particle propagator in contrast to the VN propagator in the VV sector. The function $D(x)$ is more involved and is related to the U particle propagator by Eq. (27a).

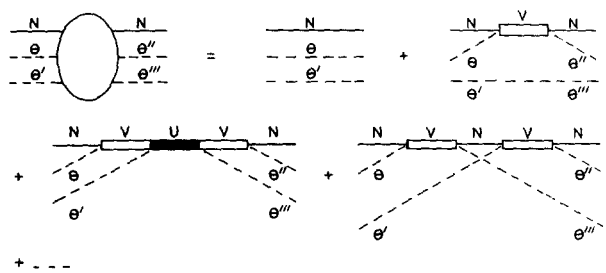

FIG. 1. $N\theta\theta'$ scattering diagram of the Bronzan–Lee model.

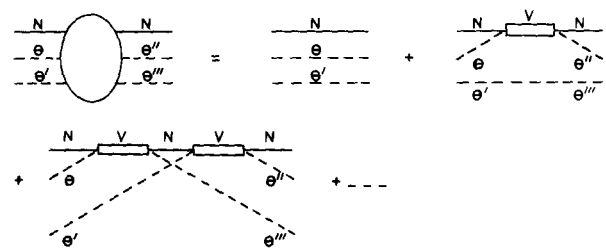

FIG. 2. $N\theta\theta'$ scattering diagram in the normal Lee model.

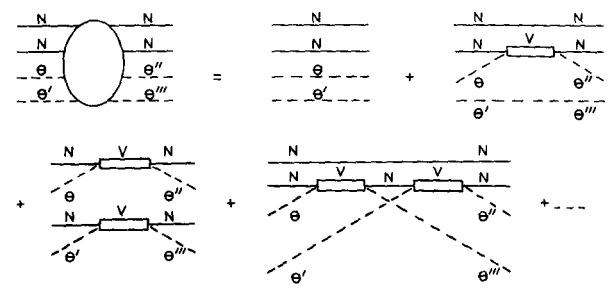

FIG. 3. $2N\theta\theta'$ scattering in the normal Lee model.

The Bronzan–Lee model can also be used to study the possibility of the existence of other bound states in the U sector. In particular, we can investigate the possibility of the existence of complex poles of the expression (26). Recently, Lee and Wick[16] introduced states with negative norm to eliminate divergent quantities from quantum electrodynamics. They suggested certain techniques to carry out contour integrations which come up in the evaluation of the Feynman graphs. They showed how their ideas work in the Lee model. It is possible to investigate how such procedures fit into the Bronzan–Lee model by slightly modifying it. Particularly, in our opinion the assumption that all states which are composed of positive and negative norm states form a complete set should be investigated more closely.

ACKNOWLEDGMENTS

I am grateful for the hospitality of Professor J. W. Straley and Professor H. C. Wilkins, who helped me constantly to get adjusted in the Department of Physics at the University of North Carolina, Chapel Hill. Thanks are due to Professor E. Merzbacher and Professor H. Van Dam for continuous encouragement and helpful discussions. This work was carried out as a Summer Institute activity of the project: "Revitalization of Freshman–Sophomore Physics Programs at Twenty Colleges" at the Department of Physics, University of North Carolina, Chapel Hill. This is a project of the National Laboratory of Higher Education founded by the National Science Foundation.

[1]H. Lehmann, K. Symanzik, and W. Zimmermann, Nuovo Cimento **1**, 205 (1955).
[2]T. D. Lee, Phys. Rev. **95**, 1329 (1954); G. Källén and W. Pauli, K. Dan. Vidensk. Selsk. Mat.-Fys. Medd. 10, No. 7 (1955).
[3]M. L. Goldberger and S. B. Treiman, Phys. Rev. **113**, 1163 (1959).
[4]R. D. Amado, Phys. Rev. **122**, 696 (1961).
[5]M. S. Maxon and R. B. Curtis, Phys. Rev. B **137**, 996 (1965).
[6]P. T. Matthews and A. Salam, Proc. R. Soc. A **221**, 128 (1953).
[7]L. M. Scarfone, J. Math. Phys. **7**, 159 (1966). See also S. Weinberg, Phys. Rev. **102**, 285 (1955) and L. M. Scarfone, Nucl. Phys. **39**, 658 (1962).
[8]M. S. Maxon, Phys. Rev. **149**, 1273 (1966).
[9]N. I. Muskhelishivili, *Singular integral equations* (Noordhoff, Groningen, The Netherlands, 1953) Chap. 5.
[10]A. Pagnamenta, J. Math. Phys. **6**, 955 (1968).
[11]L. M. Scarfone, J. Math. Phys. **9**, 246 (1968); J. Math. Phys. **9**, 977 (1968).
[12]J. B. Bronzan, Phys. Rev. B **139**, 751 (1965).
[13]Fei Shian Chen-Cheung, Phys. Rev. B **152**, 1407 (1966).
[14]P. S. Liossatos, Phys. Rev. B **172**, 1554 (1968).
[15]T. L. Trueman, Phys. Rev. B **137**, 1566 (1965).
[16]T. D. Lee and G. C. Wick, Nucl. Phys. B **7**, 209 (1969); Nucl. Phys. B **10**, 1 (1969).

Neutron transport equations with spin–orbit coupling*

L. M. Tannenwald†
Department of Physics, University of California, Berkeley, California 94720
(Received 28 July 1972)

An *ab initio* derivation of the transport equations is given for fast neutrons traversing a material having spin zero nuclei. The resulting set of equations agrees with the equations of Bell and Goad except for the treatment of the coherent terms. Application of the equations to asymmetric situations is suggested.

1. INTRODUCTION

In this article, an *ab initio* derivation of the transport equations will be given for fast neutrons scattered by a material having spin-0 nuclei, taking account of the spin–orbit interaction.

This problem was first pointed out by Wigner.[1] So far, it seems to have been dealt with only from a heuristic standpoint.[2] In this connection, however, Bell and Goad[2] were able to show some immediate practical effects resulting from the inclusion of spin–orbit coupling.

As in Ref. 2, it will be assumed that the neutrons do not have sufficient energy to cause a significant amount of nuclear excitation; on the other hand, it is assumed that their energy is far greater than any atom's lattice binding energy.

In a general sense, one first needs to give an amplitude description, i.e., a quantum mechanical description of the multiple scattering; this is contained in Sec. 2. Then the transport equations are obtained in Sec. 3 from averaged expressions which are quadratic in the spin amplitudes. The equations thus obtained, Eqs. (3.10) and (3.13)–(3.15), differ from those of Ref. 2 only in so far as that the coherent contributions are properly included.

The formalism used stems mainly from Watson.[3,4]

2. SCATTERING THEORY

The main concern of this section is in amending the multiple scattering formalism so that spin is more explicitly taken into account.

To enable a nonspecialist to follow the gist of the derivation, the more essential formulas of scattering theory are given below, though often without detailed proofs when they are readily available in Goldberger and Watson's book.[5]

The total Hamiltonian will be expressed as

$$H = h + K_i + V \equiv K + V,$$

where the quantities h, K_i, and V are, respectively, the Hamiltonian of the N nuclei, the kinetic energy operator of the neutron and the interaction potential between the neutron and all the nuclei. The eigenvalue problem associated with the unperturbed situation, i.e., $V = 0$, consists of the solutions to

$$h\phi_\gamma = W_\gamma \phi_\gamma \tag{2.1}$$

for the target, and

$$K_i \chi_{p,\nu} = \epsilon_p \chi_{p,\nu} \tag{2.2}$$

for the neutron, with the eigenfunctions obeying appropriate boundary conditions. When the target is in the ground state $\gamma = 0$, ϕ_0 will be a bound state. But when $\gamma \neq 0$, a recoiling nucleus may acquire sufficient energy to require an outgoing (incoming) wave boundary condition on the corresponding wavefunction. The free neutron's spinor function, in the coordinate representation, will be

$$\chi^{(x)}_{p,\nu} = (2\pi)^{-3/2} e^{i[p\cdot x]} u_\nu, \tag{2.3}$$

where p is the initial momentum and $\nu = \pm 1/2$ the spin eigenvalue label. The normalization for $\chi_{p,\nu}$ is

$$\int d^3x \chi^*_{p,\nu} \chi_{q,\mu} = \delta(p - q)\delta_{\nu\mu}. \tag{2.4}$$

The spin eigenfunction u_ν is normalized, therefore, according to

$$(u_\nu, u_\mu) = \delta_{\mu\nu}.$$

The fundamental equation describing the complete problem is

$$\psi^+ = \chi_I + G_0^+ V\psi^+, \tag{2.5}$$

where

$$G_0^\pm \equiv (E \pm i\eta - K)^{-1},$$

E is the total energy of the system, and η is the usual infinitesimal used to assure an outgoing (incoming) wave solution $\psi^+(\psi^-)$, and

$$\begin{aligned}\langle x, z, \nu | \chi_I \rangle &= \chi_{p,\nu}(x)\langle z | \phi_0\rangle \\ &= \chi_{p,\nu}(x)\,\phi_0(z) \\ &= \chi_I(x, z).\end{aligned} \tag{2.6}$$

In Eq. (2.6) and subsequent thereto, x and z will always refer, respectively, to the neutron and (totality of) nuclear coordinates.

Now if

$$V = \sum_{\alpha=1}^{\alpha=N} V_\alpha, \tag{2.7}$$

where here V_α is the neutron–nucleus α interaction, then there exists, according to Watson,[5] a formal solution to Eq. (2.5) given by

$$\psi^+ = \chi_I + \sum_{\alpha=1}^{\alpha=N} G_0^+ t'_\alpha \psi'_\alpha \tag{2.8}$$

with ψ'_α satisfying

$$\psi'_\alpha = \chi_I + \sum_{\beta(\neq\alpha)=1}^{\beta=N} G_0^+ t'_\beta \psi'_\beta, \tag{2.9}$$

and

$$t'_\alpha \equiv V_\alpha + V_\alpha G_\alpha^+ V_\alpha \tag{2.10}$$

with

$$G_\alpha^+ \equiv (E + i\eta - K - V_\alpha)^{-1}. \tag{2.11}$$

The total amplitude of the exact solution to (2.5) consists of an unattenuated amplitude $\chi_I(x,z)$ plus a superposition of amplitudes for scatterings stemming in part from $\chi_I(x,z)$ as well as from rescattered waves. If the attenuation in the target can be significant, then it is much preferable to envision the amplitude ψ^+ made up of an attenuated, coherent[6] wave ψ_c, plus a superposition of amplitudes describing the incoherent scattering.

Again an implicit, though exact, solution exists.[5] If

$$\psi^+ \equiv F\psi_c, \tag{2.12}$$

then

$$F = 1 + P_0 \mathcal{G} \sum_{\alpha=1}^{\alpha=N} t_\alpha F_\alpha \tag{2.13}$$

and

$$F_\alpha = 1 + P_0 \mathcal{G} \sum_{\beta(\neq\alpha)=1}^{\beta=N} t_\beta F_\beta \tag{2.14}$$

with

$$t_\alpha = V_\alpha + V_\alpha \mathcal{G} t_\alpha, \tag{2.15}$$

and

$$\mathcal{G}^{-1} = (G_0^+)^{-1} - \mathcal{U},$$

where $\mathcal{U}$ is the pseudopotential which depends on the state of the target (medium) γ and P_0 is the operator which, in the iterative solution of Eqs. (2.5) and (2.13)–(2.15), prevents the recurrence of the initial state or any other states of the target which have ever been a bra or ket for any t_α. For the present application (where $N \to \infty$)

$$\mathcal{U}(x) = \langle \gamma \mid \mathcal{U} \mid \gamma \rangle = \Big\langle \gamma \Big| \sum_{\alpha=1}^{\alpha=N} t_\alpha F_\alpha \Big| \gamma \Big\rangle. \tag{2.16}$$

The pseudopotential obeys the equation

$$\xi^+_{p,\nu} = \chi_{p,\nu} + g\mathcal{U}\xi^+_{p,\nu}, \tag{2.17}$$

where

$$\langle x, z, \nu \mid \psi_c \rangle = \xi_{p,\nu}(x)\,\phi_0(z) \quad \text{(spin 0)}$$

and g is the unperturbed propagator $(\epsilon_p + i\eta - K_i)^{-1}$.[7]

An essential simplification (the impulse approximation) is made possible by the condition, mentioned in Sec. 1, that the neutrons energy, say 1 MeV, is much greater than the lattice binding energy of the individual target atom. In this regime one may consider the individual collions as practically free, and, thus, it becomes possible to replace t_α by $\mathcal{T}_\alpha$, which denotes the corresponding free particle scattering matrix. The calculations are much simplified because the $\mathcal{T}_\alpha$ depend only on the two coordinates x and z (in a coordinate representation) instead of on all the coordinates. With

$$g_\alpha^+ = (\epsilon + i\eta - K_i - K_\alpha - V_\alpha)^{-1}, \tag{2.18}$$

where ϵ is the energy of the neutron i plus nucleus α,

$$\mathcal{T}_\alpha \equiv V_\alpha + V_\alpha g_\alpha V_\alpha. \tag{2.19}$$

If $\psi_\alpha \equiv F_\alpha \psi_c$, then Eqs. (2.14) and (2.15), using the impulse approximation, take on the following form:

$$\psi^+ = \psi_c^+ + \sum_{\alpha=1}^{\alpha=N} \mathcal{G}\mathcal{T}_\alpha \psi_\alpha, \tag{2.20}$$

$$\psi_\alpha = \psi_c^+ + \sum_{\beta(\neq\alpha)=1}^{\beta=N} \mathcal{G}\mathcal{T}_\beta \psi_\beta. \tag{2.21}$$

In Eqs. (2.20) and (2.21) it has also been assumed that $P_0 = 1$ due to the enormous number of states of the target.

Next, it is necessary to calculate the typical amplitude for a scattered wave from nucleus α, viz., first

$$\hat{\psi}'_\alpha \equiv \langle x, z_\alpha \mid G_0^+ \mathcal{T}_\alpha \psi'_\alpha \rangle, \tag{2.22'}$$

and then

$$\hat{\psi}_\alpha \equiv \langle x, z_\alpha \mid \mathcal{G}\mathcal{T}_\alpha \psi_\alpha \rangle \tag{2.22}$$

as $r = |x - z| \to \infty$.

In this connection, it is convenient to introduce the operators T_α to separate out the momentum conserving property of the $\mathcal{T}_\alpha$. Let the neutron have momenta p and k and the nucleus momenta P and Q, respectively, before and after the collision. Then

$$\langle k, Q \mid \mathcal{T}_\alpha \mid p, P \rangle \equiv \delta(k + Q - p - P)\langle k, Q \mid T_\alpha \mid p, P \rangle. \tag{2.23}$$

The dependence of $\mathcal{T}_\alpha$ and T_α on spin of the neutrons ν is still suppressed. Finally, the positional dependence of $\mathcal{T}_\alpha$ on z_α can be separated out to the following extent:

$$T_\alpha \equiv e^{-ip z_\alpha} T,$$

where

$$\rho \equiv k - p$$

is the neutron's momentum transfer; T is an operator with respect to z through $iP = \nabla_{z_\alpha}$ as given in

$$\langle k, P - \rho \mid T_\alpha \mid p, P \rangle = T.$$

But this z_α dependence exists only if terms of the order of magnitude of the ratio of final nucleon to neutron velocity are considered.[5] Such terms will be ignored here and hence T, hereafter, will be considered not to depend on z_α.

For the unperturbed amplitude ($I = \{p, \gamma = 0\}$)

$$\chi_{I,\nu} = \frac{e^{ip\cdot x}}{(2\pi)^{3/2}}\, u_\nu\, \phi_0(z), \tag{2.24}$$

the scattered amplitude is

$$\langle x, z, \nu' \mid G_0^+ \mathcal{T}_\alpha \chi_{I_\nu} \rangle \equiv (2\pi)^{-3/2} (e^{ik^0 R_\alpha}/R_\alpha)\, e^{iQ_\alpha^0 (z_\alpha - z_\alpha^0)}\, e^{i(p\cdot z_\alpha^0)} f_{\nu'\nu} \phi_0(z_\alpha^0). \tag{2.25}$$

In the last equation $R_\alpha \equiv |x - z_\alpha^0|$, where z_α^0 is the location of the nucleus α before the collision; Q_α^0 and k^0 are, respectively, Q_α and k for $P = 0$. $f_{\nu'\nu''}$ is defined by

$$f_{\nu'\nu''} = -(2\pi)^2 \mu T_{\nu'\nu''} (v_n/\Delta v) F. \tag{2.26}$$

The matrix elements of the proper scattering amplitude consists of only $-(2\pi)^2 \mu T_{\nu\nu'}$, where μ is the reduced mass; $(v_n/\Delta v)_F$ denotes the ratio of the neutron velocity to the relative velocity of neutron and nucleus, both calculated after the collision.

To find $\hat{\psi}'_{\alpha,\nu'}$ stemming from $\psi'_{\alpha,\nu'}$ [see Eqs. (2.22') and (2.25)] requires taking

$$\hat{\psi}'_{\alpha,\nu'} = \sum_{\nu''} \int \langle x, z, \nu' \mid G_0^+ \mathcal{T}_\alpha \chi_{p',\nu''} \rangle\, \psi'_{\alpha,\nu''}(p')\, d^3p', \tag{2.27}$$

where

$$\langle x,z,\nu \mid \psi'_\alpha\rangle \equiv \int \frac{e^{ip\cdot x}}{(2\pi)^{3/2}}\ \psi'_{\alpha,\nu}(p')\,d^3p' \equiv \psi'_{\alpha,\nu}(x,z). \tag{2.28}$$

If it is now assumed, quite realistically, that each scattering is an isolatable event, then (for spin-0 nuclei) the ansatz

$$\psi'_{\beta,\nu}(p') = \frac{1}{(2\pi)^{3/2}} \sum_{\beta(\neq\sigma)=0}^{\beta=N} \psi'_{\beta\sigma,\nu}(p')\,\phi_0(z_1\cdots z^0_\beta\cdots z^0_\sigma\cdots z_N) \tag{2.29}$$

with

$$\psi'_{\beta 0,\nu}(p') = \delta(p-p')u_\nu,$$

which is in harmony with the result of Eq. (2.25), permits one to interpret $\psi_{\beta\sigma,\nu}$ as the amplitude for scattering of neutrons, with spin along (opposite) to the quantization direction, from nucleus σ toward nucleus β. For the ($\langle x,z' \mid$) "bracket" version of Eq. (2.9), using ansatz (2.29) one obtains

$$\psi'_{\alpha,\nu}(x,z) = \chi_{p,\nu}(x)\,\phi_0(z) + \psi'_{\text{scatt}} \tag{2.30}$$

with

$$\psi'_{\text{scatt}} = (2\pi)^{-3/2} \sum_{\nu'} \sum_{\beta(\neq\alpha)=1} \sum_{\sigma(\neq\beta)=1} E'_{\alpha\beta\sigma} f_{\nu\nu'}(\alpha\beta,\beta\sigma) \times \psi'_{\beta\sigma,\nu'}(x)\,\phi_0(z_1\cdots z^0_\beta,\cdots z^0_\sigma,\cdots),$$

where

$$\psi'_{\beta\sigma,\nu}(x) = (2\pi)^{-3/2}\int d^3p'\psi'_{\beta\sigma,\nu}(p'),$$
$$R_{\alpha\beta}E'_{\alpha\beta\sigma} \equiv \exp\{i[k_\beta R_{\alpha\beta} - \rho_{\sigma\beta}\cdot(z_\beta - z^0_\beta)]\}, \tag{2.31}$$
$$f_{\nu\nu'}(\alpha\beta,\beta\sigma) \equiv f_{\nu\nu'}(\hat n_{\alpha\beta},\hat n_{\beta\sigma}),$$

and, in turn,

$$R_{\alpha\beta} \equiv |z^0_\alpha - z^0_\beta|,$$
$$\rho_{\sigma\beta} \equiv k_\beta - k_{\beta\sigma}\hat n_{\beta\sigma}, \quad \beta\neq 0 \tag{2.32}$$
$$\hat n_{\beta\sigma} \equiv (z^0_\beta - z^0_\sigma)/|z^0_\beta - z^0_\sigma|, \quad \hat p \equiv p/|p|,$$

i.e., $\rho_{\sigma\beta}$ is the difference between the neutrons momentum after the collision (near z^0_β) k_β and the corresponding quantity before collision. For $\sigma = 0$, $\rho_{\beta 0} \equiv k_\beta - p$ and $f(\alpha\beta,\beta 0) \equiv f(\hat n_{\alpha\beta},\hat p)$.

A second equation for $\psi'_{\alpha,\nu}(x,z)$ stems from the Fourier transform of Eq. (2.29). It is

$$\psi'_{\alpha,\nu}(x,z) = \chi_{p,\nu}(x)\,\phi_0(z^0_\alpha,z') + (2\pi)^{-3/2} \times \sum_{\beta(\neq\alpha)=1}^{\beta=N} \psi'_{\alpha\beta,\nu}(x)\,\phi_0(z^0_\alpha,\cdots z^0_\beta\cdots z_N), \tag{2.33}$$

where z' denotes all z_γ with $\gamma\neq\alpha$.

Now is is possible to eliminate the ϕ_0, which so far appear with different arguments z and z^0 so that only an equation for $\psi'_{\alpha\beta,\nu}$ remains. This is achieved by first setting in Eqs. (2.30) and (2.31), $x = z^0_\alpha$, and $z_\iota = z^0_\iota$ for all the z'_ι referring to sites where no collision took place. The two resulting expressions for $\psi'_{\alpha,\nu}(z^0_\alpha,z^0)$ are then equated, yielding

$$\psi'_{\alpha\beta,\nu}(z^0_\alpha) = \sum_{\nu'} (E'_{\alpha\beta 0} f_{\nu\nu'}(\alpha\beta,\beta 0)\chi_{p,\nu'}(z^0_\beta) + \sum_{\sigma(\neq\beta)=1}^{\sigma=N} E'_{\alpha\beta\sigma} f_{\nu\nu'}(\alpha\beta,\beta\sigma)\psi'_{\beta\sigma,\nu'}(z^0_\beta)). \tag{2.34'}$$

Equation (2.34') represents a major simplification of the multiple scattering equations because it consists only of linear algebraic equations.

Before proceeding to the transport equations, however, it is preferable, as has already been indicated, to make use of a separation into coherent and incoherent waves,[6] i.e., to work with Eqs. (2.20) and (2.21) and their consequences, instead of Eqs. (2.8) and (2.9) (with t_α replaced by $\mathcal{T}_\alpha$).

Formally, the attendant changes are simple. For a parity conserving interaction, which is the case here, $\mathcal{U}$ has no spin dependence; for a uniform medium as is considered here $\mathcal{U}$ is a complex constant. Hence in going from G^+_0 to $\mathcal{G}$ [see Eq. (2.16)], in the process of changing from a $\psi'_{\alpha,\nu}$ to a $\psi_{\alpha,\nu}$ description, one usually introduces a complex refractive index n

$$n^2 = 1 - (\mathcal{U}/\epsilon_{k_\beta}). \tag{2.35}$$

According to Eq. (2.17), $\xi_{p,\nu}$ may, therefore, be expressed as

$$\xi_{p,\nu}(x) = \xi_{p,\nu}(0)\,e^{i[npD(x)+p(x-D)]}, \tag{2.36}$$

where $\xi_{p,\nu}(0)$ is the spinor amplitude outside the medium, which is normalized such that

$$|\xi_{p,1/2}(0)|^2 + |\xi_{p,-1/2}(0)|^2 = 1,$$

and $D(x)$ is the distance of x into the medium, as measured along a line through x and parallel to $\hat p$.

According to Eqs. (2.20) and (2.21), one change in Eq. (2.34), below, which is the $\psi_{\alpha\beta,\nu}(z^0_\alpha)$ counterpart of Eq. (2.34'), is the replacement of $\chi_{p,\nu}(x)$ [see Eq. (2.3)] by $\xi_{p,\nu}(x)$. A second change involves the replacement of $E'_{\alpha\beta\sigma}$ by $E_{\alpha\beta\sigma}$, which according to Eqs. (2.25), and (2.31) are related to each other by

$$(E_{\alpha\beta\sigma}/E'_{\alpha\beta\sigma}) = \exp[ik_\beta(n-1)R_{\alpha\beta}].$$

Consequently,

$$\psi_{\alpha\beta,\nu}(z^0_\alpha) = \sum_{\nu'} (E_{\alpha\beta 0} f_{\nu\nu'}(\alpha\beta,\beta 0)\,\xi_{p,\nu'}(z^0_\beta) + \sum_{\sigma(\neq\beta)=1}^{\sigma=N} E_{\alpha\beta\sigma} f_{\nu\nu'}(\alpha\beta,\beta\sigma)\psi_{\beta\sigma,\nu'}(z^0_\beta)). \tag{2.34}$$

3. TRANSPORT EQUATIONS

The local neutron density is commonly the dependent variable of the transport equations. Consequently, equations in variables quadratic in $\xi_{p,\nu}$ and $\psi_{\alpha\beta,\nu}$ (and the complex conjugates) will be sought here. The variables used are[8-10]

$$n_c(x) \equiv \sum_\nu |\xi_{p,\nu}(x)|^2, \quad \Pi^{(j)}_c(x) \equiv \sum_{\nu\nu'} \xi^*_{p,\nu}(x)\,\sigma^{(j)}_{\nu\nu'}\,\xi_{p,\nu'}(x),$$
$$n_{\text{inc}}(z'_\alpha,\hat k)\,d\Omega_{\hat k} \equiv \sum'_\beta \sum_\nu |\psi_{\alpha\beta,\nu}(z^0_\alpha)|^2, \tag{3.1}$$
$$\Pi^{(j)}_{\text{inc}}(z^0_\alpha,\hat k)\,d\Omega_{\hat k} \equiv \sum'_\beta \sum_{\nu\nu'} \psi^*_{\alpha\beta,\nu}(z^0_\alpha)\,\sigma^{(j)}_{\nu\nu'}\psi_{\alpha\beta,\nu'}(z^0_\alpha).$$

The "c" and "inc" subscripts are attached, respectively, to quantities associated with coherent and incoherent processes. Thus, n_c is the number of particles per unit volume which have not undergone a (quasi-free) scattering, and $\Pi^{(j)}_c$ is the jth component of their polarization per unit volume. Outside the scatterer, the wave vector

of the neutrons was p. Inside the scatterer consider a point z_α^0, and neutrons arriving there, directed along $\hat{k}$; i.e., arriving in a infinitesimal solid angle $d\Omega_k$, whose axis is $-\hat{k}$. Thus $n_{\rm inc}(z_\alpha^0, \hat{k})$ gives the number of neutrons per unit volume and per unit solid angle (along $-k$) who have undergone at least one nuclear scattering; $\Pi_{\rm inc}^{(j)}(z_\alpha^0, \hat{k})$ is the jth component of the polarization per unit volume and per unit solid angle of these neutrons.

Finally, $\sigma^{(j)}$ in Eq. (3.1) is the jth Pauli spin matrix, which obeys

$$\sum_{\nu'} \sigma_{\nu\nu'}^{(i)} \sigma_{\nu'\nu''}^{(j)} = i\epsilon_{ijk} \sigma_{\nu\nu''}^{(k)} + \delta_{ij}\delta_{\nu\nu''}. \tag{3.2}$$

The total neutron density at x per unit solid angle is, therefore,

$$n(x, \hat{k}) = n_c\, \delta(\hat{k}, \hat{p}) + n_{\rm inc}(x, \hat{k}) \tag{3.3}$$

with

$$(2\pi)^3 n_c = \exp[-D(x)/\lambda],$$

where λ is the mean free path for scattering given by Eq. (3.6) and $\delta(\hat{k}, \hat{p})$ is a directional δ function, such that for any function $F(\hat{k})$ to be used

$$\int \delta(\hat{k}, \hat{p}) F(\hat{k})\, d\Omega_{\hat{k}} = F(\hat{p}).$$

Corresponding to (3.3), there will also be defined the total polarization per unit volume and solid angle according to

$$\Pi^{(j)}(x, \hat{k}) = \Pi_c^{(j)}(x)\, \delta(\hat{k}, \hat{p}) + \Pi_{\rm inc}^{(j)}(x, \hat{k}), \tag{3.4}$$

where, according to Eq. (2.36),

$$\Pi_c^j(x) = \Pi_c^j(0)\, e^{-D(x)/\lambda}.$$

In order to treat the spin–orbit coupling explicitly, the spin matrix elements of the effective scattering amplitude [see Eq. (2.26)] is customarily written[5]

$$f_{\nu\mu}(k, p) = g\, \delta_{\nu\mu} + i h (\sigma \cdot u)_{\nu\mu},$$

where $u \equiv \hat{u} \sin\theta \equiv (\hat{p} \wedge \hat{k})$, and the newly introduced (complex) g and h are given functions of $(\hat{k}, \hat{p})$ and energy ϵ_p. If $h = 0$, then, of course, there is no spin–orbit interaction.

The derivation of the transport equations is now at hand. The first integral equation, the one that will remain even if the spin orbit coupling is omitted, is obtained by computing $n_{\rm inc}\, d\Omega$ from Eq. (2.34) and its complex conjugate version. Thus

$$\begin{aligned} n_{\rm inc}(z_\alpha^0, \hat{k})\, d\Omega_k = \sum_\beta{}' \Gamma_{\alpha\beta} \{ \sum_{\nu\nu'} , [\xi_{p,\nu}^*(z_\beta^0) \\ \times O_{\nu\nu'}^{(n)}(\alpha\beta, \beta O) \xi_{p,\nu'}(z_\beta^0) \\ + \sum_\sigma \psi_{\beta\sigma,\nu}^*(z_\beta^0) O_{\nu\nu'}^{(n)}(\alpha\beta, \beta\sigma) \psi_{\beta\sigma,\nu'}(z_\beta^0)]\} \end{aligned} \tag{3.5}$$

since the $\psi_{\beta\sigma,\nu}$ do not interfere with the $\xi_{p,\nu}$ and (in this energy range) with the $\psi_{\beta\tau,\nu}$. In Eq. (3.5)

$$\Gamma_{\alpha\beta} \equiv |E_{\alpha\beta\sigma}|^2 = e^{-R_{\alpha\beta}/\lambda}/R_{\alpha\beta}$$

where

$$1/2\lambda \equiv {\rm Im}(nk_p). \tag{3.6}$$

Moreover

$$O_{\nu\nu'}^{(n)} \equiv \sum_\mu f_{\nu'\mu} f_{\nu\mu}^* \equiv C_0 \delta_{\nu\nu'} + C_1 (\sigma \cdot \hat{u})_{\nu\nu'} \tag{3.7}$$

with

$$C_0 = |g|^2 + |h|^2 \sin^2\theta$$

and

$$C_1 = i(hg^* - gh^*) \sin\theta.$$

It should be noted that these C are functions of $\hat{k} = \hat{n}_{\alpha\beta}$, $\hat{k}' = \hat{n}_{\beta\sigma'}$, and $|k_\beta|$.

The first term on the right-hand side of Eq. (3.5) is the density of particles moving in the direction $\hat{k}$ as a result of undergoing their first collision at a certain $z_{\beta_j}^0$; similarly the second term gives the contribution to the $n_{\rm inc}(z_\alpha^0, \hat{k})$ as a result of undergoing not their first collision at a certain $z_{\beta_j}^0$, having previously traveled along $\hat{k}'$, i.e., been confined to a $d\Omega_{\hat{k}'}$.

Next Eq. (3.5) will be turned into an integral equation by converting the β and σ sums to integrals. According to (3.1) and (3.7)

$$\begin{aligned} \sum_{\nu\nu'} \sum_\sigma \psi_{\beta\sigma,\nu} O_{\nu\nu'}^{(n)} \psi_{\beta\sigma,\nu'} \\ = \int d\Omega_{\hat{k}'} [C_0(\hat{k}, \hat{k}') n'_{\rm inc} + C_1(\hat{k}, \hat{k}')(\Pi'_{\rm inc} \cdot \hat{u}')], \end{aligned} \tag{3.8}$$

where

$$\hat{u}' = \hat{k}' \wedge \hat{k} / |\hat{k}' \wedge \hat{k}|;$$

primes on $n_{\rm inc}$ and $\Pi_{\rm inc}$ indicate that they are evaluated at a space point $z_\sigma^0 \equiv x' \neq z_\alpha^0 \equiv x$. The sum over β can be written

$$(1/d\Omega_{\hat{k}}) \sum_\beta{}'(\ ) = \int (\ )\rho R^2\, dR, \tag{3.9}$$

where $R \equiv |z_\alpha^0 - z_\beta^0|$ and $\rho(x')$ is the particle density of atoms (nuclei). By using the substitutions (3.8) and (3.9), together with Eqs. (3.3) and (3.4), Eq. (3.5) with $z_\alpha^0 \equiv x$ becomes finally

$$\begin{aligned} n(x, \hat{k}) = n_c(x)\, \delta(\hat{k}, \hat{p}) \\ + \int_0^\infty dR \int d\Omega_{\hat{k}'} e^{-R/\lambda} \rho(R + x) \\ \times [C_0(\hat{k}, \hat{k}') n'(x', \hat{k}') + C_1(\hat{k}, \hat{k}')(u' \cdot \Pi')]. \end{aligned} \tag{3.10}$$

This is the first of four integral equations in n and Π. In Eq. (3.10) the first term is the contribution from particles which have not been scattered. The second gives the number of neutrons scattered into the $(x, \hat{k})$ element. Equation (3.10) is an integral equation in n and Π; hence, to complete the system of equations, $\Pi_{\rm inc}\, d\Omega_k$ will be calculated, analogously to Eq. (3.5).

From Eqs. (3.1) and (2.34) one obtains

$$\begin{aligned} \Pi_{\rm inc}^{(j)}(z_\alpha^0, k)\, d\Omega_{\hat{k}} = \sum_\beta{}' \Gamma_{\alpha\beta} [\sum_{\nu\nu'}{}' \xi_{p,\nu}^* O_{\nu\nu'}^{(\Pi)} \xi_{p,\nu} \\ + \sum_\sigma \psi_{\beta\sigma,\nu}^* O_{\nu\nu'}^{(\Pi)} \psi_{\beta\sigma,\nu'}] \end{aligned} \tag{3.11}$$

with the functions on the right-hand side of Eq. (3.11) depending on the same variables as the corresponding functions in Eq. (3.5). However,

$$O_{\nu m,\nu''}^{(\Pi)} \equiv \sum_{\nu\nu'} f_{\nu\nu m}^* \sigma_{\nu\nu'} f_{\nu'\nu''}$$

or

$$O^{(\Pi)} = [C_0 \sigma - C_2(\hat{u}' \wedge \sigma) + C_3(\hat{u}' \wedge (\hat{u}' \wedge \sigma)) + C_1 \hat{u}' I] \tag{3.12}$$

with

$$C_2 = +(hg^* + gh^*) \sin\theta,$$

$$C_3 = 2|h|^2 \sin^2\theta,$$

I being the unit matrix of two dimensions.

Using, as before, Eqs. (3.1), (3.6), (3.9), (3.11), and (3.12), one finds

$$\Pi^j(x,\hat{k}) = \Pi_c^j(x,\hat{k})\,\delta(\hat{k},\hat{p}) + \int dR\rho(R+x)\,e^{-R/\lambda} \times \int d\Omega_{\hat{k}'}[C_1\hat{u}'n' + C_0\Pi' - C_2(\hat{u}'\wedge\Pi') + C_3(\hat{u}'\wedge\hat{u}'\wedge\Pi')]. \quad (3.13)$$

The three equations (3.13), plus Eq. (3.10), form the complete set of transport equations which were sought.

The first term in Eq. (3.13) is the contribution to the polarization density of neutrons from neutrons which have not been scattered; the second term gives the corresponding contribution of neutrons scattered into the $(x,\hat{k})$ element.

If Eqs. (3.10) and (3.13) are differentiated with respect to x in the direction $\hat{k}$, one obtains

$$\hat{k}\cdot\frac{\partial}{\partial x}\,n(x,\hat{k}) - \hat{k}\cdot\frac{\partial}{\partial x}\,n_c(x)\,\delta(\hat{k},\hat{p}) = -\frac{n(x,\hat{k})}{\lambda} + \int \rho d\Omega_{\hat{k}'}[C_0(\hat{k},\hat{k}')n(x,\hat{k}') + C_1(\hat{k},\hat{k}')\,\hat{u}'\cdot\Pi(x,\hat{k}')] \quad (3.14)$$

$$\hat{k}\cdot\frac{\partial\Pi}{\partial x} - \hat{k}\cdot\frac{\partial\Pi_c}{\partial x}\,\delta(\hat{k},\hat{p}) = -\frac{\Pi}{\lambda} + \int \rho d\Omega_{\hat{k}'}[C_0\Pi + C_1 - C_2(\hat{u}'\wedge\Pi) + C_3(\hat{u}'n\wedge\hat{\mu}'\wedge\Pi)]. \quad (3.15)$$

which is the more common form of writing the transport equations.

With the exception of the coherent terms, which are needed to describe the contribution from neutrons which have not been scattered, Eqs. (3.14) and (3.15) are equivalent to their counterparts of Ref. 2.

Methods for solving the transport equations, for special symmetries, are given in Ref. 2. Because of the presence of left–right asymmetries in scattering involving spin–orbit coupling, it should be more interesting to investigate the influence of spin–orbit coupling on neutron transport in asymmetric situations. This would generally require extensive numerical work, but it ought to give rise to more significant manifestations of spin–orbit coupling.

ACKNOWLEDGMENTS

The great merit of Professor K. M. Watson's suggestions and the beneficial criticism of the manuscript by Professor A. N. Kaufman are gratefully acknowledged.

---

*Research supported by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under Contract number F44620-70-C-0028. This document has been approved for public release and sale; its distribution is unlimited.

†Permanent address: Lockheed Palo Alto Research Laboratory, Palo Alto, California

[1]E. Wigner, Symp. Appl. Math., 9. 92 (1961).
[2]G. I. Bell and W. B. Goad, Nucl. Sci. Eng. **23**, 380 (1965).
[3]K. M. Watson, Phys. Rev. **105**, 1388 (1957).
[4]K. M. Watson, Phys. Rev. **118**, 886 (1960).
[5]M. L. Goldberger and K. M. Watson, *Collision theory* (Wiley, New York, 1965).

[6]By definition, the coherent wave can interfere with primary wave in the medium. The incoherent wave is associated with excitation of the medium.

[7]The mass of the medium is here taken to be infinite.

[8]Our normalization and that of Ref. 5 differ by a factor $(2\pi)^3$.

[9]In order to avoid excessively complicated expressions, an average over the spin states of the individual beam particles is not explicitly shown.

[10]The prime on the sum Σ'_β is restricted to scatterers β in $d\Omega_{\hat{k}}$.

Maxwell's equations and complex Minkowski space

Ezra T. Newman*
Department of Physics, University of Pittsburgh, Pittsburgh, Pennsylvania 15213
(Received 18 July 1972)

It is shown how, from the invariance of Maxwell's equations under the complex Poincaré group, new solutions can be obtained from already known ones. The technique is illustrated with the Coulomb solution. If the same ideas are applied to the linearized Einstein–Maxwell equations a simple classical "derivation" of the Dirac value of the gyromagnetic ratio is obtained.

I. INTRODUCTION

It is the purpose of this note to show that one can obtain new solutions of Maxwell's equations by considering their extension into complex Minkowski space and their consequent invariance under the complex Poincaré group.

In the second section (using the complex 3-vector $\mathbf{E} + i\mathbf{B}$ as the basic field variable), we show how new solutions may be obtained by a complex translation and then illustrate the technique with the Coulomb field. In Sec. III (using the self-dual bivector $F^{\mu\nu} + iF^{*\mu\nu}$ as the basic field variable), we show the invariance of Maxwell's equations under the ten (complex) parameter Poincaré group and how new solutions may be obtained. Finally, in the last section we point out that the ideas used here can be applied to many Lorentz invariant fields. In particular, this applied to both the spin-1 and spin-2, rest-mass zero fields yields a classical "explanation" for the Dirac value of the gyromagnetic ratio.

II. COMPLEX TRANSLATIONS

It is well known[1] that Maxwell's equations can be written in the form

$$\mathrm{curl}\mathbf{W} = i\dot{\mathbf{W}}, \qquad \mathrm{div}\mathbf{W} = 0, \tag{1}$$

where $\mathbf{W} = \mathbf{E} + i\mathbf{B}$. We consider the extension of this equation into complex Minkowski space by allowing the coordinates $z^\mu = (x, y, z, t)$ to take on complex values. (Solutions are now to be holomorphic functions of z^μ.) It is clear that any solution $\mathbf{W}$, of these complexified equations will yield a solution of the real Maxwell equations by restricting the coordinates to the real domain and then taking the real and imaginary parts $\mathbf{W}$ to be $\mathbf{E}$ and $\mathbf{B}$. Furthermore, it is also clear that given a solution $\mathbf{W}(z^\mu)$, then $\mathbf{W}_T(z^\mu) \equiv \mathbf{W}(z^\mu - b^\mu)$ is also a solution. If b^μ is real then the "new" real solution is the "old" one with simply a shift of the origin of real Minkowski space. If b^μ is however complex, the new real solution is fundamentally different from the original.

We illustrate the procedure with the Coulomb field. Since $\mathbf{B} = 0$, $W = (e^3/r^3)(x, y, z)$, where $r = (x^2 + y^2 + z^2)^{1/2}$. By a shift in the origin in the complex z direction we obtain $W_T = (e/r_T^3)(x, y, z - ia)$, where $r_T = [x^2 + y^2 + (z - ia)^2]^{1/2}$.

The $\mathbf{E}$ and $\mathbf{B}$ obtained from the real and imaginary parts of W_T has a multipole expansion of the form, *electric* monopole moment (e), *magnetic* dipole moment (ea), *electric* quadrupole moment (ea^2), *magnetic* octupole, etc. This field is already well known,[2] it being the electromagnetic part of the charged Kerr solution of the Einstein–Maxwell equations.

III. COMPLEX HOMOGENEOUS TRANSFORMATIONS

The invariance of Maxwell's equations under the complex homogeneous Lorentz group is easily demonstrated by writing the covariant form of Maxwell's equations as

$$W^{\mu\nu}{}_{,\nu} = 0, \tag{2}$$

where $W^{\mu\nu} = F^{\mu\nu} + iF^{*\mu\nu}$ and the coordinates are complex. A complex Lorentz transformation ($z'^\mu = \alpha^\mu_\nu z^\nu$, with complex α^μ_ν satisfying $\alpha^\mu_\alpha \alpha^\nu_\beta \eta_{\mu\nu} = \eta_{\alpha\beta}$) acting on $W^{\mu\nu}$ ($W'^{\mu\nu} = \alpha^\mu_\alpha \alpha^\nu_\beta W^{\alpha\beta}$) obviously leaves (3.1) invariant. A new solution of the real Maxwell equations is obtained by restricting the coordinates (*in* $W'^{\mu\nu}$) to real values and then taking the real part of $W'^{\mu\nu}$.

A complex boost applied to the Coulomb solution yields a time dependent solution with the following structure; an *electric* monopole, a *magnetic* dipole with linear time dependence, an *electric* quadrupole with quadratic time dependence, etc.

IV. DISCUSSION

The ideas discussed in the preceding sections appear to be generalizable to most noninteracting Lorentz invariant fields. (When interactions are involved, difficulties arise in the complexification of the potentials.) They certainly apply to the noninteracting rest-mass zero fields.

In particular if the translation in the complex z direction is applied to the monopole solution of the spin-2, rest-mass zero field (i.e., to the linearized version of the Schwarzschild solution of the Einstein equations), a new solution is obtained (the linearized Kerr[3,4] solution). (We emphasize that the complexification is on the Weyl tensor and not on the metric.) This solution has the following source structure: mass monopole moment (m), spin angular momentum (ma), mass quadrupole moment (ma^2), spin octupole moment (ma^3), etc.

[This solution, combined with the complex displaced Coulomb solution (Sec. II) of Maxwell's equations, is the linearized version of the charged Kerr solution[2] of the Einstein–Maxwell equations.]

It is interesting to note that if we begin with the monopole solutions of the spin-1 and spin-2 equations and perform the same complex translation on them, the new solutions have, respectively, a magnetic moment and angular momentum for their sources. The gyromagnetic ratio is independent of the translation parameter and is precisely (e/m), $(g = 2)$ the Dirac value.

It is not at all clear whether this classical "prediction" of the Dirac gyromagnetic ratio (or any of the material of this paper) is nothing but a mathematical exercise or if it has deeper physical meaning. We nevertheless wish to point out that in recent years complex Minkowski space has been playing an increasingly important

role in both particle physics[5] and in relativity (the Penrose theory of twistors).[6,7]

Lastly, we mention that though we have here used the source-free Maxwell equations, the work described could be generalized to include sources.

ACKNOWLEDGMENTS

We would like to thank the Aspen Institute for Physics for its hospitality while this work was being completed, and Dara Newman for her valuable aid.

*This research has been supported by the NSF, Grant No. GP 22789

[1] H. Bateman, *The Mathematical Analysis of Electric and Optical Wave Motion on the Basis of Maxwell's Equations* (Dover, New York, 1955).
[2] E. T. Newman, E. Couch, K. Chinnapared, A. Exton, A. Prakash, and R. Torrence, J. Math. Phys. **6**, 918 (1965).
[3] R. P. Kerr, Phys. Rev. Lett. **11**, 237 (1963).
[4] E. T. Newman and A. I. Janis, J. Math. Phys. **6**, 915 (1965).
[5] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Benjamin, New York, 1964).
[6] R. Penrose, J. Math. Phys. **8**, 345 (1967).
[7] Added in Proof: It has just been pointed out to us that with different motivation, A. Trautman, Proc. Roy. Soc. A 270, 326 (1962), has considered the extension of the Maxwell equations into complex Minkowski space.

Ether flow through a drainhole: A particle model in general relativity

Homer G. Ellis
Department of Mathematics, University of Colorado, Boulder, Colorado 80302
(Received 2 December 1969; revised manuscript received 4 June 1971)

The Schwarzchild manifold of general relativity theory is unsatisfactory as a particle model because the singularity at the origin makes it geodesically incomplete. A coupling of the geometry of space–time to a scalar field ϕ produces in its stead a static, spherically symmetric, geodesically complete, horizonless space–time manifold with a topological hole, termed a drainhole, in its center. The coupling is $R_{\mu\nu}=2\phi_{,\mu}\phi_{,\nu}$; its polarity is reversed from the usual to allow both the negative curvatures found in the drainhole and the completeness of the geodesics. The scalar field satisfies the scalar wave equation $\Box\phi=0$ and has finite total energy whose magnitude, expressed as a length, is comparable to the drainhole radius. On one side of the drainhole the manifold is asymptotic to a Schwarzschild manifold with positive mass parameter m, on the other to a Schwarzschild manifold with negative mass parameter $\bar{m}$, and $-\bar{m}>m$. The two-sided particle thus modeled attracts matter on the one side and, with greater strength, repels it on the other. If m is one proton mass, then $-\bar{m}/m\approx 1+10^{-19}$ or $1+10^{-39}$, according as the drainhole radius is close to 10^{-33}cm or close to 10^{-13} cm; the ratios of total scalar field energy to m in these instances are 10^{19} and 10^{39}. A radially directed vector field which presents itself is interpreted, for purposes of conceptualization, as the velocity of a flowing "substantial ether" whose nonrigid motions manifest themselves as gravitational phenomena. When the ether is at rest, the two-sided particle has no mass on either side, but the drainhole remains open and is able to trap test particles for any finite length of time, then release them without ever accelerating them; some it can trap for all time without accelerating them. This massless, chargeless, spinless particle can, if disturbed, dematerialize into a scalar-field wave propagating at the wave speed characteristic of the space–time manifold.

I. INTRODUCTION

Ever since Schwarzschild presented his spherically symmetric solution of the Einstein vacuum gravitational field equations,[1] it has been a common practice to think of space–time manifolds with "point singularities" as the most appropriate models for mass particles within general relativity theory. Such manifolds, however, are unsatisfactory as models because they are not geodesically complete, failing to provide complete histories for test particles and light rays that encounter the singularities. Einstein and Rosen attempted to do away with the Schwarzschild point singularity by connecting together two Schwarzschild exteriors by a "bridge" at the Schwarzschild horizon.[2] They hoped by thus picturing elementary particles as topological holes in space to explain the atomistic character of matter. They also held out the possibility of explaining quantum phenomena in the same way. The manifold that they constructed, however, not only carried a degenerate metric, which they were prepared to accept, it also suffered the defect of being geodesically incomplete. In cutting away the Schwarzschild interiors they had taken portions of geodesics whose remaining parts they had not subsequently pieced out to completeness.

In more recent times Kruskal has shown,[3] and Fronsdal independently has shown,[4] that the maximal analytic extension of the Schwarzschild manifold has in it a hole, associated with the Schwarzschild horizon, that is topologically but not metrically like the hole in the Einstein-Rosen manifold. This hole Wheeler has termed a "wormhole"[5]; it connects the two Schwarzschild exteriors found in the maximal analytic extension. Some of the geodesics that in the Schwarzschild manifold terminate abruptly at the horizon are, in the maximal extension, completed through the wormhole. However, there are others in the extended manifold that arrive at one of its two point singularities without having exhausted their affine parameters. Hence the maximal analytic extension is geodesically incomplete because of the point singularities.

To get a geodesically complete space–time manifold with a hole in it by which to represent a mass particle, one must find a way to force open the Schwarzschild singularity and there to connect on an additional chunk of space–time, taking care to preserve those features of the original manifold that bring it into agreement with the observable properties of the mass particle. The main object of this writing is to show how that may be done. The hole that replaces the singularity will differ in important respects from the Einsten–Rosen bridge and from the Kruskal–Fronsdal wormhole. At the risk of superadding coinage I shall refer to this hole as a "drainhole." The rationale for this name is that on the space–time manifold containing the hole there is a vector field that can be interpreted as a velocity field for an "ether" draining through the hole. The existence of the hole permits this ether to be conserved in the sense that its streamlines, which are timelike geodesics, never abruptly terminate. It is intriguing that the manifolds that contain one of these drainholes have among them not only reasonable models of mass particles, but also novel models of massless particles with the ability to hold test particles in close orbit for arbitrary lengths of time without accelerating them. These particles, both the massive and the massless, could serve as nuclear glue.

It is clear that these drainhole manifolds, if spherically symmetric, cannot satisfy Einstein's vacuum field equations. Indeed, according to a theorem of Birkhoff, the only spherically symmetric space–time manifold that does so is Schwarzschild's.[6] A "plumber's friend" is needed to open up the Schwarzschild singularity with. The device that will be used is a scalar field. This field ϕ will satisfy the scalar wave equation $\Box\phi = 0$ and will be coupled to the metric of the space–time manifold through the field equations $R_{\mu\nu} = 2\phi_{,\mu}\phi_{,\nu}$, the $R_{\mu\nu}$ being the components of the Ricci tensor field. The polarity of the coupling, which is opposite to the customarily accepted polarity, will be seen to be fixed by the requirement that these field equations

have a static, spherically symmetric, and geodesically complete solution manifold.

It will be convenient to begin with a discussion of a generalized drainhole line element and the geometrical and physical entities that can be associated with it, without at first imposing the field equations (Secs. II–V). After an argument to motivate the choice of field equations (Sec. VI), there will come a description of all their solution manifolds that carry such a line element (Sec. VII), a proof that some of these manifolds are geodesically complete and a description of their geodesics (Sec. VIII), and a final discussion, devoted mainly to the choice of coupling polarity in the field equations and including a proof that every static and spherically symmetric line element can be brought into the adopted form (Sec. IX). The computational framework to be used will be found outlined in the Appendix.

II. THE DRAINHOLE LINE ELEMENT

When referred to a certain nicely adapted coordinate system, the general line element in question takes the spherically symmetric form

$$d\tau^2 = dt^2 - [d\rho - f(\rho)dt]^2 - r^2(\rho)[d\vartheta^2 + (\sin\vartheta)^2 d\varphi^2]$$
$$= dt^2 - [d\rho - f(\rho)dt]^2 - r^2(\rho)d\Omega^2. \quad (1)$$

The function f and the nonnegative function r are to be determined by the field equations. The coordinate ranges are given by

$$-\infty < t < \infty,\ -\infty < \rho < \infty,\ 0 < \vartheta < \pi,\ -\pi < \varphi < \pi, \quad (2)$$

and the additional stipulation that $\rho \in \mathrm{dmn}\, f \cap \mathrm{dmn}\, r - r^{-1}(0)$. The determinant of the metric tensor in this coordinate system is $-[r^2(\rho)\sin\vartheta]^2$; it is, as a result, independent of f. Because $r^{-1}(0)$ is excluded from the range of ρ, the line element is regular for all values of the coordinates.

Once the functions f and r have been specified, the line element may be considered to lie upon a manifold $\mathfrak{M}$ that is almost globally coordinatized by the coordinate system $[t, \rho, \vartheta, \varphi]$, the points without coordinates being those at which $\liminf r(\rho) = 0$, $\lim \vartheta = 0$ or π, or $\lim \varphi = \pm\pi$. Because the metric coefficients in Eq. (1) are independent of t, all translations of $\mathfrak{M}$ along the t coordinate curves are isometries; hence $\partial/\partial t$ is a Killing vector field. Inasmuch as $|\partial/\partial t|^2 = 1 - f^2$, $\partial/\partial t$ is timelike, null, or spacelike according as $f^2 < 1$, $f^2 = 1$, or $f^2 > 1$. Consequently, those regions of $\mathfrak{M}$ where $f^2 < 1$ are stationary. Because $2f(\rho)d\rho dt$ is the only cross term in $\mathfrak{M}$'s line element, $\partial/\partial t$ is not everywhere orthogonal to the hypersurfaces of constant t unless $f = 0$, in which event $\mathfrak{M}$ is static. Actually, $\mathfrak{M}$ is static wherever $f^2 < 1$. This is established in Sec. V, where it is shown that $\partial/\partial t$ is orthogonal to other hypersurfaces.

Let Σ_t denote the cross section of $\mathfrak{M}$ on which the time coordinate has the constant value t.[7] Σ_t is spacelike and inherits from $\mathfrak{M}$ the Riemannian line element given by

$$d\sigma^2 = d\rho^2 + r^2(\rho)d\Omega^2. \quad (3)$$

If it were the case that $r(\rho) = \rho$, then this would be the line element of Euclidean 3-space E^3, cast in polar coordinates ρ, ϑ, and φ. In the general case Σ_t may be thought of as a warped portion of E^3. The warping, caused by deviations of $r(\rho)$ from the Euclidean value ρ, does not destroy the spherical symmetry. The cross section $S_{t,\rho}$ on which the radial coordinate has the constant value ρ is simply a geometrical 2-sphere of radius $r(\rho)$. If $r(\rho)$ has a positive minimum value, then Σ_t has a central hole of that radius, it being the radius of the smallest such 2-sphere $S_{t,\rho}$ in Σ_t.

A case that will arise later has $r(\rho) = (\rho^2 + n^2)^{1/2}$, where n is a positive constant and is the radius of the hole, a particular instance of the drainhole. In this case the equatorial cross section of Σ_t, typical of all great-circle cross sections of Σ_t, may be pictured as in Fig. 1. It is isometrically embeddable in E^3 as $\{[x,y,z] | (x^2 + y^2)^{1/2} = n\cosh(z/n)\}$, a catenoid. Σ_t itself is congruent to $\{[x,y,z,w] | (x^2 + y^2 + z^2)^{1/2} = n\cosh(w/n)\}$ in E^4. Σ_t is asymptotic to E^3, in a sense that can be made precise, both as $\rho \to \infty$ and as $\rho \to -\infty$. This is primarily because, in each instance, $\lim[r(\rho)/|\rho|] = 1$.

III. THE ETHER FLOW

The vector field u on the manifold $\mathfrak{M}$, defined by

$$u = \frac{\partial}{\partial t} + f(\rho)\frac{\partial}{\partial \rho}, \quad (4)$$

has many interesting properties. To begin with, it is everywhere timelike, of unit length, and orthogonal to a cross section Σ_t. Thus it may serve as the timelike vector field in an orthonormal frame system whose spacelike vector fields are tangent to these hypersurfaces Σ_t. One such frame system is $\{e_\mu\}$ defined as follows:

$$e_0 = u = \frac{\partial}{\partial t} + f(\rho)\frac{\partial}{\partial \rho},\ e_1 = \frac{\partial}{\partial \rho},$$
$$e_2 = \frac{1}{r(\rho)}\frac{\partial}{\partial \vartheta},\ e_3 = \frac{1}{r(\rho)\sin\vartheta}\frac{\partial}{\partial \varphi}. \quad (5)$$

The system coframe $\{\omega^\mu\}$ dual to $\{e_\mu\}$ is given by

$$\omega^0 = dt,\ \omega^1 = d\rho - f(\rho)dt,$$
$$\omega^2 = r(\rho)d\vartheta,\ \omega^3 = r(\rho)(\sin\vartheta)d\varphi. \quad (6)$$

Determining the unique torsion-free covariant differentiation $\mathbf{d}$ that is consistent with the metric is made easy by the use of this orthonormal frame system. The connection forms are found to be expressed by[8]

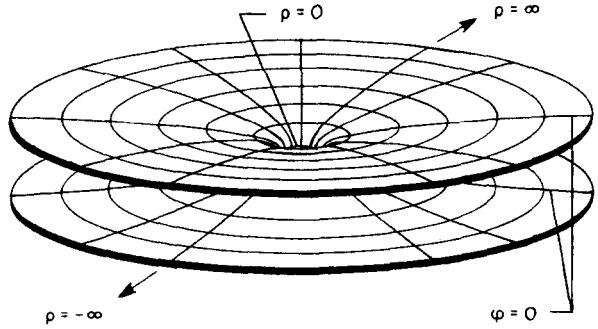

FIG. 1. The equatorial cross section of the typical spatial cross section Σ_t of the space–time manifold $\mathfrak{M}$ in a special case. The line element of this surface is given by $d\sigma^2 = d\rho^2 + (\rho^2 + n^2)d\varphi^2$. The surface is isometric to the catenoid $\{[x,y,z] | (x^2 + y^2)^{1/2} = n\cosh(z/n)\}$ in E^3. The radius of the central hole, where $\rho = 0$, is n. The surface is asymptotic to E^2, both as $\rho \to \infty$ and as $\rho \to -\infty$.

$$[\omega_\kappa{}^\mu] = \begin{array}{c} \\ \kappa\downarrow \end{array} \overset{\mu\rightarrow}{\left[\begin{array}{c|c|c|c} 0 & f'\omega^1 & (r'/r)f\omega^2 & (r'/r)f\omega^3 \\ \hline f'\omega^1 & 0 & (r'/r)\omega^2 & (r'/r)\omega^3 \\ \hline (r'/r)f\omega^2 & -(r'/r)\omega^2 & 0 & [(\mathrm{ctn}\,\vartheta)/r]\omega^3 \\ \hline (r'/r)f\omega^3 & -(r'/r)\omega^3 & -[(\mathrm{ctn}\,\vartheta)/r]\omega^3 & 0 \end{array}\right]}. \tag{7}$$

Because $u = e_0$,

$$\mathbf{d}u = f'(\omega^1 \otimes e_1) + (r'/r)f(\omega^2 \otimes e_2 + \omega^3 \otimes e_3) \tag{8}$$

and

$$\mathbf{d}uu = f'(\omega^1 e_0)e_1 + (r'/r)f[\omega^2 e_0)e_2 + (\omega^3 e_0)e_3] = 0. \tag{9}$$

Another property of the vector field u now becomes apparent: each of its integral paths is geodesic. If p is an integral path of u, then

$$\dot{p} = u(p) = e_0(p); \tag{10}$$

hence

$$\ddot{p} = \mathbf{d}u(p)\dot{p} = (\mathbf{d}uu)(p) = 0. \tag{11}$$

Thus p is geodesic and is parametrized by an affine parameter, which is, because $|\dot{p}|^2 = |u(p)|^2 = 1$, the proper time along p measured from some initial point. Therefore, u generates a congruence of timelike geodesics parametrized by proper time, filling up the space-time manifold $\mathfrak{M}$.

In attempting to understand gravity, I have found it useful to accept as a working hypothesis the existence of a more or less substantial "ether," pervading all of space–time. The ether that I imagine is more than a mere inert medium for the propagation of electromagnetic waves; it is a restless, flowing continuum whose internal, relative motions manifest themselves to us as gravity. Mass particles appear as sinks or sources of this flowing ether. In the case of the space–time manifold $\mathfrak{M}$ under discussion here the velocity that I associate with the ether flow is the vector field u. The geodesic property of u just now established I interpret as saying that every observer or test particle drifting with the ether, following its flow, is absolutely unaccelerated. In this sense my hypothetical ether provides a universal system of inertial observers, just as did the nineteenth-century luminiferous ether, and as must every ether worthy of the name.

It was in pursuing the consequences of this hypothesis that I became convinced of the need to replace the Schwarzschild singularity with a drainhole. Telling how to do that is the principal aim here, and I shall therefore make no effort to justify the ether-flow hypothesis.[9,10] Although henceforth I shall refer to u as "the ether flow velocity" and speak of "the ether" as if it really does exist and flow about, I shall do so not because I expect the reader to adopt this hypothesis, rather because the concepts and terminology provide an expressive and stimulating vehicle of thought that I am accustomed to using. Whether there is such an ether is a question that requires clarification if it is to be answered with confidence.

Returning now to the discussion of geodesics associated with the ether flow velocity u, let us first note that if p is any path in the manifold $\mathfrak{M}$, then, in terms of the orthonormal frame system $\{e_\mu\}$,

$$\dot{p} = \dot{t}e_0(p) + [\dot{\rho} - f(\rho)\dot{t}]e_1(p) + r(\rho)\dot{\vartheta}e_2(p) + r(\rho)(\sin\vartheta)\dot{\varphi}e_3(p), \tag{12}$$

where by abuse of notation t, ρ, ϑ, and φ stand for $t(p)$, $\rho(p)$, $\vartheta(p)$, and $\varphi(p)$. If the path p is that of an observer drifting with the ether, parametrized by his proper time, then Eq. (12) must agree with Eq. (10). Therefore $\dot{\vartheta} = \dot{\varphi} = 0$, meaning that the drift is radial. Further, $\dot{t} = 1$, which says that coordinate time elapses at the same rate as the proper time of the drifting observer. Finally, $\dot{\rho} = f(\rho)$, and it then follows that the coordinate 3-velocity of the drifting ether is $f(\rho)\partial/\partial\rho$, and that the coordinate 3-speed is $|f(\rho)|$.

For the discussion of horizons in Sec. V it will be required that a little about the paths of light rays be known. If p is any null path, then from Eq. (12) it follows that

$$\left(\frac{d\rho}{dt} - f(\rho)\right)^2 + r^2(\rho)\left(\frac{d\Omega}{dt}\right)^2 = 1, \tag{13}$$

unless $\dot{t} = 0$, in which case also $\dot{\rho} = \dot{\vartheta} = \dot{\varphi} = 0$ and p is not a path of a light ray. Inasmuch as the relative coordinate 3-velocity of the path p with respect to the ether flow is

$$\left(\frac{d\rho}{dt} - f(\rho)\right)\frac{\partial}{\partial\rho}(p) + \frac{d\vartheta}{dt}\frac{\partial}{\partial\vartheta}(p) + \frac{d\varphi}{dt}\frac{\partial}{\partial\varphi}(p), \tag{14}$$

the import of Eq. (13) is that the square of the speed of light with respect to the ether, as measured in the coordinate system $[t, \rho, \vartheta, \varphi]$, is 1.

Each of the vector fields $u \pm \partial/\partial\rho$ generates a congruence of null geodesics, for if

$$\begin{aligned} \dot{p} &= u(p) \pm \frac{\partial}{\partial\rho}(p) \\ &= e_0(p) \pm e_1(p), \end{aligned} \tag{15}$$

then $\dot{p}$ is null, and, in view of Eq. (7),

$$\begin{aligned} \ddot{p} &= \mathbf{d}e_0(p)\dot{p} \pm \mathbf{d}e_1(p)\dot{p} \\ &= (\mathbf{d}e_0e_0 \pm \mathbf{d}e_0e_1 \pm \mathbf{d}e_1e_0 + \mathbf{d}e_1e_1)(p) \\ &= f'\dot{p}. \end{aligned} \tag{16}$$

The coordinate 3-velocities of the null geodesics in these two congruences, and their coordinate 3-velocities with respect to the ether flow, are, respectively,

$$[f(\rho) \pm 1]\,\partial/\partial\rho \quad \text{and} \quad \pm\,\partial/\partial\rho. \tag{17}$$

Light rays following these paths are moving in the radial direction if they are moving at all. Those in one group move upstream in the ether, and those in the other group go downstream.

IV. THE CURVATURE TENSOR FIELDS

It is a simple matter to calculate the curvature forms $\Theta_\kappa{}^\mu$ from the formulas (7) for the connection forms $\omega_\kappa{}^\mu$. The result is that

$$[\Theta_\kappa{}^\mu] = \begin{array}{c} \\ \kappa\downarrow \end{array} \overset{\mu\rightarrow}{\begin{bmatrix} 0 & + & + & + \\ \left(\frac{f^2}{2}\right)''\left(\omega^0\wedge\omega^1\right) & 0 & - & - \\ \frac{(r'f)'f}{r}\left(\omega^0\wedge\omega^2\right)+\frac{r''}{r}f\left(\omega^1\wedge\omega^2\right) & -\frac{r''}{r}f\left(\omega^0\wedge\omega^2\right)+\left[\frac{r'}{r}\left(\frac{f^2}{2}\right)'-\frac{r''}{r}\right]\left(\omega^1\wedge\omega^2\right) & 0 & - \\ \frac{(r'f)'f}{r}\left(\omega^0\wedge\omega^3\right)+\frac{r''}{r}f\left(\omega^1\wedge\omega^3\right) & -\frac{r''}{r}f\left(\omega^0\wedge\omega^3\right)+\left[\frac{r'}{r}\left(\frac{f^2}{2}\right)'-\frac{r''}{r}\right]\left(\omega^1\wedge\omega^3\right) & \frac{1-(1-f^2)r'^2}{r^2}\left(\omega^2\wedge\omega^3\right) & 0 \end{bmatrix}}. \tag{18}$$

The isolated + and − signs are meant to reflect the symmetry $\Theta_0{}^m = \Theta_m{}^0$ and the antisymmetry $\Theta_k{}^m = -\Theta_m{}^k$ for $k,m = 1,2,3$.

A brief additional calculation finds the nonvanishing components of the Ricci curvature tensor field to be given by

$$\begin{aligned} R_{00} &= \nabla^2(\tfrac{1}{2}f^2) + 2(r''/r)f^2, \\ R_{01} = R_{10} &= 2(r''/r)f, \\ R_{11} &= -\nabla^2(\tfrac{1}{2}f^2) + 2r''/r, \\ R_{22} = R_{33} &= \{[(\tfrac{1}{2}r^2)'(1-f^2)]' - 1\}/r^2. \end{aligned} \tag{19}$$

Here ∇^2 is the Laplacian for any one of the spacelike hypersurfaces Σ_t orthogonal to u; it is determined by the Riemannian line element (3). For a function $h(\rho)$,

$$\nabla^2[h(\rho)] = [1/r^2(\rho)]\,(r^2h')'(\rho). \tag{20}$$

The scalar field $f^2/2$ that appears in these formulas is $\frac{1}{2}(1-g_{00})$, as calculated in the coordinate system $[t,\rho,\vartheta,\varphi]$. As such, it is the conventional general-relativistic analog, for the gravitational field described by the line element (1), of the negative of the Newtonian gravitational potential. By the same token $-\nabla(f^2/2)$ is the analog of Newton's force of gravity. If the ether flow rate $|f|$ is constant, then this gradient is 0, and, following convention, one has to say that in this case the gravitational field exerts on test particles no force in the Newtonian sense. It is this observation which provides the rationale to identify "gravity" with the internal, relative motions of the postulated ether, as distinguished from its overall rigid motions.

V. HORIZONS

The line element (1) assumes a familiar form upon introduction of a new coordinate T satisfying

$$dT = dt + f(\rho)[1-f^2(\rho)]^{-1}d\rho. \tag{21}$$

It is

$$d\tau^2 = [1-f^2(\rho)]dT^2 - [1-f^2(\rho)]^{-1}d\rho^2 - r^2(\rho)d\Omega^2. \tag{1_S}$$

This is analogous to the usual orthogonal form of Schwarzschild's vacuum line element and reduces to it when $r(\rho)=\rho$ and $f^2(\rho) = 2m/\rho$, m being the mass parameter. It is clear from Eq. (1_S) that the translations along the T coordinate curves are isometries, hence that $\partial/\partial T$ is a Killing vector field. It is also clear that $\partial/\partial T$ is everywhere orthogonal to the hypersurfaces on which T is constant. Therefore, wherever $f^2<1$, so that $\partial/\partial T$ is timelike, the manifold is static. This was stated in Sec. II to be the case; it was also said at that point that $\partial/\partial t$ is hypersurface orthogonal, and this now follows from the determination that $\partial/\partial t = \partial/\partial T$.

The Schwarzschild horizon, where $\rho = 2m$, corresponds in the general case to 2-spheres $S_{t,\rho}$ on which $f(\rho) = \pm 1$. The ether-flow picture includes a graphic interpretation of such horizons. On each such sphere the coordinate speed of the drifting ether, which is $|f(\rho)|$, just matches the speed of light with respect to the ether. From Eq. (13) it follows that if $S_{t,\rho}$ is intersected by the null path p, then $0 \le d\rho/dt \le 2$ if $f(\rho) = 1$, but $-2 \le d\rho/dt \le 0$ if $f(\rho) = -1$; therefore, if p crosses $S_{t,\rho}$, its radial velocity component and that of the ether flow cannot be oppositely directed at the crossing point. Thus light rays can only cross a horizon in the downstream direction of the flow. One can easily check that the only paths of light rays that contact a horizon without crossing it belong to the upstream member of the pair of radial null congruences mentioned in Sec. III; these light rays remain forever on the horizon, struggling to go nowhere. In regions where $f^2(\rho) > 1$, such as Schwarzschild interiors, all light rays are swept along downstream, even those whose motion relative to the ether is upstream. In regions where $f^2(\rho) < 1$, such as Schwarzschild exteriors, some are able to progress upstream, but only with difficulty when near the horizon. People in light canoes should avoid ethereal rapids!

Another space–time manifold whose line element can assume the forms (1) and (1_S), and that possesses a horizon, is the de Sitter cosmological model,[11,12] for which $r(\rho)=\rho$ and $f^2(\rho) = (\rho/R)^2$, R being a positive parameter. It models a universe that is devoid of mat-

ter yet exhibits gravitational effects. Test particles in this universe cannot remain at rest with respect to one another, for they have to share in a cosmological expansion or contraction, which may be identified with a linear expansion or contraction of the ether, reflected in the form that f has. The 2-spheres $S_{t,R}$ constitute a horizon which is the edge of the field of vision for all observers on the upstream side of it. Here, again, gravity corresponds to nonrigid motions of the ether.

VI. THE SCALAR FIELD AS PLUMBER'S FRIEND

We come now to the task of opening up the Schwarzschild singularity so that the ether may flow through unimpededly. To discover the cause of the constriction, let us refer to the formulas (19) for the components of the Ricci curvature tensor field and observe that the Einstein vacuum field equations $R_{00} = R_{01} = R_{11} = 0$ imply that $r'' = 0$, hence that $r(\rho)$ is a linear function of ρ, which in view of the field equation $R_{22} = 0$ cannot be constant, that r must therefore have a zero, and that as ρ approaches this zero the 2-spheres $S_{t,\rho}$ shrink to points, these points constituting the Schwarzschild singularity. In this way we may identify as the cause of constriction an excess of strength in the Einstein vacuum field equations. To weaken these equations and thereby to remove the constriction, an aid is required, a plumber's friend so to speak. Let us find one.

The Ricci tensor field is $\omega^\kappa \otimes R_{\kappa\lambda}\omega^\lambda$, where the $R_{\kappa\lambda}$ are given by Eqs. (19). Look at the terms that involve r''. Their sum can be factored:

$$\begin{aligned} &2(r''/r)[f^2(\omega^0 \otimes \omega^0) + f(\omega^0 \otimes \omega^1 + \omega^1 \otimes \omega^0) + \omega^1 \otimes \omega^1] \\ &\quad = 2(r''/r)(f\omega^0 + \omega^1) \otimes (f\omega^0 + \omega^1) \\ &\quad = 2(r''/r)(d\rho \otimes d\rho). \end{aligned} \tag{22}$$

Now let α be a nonconstant, differentiable, real-valued function on the real line, and let $\phi = \alpha(\rho)$. Then the square of the gradient of the scalar field ϕ is given by

$$d\phi \otimes d\phi = \alpha'^2(d\rho \otimes d\rho). \tag{23}$$

Upon comparing Eq. (23) with Eq. (22) we see that a field equation of the form

$$\text{Ricci tensor field} = K(d\phi \otimes d\phi), \tag{24}$$

with nonzero coupling constant K, will replace the unwanted condition $r'' = 0$ with the less restrictive condition $r'' = \frac{1}{2}K\alpha'^2 r$. This latter condition implies that the radius function r is convex if $K > 0$, but concave if $K < 0$. If r is concave, then it is impossible for the space–time manifold $\mathfrak{M}$ to have a central hole such as the one that Fig. 1 shows a cross section of. The reason is that on each great-circle cross section of a typical spatial cross section Σ_t of $\mathfrak{M}$ the induced Gaussian curvature is given by the scalar field $-r''/r$. Concavity of r renders this curvature everywhere nonnegative, which in a hole of the kind envisioned it cannot be. To enlarge the Schwarzschild singularity into a proper hole, we must therefore take $K > 0$ so that r will be convex.[13]

As it happens, the coupling expressed by Eq. (24) is known to derive from the simple variational principle

$$0 = \delta\int(-g)^{1/2}(R^\kappa{}_\kappa - K\phi^{,\kappa}\phi_{,\kappa})d^4x. \tag{25}$$

Once K has been made positive in Eq. (25), it may be replaced by the number 2, for this involves at most a rescaling of ϕ. Then the Euler equations that together are equivalent to the variational principle are

$$R_{\mu\nu} - \tfrac{1}{2}R^\kappa{}_\kappa g_{\mu\nu} = 2(\phi_{,\mu}\phi_{,\nu} - \tfrac{1}{2}\phi^{,\kappa}\phi_{,\kappa}g_{\mu\nu}) \tag{26}$$

and the scalar wave equation

$$\Box\phi \equiv \mathrm{Tr}\,\mathbf{d}(G^{-1}d\phi) = \phi^{,\kappa}{}_{;\kappa} = 0, \tag{27}$$

in which G is the metric tensor field. Equation (26) is equivalent to

$$R_{\mu\nu} = 2\phi_{,\mu}\phi_{,\nu}, \tag{28}$$

which in turn is equivalent to Eq. (24) with $K = 2$.

These are the field equations by way of which the scalar field ϕ will be applied to the ethereal plumbing problem. The next section will present all of their solution manifolds that have line elements of the form in Eq. (1), as well as the analogous solutions for the equations that would have resulted had K been taken negative. Of the former, one will turn out to be geodesically complete (also static and possessed of a central hole); of the latter, none will.

VII. THE ETHER-FLOW, DRAINHOLE, PARTICLE MODEL

Under the assumption that $\phi = \alpha(\rho)$, the wave equation (27) is equivalent to

$$[r^2(f^2 - 1)\alpha']' = 0, \tag{29}$$

and the field equations (28) are equivalent to the three equations

$$r''/r = \alpha'^2, \tag{30}$$

$$[r^2(f^2/2)']' = 0, \tag{31}$$

$$[(r^2/2)'(1 - f^2)]' = 1. \tag{32}$$

The last two yield, upon integration and rearrangement,

$$r^2(1 - f^2)' = 2m \tag{33}$$

and

$$r^{2\prime}(1 - f^2) = 2(\rho - m), \tag{34}$$

where without loss of generality the zero point of ρ has been adjusted to equalize the integration constants. Combined, these equations produce, after integration,

$$r^2(1 - f^2) = \rho^2 + C. \tag{35}$$

Also, Eq. (29) integrates to

$$\alpha' = -n/[r^2(1 - f^2)] = -n/(\rho^2 + C). \tag{36}$$

From Eqs. (34) and (35) it follows that

$$r'/r = (\rho - m)/(\rho^2 + C), \tag{37}$$

hence that

$$r''/r = (r'/r)' + (r'/r)^2 = (C + m^2)/(\rho^2 + C)^2. \tag{38}$$

Thus Eq. (30) adds only the information that $C = n^2 - m^2$. Equations (36) and (37) now imply that

$$r^2(\rho) = |\rho^2 + n^2 - m^2|\, e^{(2m/n)\alpha(\rho)}, \tag{39}$$

provided α is made to absorb additively the integration constant, which it can do with no change in α', hence with no effect on the field equations.

At this stage Eq. (1_S) for the line element has been specialized to

$$d\tau^2 = \operatorname{sgn}(\rho^2 + n^2 - m^2)\{e^{-(2m/n)\alpha(\rho)}d\tau^2 - e^{(2m/n)\alpha(\rho)}[d\rho^2 + (\rho^2 + n^2 - m^2)d\Omega^2]\}, \tag{$1_{S,\alpha}$}$$

and the only integration left to be done is that of Eq. (36), which now reads

$$\alpha' = -n/(\rho^2 + n^2 - m^2). \tag{36'}$$

This requires consideration of three cases: (I) $n^2 < m^2$; (II) $n^2 = m^2$; (III) $n^2 > m^2$. In each case n will be taken to be nonnegative, for at the end only n^2 will appear. Also, the boundary condition that

$$\lim_{\rho\to\infty} \phi \equiv \lim_{\rho\to\infty} \alpha(\rho) = 0 \tag{40}$$

will be applied. This limit always exists; requiring it to be 0 is equivalent to requiring that the line element in the form $(1_{S,\alpha})$ be asymptotic to a Schwarzschild vacuum line element (with mass parameter m, it turns out). Within isometric equivalence this boundary condition does not reduce the set of solution manifolds, the reason being that it has no effect on α', and only α' appears in the field equations.

Case I $(n^2 < m^2)$

Let $a = (m^2 - n^2)^{1/2}$. Then

$$\begin{aligned} \alpha'(\rho) &= -n/(\rho^2 - a^2), \\ \alpha(\rho) &= (n/2a)\log|(\rho + a)/(\rho - a)|, \\ r^2(\rho) &= |\rho^2 - a^2|\cdot|(\rho + a)/(\rho - a)|^{m/a} \\ &= |\rho + a|^{(m/a)+1}/|\rho - a|^{(m/a)-1}, \\ f^2(\rho) &= 1 - \operatorname{sgn}(\rho^2 - a^2)|(\rho - a)/(\rho + a)|^{m/a}. \end{aligned} \tag{41}$$

When $n > 0$, there is a separation of the space–time manifold $\mathfrak{M}$ to which these formulas apply into three connected submanifolds, corresponding to the radial coordinate ranges $\rho < -a$, $-a < \rho < a$, and $a < \rho$. If $m \neq 0$, the formula for $f^2(\rho)$ implies that $f^2 > 0$ on two of these submanifolds, but that $f^2 < 0$ on the other one, namely, the one corresponding to $\rho < -a$, if $m > 0$, but the one corresponding to $a < \rho$, if $m < 0$. Because f is imaginary, the line element on this submanifold, though real in the form (1_S), is complex in the form (1), and t must be interpreted as a complex coordinate, related to the real coordinates T and ρ by Eq. (21). The computations that have gone before all remain valid, but the description of the geometry and the interpretation of the vector field u must be modified. In particular the cross sections Σ_t are two-dimensional instead of three-dimensional, and u is complex instead of real. There is no horizon of the Schwarzschild type on this submanifold, for these occur only where $f^2(\rho) = 1$.

In the two submanifolds of $\mathfrak{M}$ on which $f^2 > 0$ the typical spatial cross section Σ_t is three-dimensional, and its shape is determined by the function r, whose graph when $m > 0$ is shown in Fig. 2. (Reflection of this graph through the vertical axis produces the graph of r when $m < 0$.) If, let us say, $m > 0$, then in the submanifold on which $a < \rho$ the radius $r(\rho)$ of the 2-sphere $S_{t,\rho}$ decreases from ∞ to a positive minimum value $r(m)$ as ρ decreases from ∞ to m, after which it returns to ∞ as $\rho \to a$. Therefore, each cross section Σ_t in this submanifold has a central hole of positive minimum radius $r(m)$. In the other submanifold, $S_{t,\rho}$ undergoes infinite expansion as $\rho \to a$, but shrinks toward point size as $\rho \to -a$; the cross section Σ_t thus has in its center only a pinhole and not a hole of positive radius. Neither of these two submanifolds has a horizon except in the asymptotic sense that $f^2(\rho) \to 1$ as $\rho \to a$. When $m < 0$, their geometry is demonstrably the same.

The Schwarzschild manifold occurs when $n = 0$, in which case $a = |m|$ and m is the Schwarzschild mass. The graph of r when $m > 0$ is included in Fig. 2. The Schwarzschild singularity, where $r(\rho) = 0$, corresponds to $\rho = -m$, and the horizon, where $f^2(\rho) = 1$, to $\rho = m$.

An illumination is cast upon the Schwarzschild solution by the observation that it is unstable as a solution of the field equations (26) and (27) in that, as $n \to 0$, r converges pointwise to the Schwarzschild form, but not uniformly. The two submanifolds on which $f^2 > 0$ coalesce, but only reluctantly, at the Schwarzschild horizon. This phenomenon is another aspect of the behavior of the Schwarzschild horizon under perturbations, discussed by Janis, Newman, and Winicour,[14] and by Penney.[15] They have found and examined a solution of the field equations used here, but with the coupling constant negative rather than positive—for them $K < 0$ in the variational principle (25). Their line element is the same, but for choice of coordinate system and parameter names, as the one given by Eqs. (41)

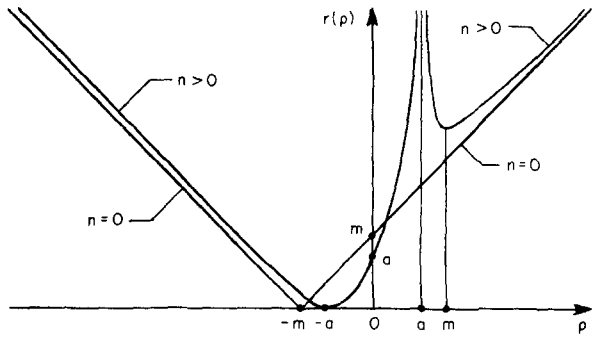

FIG. 2. The graph, for $m > 0$, of the radius function r in Case I $(n^2 < m^2)$. Here $r^2(\rho) = |\rho + a|^{1+m/a}|\rho - a|^{1-m/a}$, and $a = (m^2 - n^2)^{1/2}$. For $m < 0$, reflect the graph through the vertical axis.

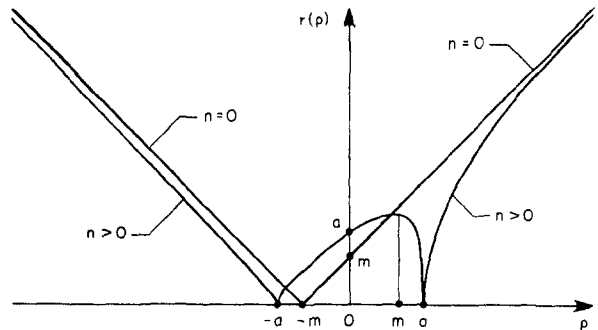

FIG. 3. The graph, for $m > 0$, of the radius function r obtained from the Case I solution by replacing n^2 with $-n^2$. Here $r^2(\rho) = |\rho + a|^{1+m/a} \times |\rho - a|^{1-m/a}$, and $a = (m^2 + n^2)^{1/2}$. The corresponding line element satisfies the field equations generated by the variational principal of Eq. (26) with $K = -2$.

with $a = (m^2 + n^2)^{1/2}$. As such, it is the only solution of the negatively coupled field equations that can take the form of Eq. (1), except the solution, falling under Case II, in which $m = n = 0$. The graph of the radius function r for this solution is depicted in Fig. 3, again under the assumption that $m > 0$. This r likewise converges pointwise but not uniformly to the Schwarzschild r as $n \to 0$. And here, also, there is, when $n > 0$, separation of $\mathfrak{M}$ at $-a$ and a into three connected submanifolds, none containing a horizon. As was forecast in Sec. VI, r is concave; hence none of these submanifolds has more than a pinhole at its center. This solution was earlier discovered by Bergmann and Leipnik.[16]

Using Eq. (18), and assuming either coupling polarity, one can easily see that if $n > 0$, then at each of the edges where $\rho^2 \to a^2$ some of the curvature components become infinite, but that if $n = 0$, this happens only at the edges where $\rho \to -m$. Because the frame system $\{e_\mu\}$ is orthonormal, these apparent singularities in curvature are real. Owing to their presence, it is impossible to extend metrically across one of these edges any submanifold of $\mathfrak{M}$. For this reason neither $\mathfrak{M}$ nor any possible metric extension of $\mathfrak{M}$ will be geodesically complete if there is a geodesic in $\mathfrak{M}$ that arrives at one of these edges without using up its affine parameter. That there are such geodesics will be established in Sec. VIII.

Case II ($n^2 = m^2$)

Here

$$\alpha'(\rho) = -n/\rho^2, \quad \alpha(\rho) = n/\rho$$
$$r^2(\rho) = \rho^2 e^{2m/\rho}, \quad f^2(\rho) = 1 - e^{-2m/\rho}. \tag{42}$$

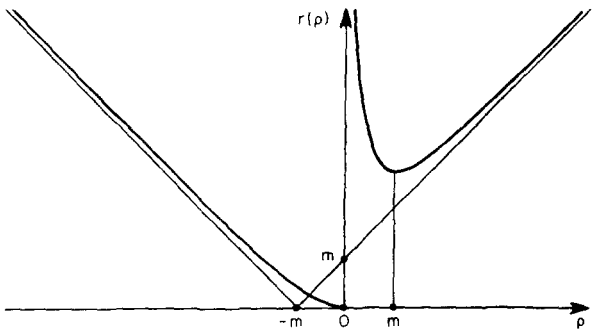

FIG. 4. The graph, for $m > 0$, of the radius function r in Case II ($n^2 = m^2$). Here $r^2(\rho) = \rho^2 e^{2m/\rho}$. For $m < 0$, reflect the graph through the vertical axis.

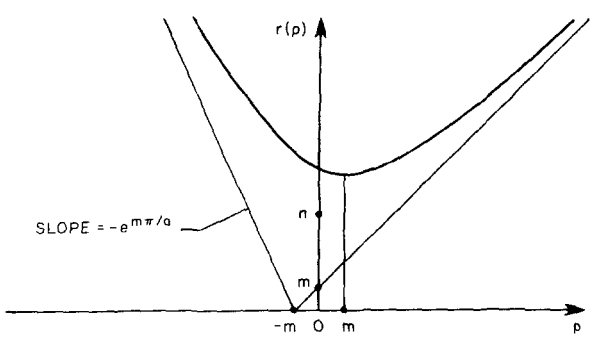

FIG. 5. The graph for $m \geq 0$, of the radius function r in Case III ($n^2 > m^2$). Here $r^2(\rho) = (\rho^2 + a^2)e^{(2m/n)\alpha(\rho)}$, $\alpha(\rho) = (n/a)[\pi/2 - \tan^{-1}(\rho/a)]$, and $a = (n^2 - m^2)^{1/2}$. The minimum value of r, namely $r(m)$, is the radius of the drainhole; it ranges from n up to $ne-$ as m goes from 0 to $n-$, and it always exceeds $2m$. That the associated manifold $\mathfrak{M}_{m,n}$ is asymptotically Schwarzschildean as $\rho \to \infty$ is reflected in the relation $r(\rho) \sim \rho + m$ as $\rho \to \infty$.

This is the limiting case of Case I as $a \to 0$. The manifold on which $-a < \rho < a$ has been squeezed out. The two remaining manifolds, corresponding now to $\rho < 0$ and $0 < \rho$, are in all qualitative aspects, including infinite edge curvatures, unchanged (unless $m = n = 0$, which results in two copies of flat Minkowski space-time). In particular, if $m \neq 0$, neither of them possesses a horizon, is metrically extendible, or is geodesically complete. The graph of r for $m > 0$ is shown in Fig. 4. The line element has been exhibited by Yilmaz in the form $(1_{S,\alpha})$.[17]

Case III ($n^2 > m^2$)

This is the case of greatest interest, for the line element measures a connected and geodesically complete space-time manifold with a drainhole. Let $a = (n^2 - m^2)^{1/2}$. Then

$$\alpha'(\rho) = -n/(\rho^2 + a^2), \quad \alpha(\rho) = (n/a)[\tfrac{1}{2}\pi - \tan^{-1}(\rho/a)],$$
$$r^2(\rho) = (\rho^2 + a^2)e^{(2m/n)\alpha(\rho)}, \quad f^2(\rho) = 1 - e^{-(2m/n)\alpha(\rho)}. \tag{43}$$

Because r^2 and $1 - f^2$ are everywhere analytic and positive, these formulas determine an analytic line element of the form (1_S), which now becomes $(1_{S,\alpha})$ with $n^2 - m^2 = a^2$. This line element fits a manifold on which the coordinate ρ ranges from $-\infty$ to ∞, as does also the coordinate T. This manifold will be shown in the next section to be geodesically complete. Let it be denoted $\mathfrak{M}_{m,n}$.

If $m \geq 0$, then $0 \leq f^2 < 1$, and the form (1) of the line element of $\mathfrak{M}_{m,n}$ is real. The relation between the time coordinates t and T, expressed by Eq. (21), depends upon whether $f \geq 0$ or $f \leq 0$, but in either event t is real and ranges from $-\infty$ to ∞. If $m = 0$, then $f = 0$ (the ether is at rest). When $m > 0$, $f^2(\rho)$ decreases from $1 - e^{-2m\pi/a}$ to 0 as ρ goes from $-\infty$ to ∞. There is no horizon, because $f^2(\rho)$ is never 1.

Figure 5 displays the graph of the radius function r when $m \geq 0$. The 2-spheres $S_{t,\rho}$ of constant t and constant ρ are smallest when $\rho = m$; they undergo infinite expansion both as $\rho \to \infty$ and as $\rho \to -\infty$. It follows from Eqs. (43) that the minimum radius $r(m)$, considered as a function of m, increases from n to $ne-$ as m goes from 0 to $n-$. Thus the order of magnitude of the radius of the drainhole is determined by n, the only noticeable effect of m being to bound it below via the first two of the inequalities $m < n \leq r(m) < ne$ (actually, as Fig. 5 shows, $r(m) > 2m$).

It is not difficult to establish that the following asymptotic relations hold, whether $m > 0$, $m = 0$, or $m < 0$: as $\rho \to \infty$,

$$\alpha(\rho) = (n/\rho) + O(1/\rho^3),$$
$$r(\rho) = (\rho + m) + O(1/\rho), \tag{44}$$
$$f^2(\rho) = [2m/(\rho + m)] + O(1/\rho^2);$$

as $\rho \to -\infty$,

$$\alpha(\rho) = (n\pi/a) + (n/\rho) + O(1/\rho^3),$$
$$r(\rho) = -(\rho + m)e^{m\pi/a} + O(1/\rho), \tag{45}$$
$$f^2(\rho) = 1 - e^{-2m\pi/a}[1 - 2m/(\rho + m)] + O(1/\rho^2).$$

These relations imply that the manifold $\mathfrak{M}_{m,n}$ is, in the

usual loose sense, asymptotic as $\rho \to \infty$ to the Schwarzschild manifold with mass parameter m. They also imply that $\mathfrak{M}_{m,n}$ is asymptotically flat, both as $\rho \to \infty$ and as $\rho \to -\infty$, because, as may easily be checked, the coefficients $(f^2/2)''$, $(r'f)'f/r$, etc., in the expression (18) for the curvature forms $\Theta_\kappa^{\ \mu}$ are asymptotic to 0. (This criterion for asymptotic flatness is acceptable because the frame system $\{e_\mu\}$ is orthonormal.) It is natural to ask if $\mathfrak{M}_{m,n}$ is also asymptotically Schwarzschildean as $\rho \to -\infty$. The answer is that $\mathfrak{M}_{m,n}$ is asymptotic as $\rho \to -\infty$ to the Schwarzschild manifold with mass parameter $-me^{m\pi/a}$. This somewhat surprising conclusion is a consequence of the observation that if $\bar{m} = -me^{m\pi/a}$ and $\bar{n} = ne^{m\pi/a}$, then there is an isometry between $\mathfrak{M}_{m,n}$ and $\mathfrak{M}_{\bar{m},\bar{n}}$ which reverses the direction of increase of the radial coordinate and thus matches up opposed asymptotic regions. Such an isometry is obtained by identifying the point of $\mathfrak{M}_{m,n}$ having S-type coordinates $[T, \rho, \vartheta, \varphi]$ with the point of $\mathfrak{M}_{\bar{m},\bar{n}}$ whose S-type coordinates are $[Te^{-m\pi/a}, -\rho e^{m\pi/a}, \vartheta, \varphi]$.

Because these isometries exist, no physically useful distinction can be drawn between the manifolds with positive mass parameters m and those for which m is negative. On the other hand, in each such manifold there is a clear physical distinction between the two sides of the drainhole, for one side is asymptotic to a Schwarzschild manifold whose mass parameter is positive, while the other is asymptotic to a Schwarzschild manifold whose mass parameter is negative. In the study of the geodesics of $\mathfrak{M}_{m,n}$ it will appear that, when $m > 0$, test particles are always accelerated in the direction of decreasing ρ. This is toward the drainhole when $\rho > m$, but away from it when $\rho < m$. Thus if $m > 0$ (and likewise if $m < 0$), the manifold $\mathfrak{M}_{m,n}$ models a Janus-faced particle that attracts matter on one side and repels it (more strongly) on the other.

The curious asymmetry between the positive mass, say m, and the negative mass $\bar{m}$ of the particle, expressed by the equation

$$\bar{m}/m = -\exp[m\pi/(n^2 - m^2)^{1/2}], \tag{46}$$

is a facet of the model that is especially eye-catching. It is even more so in light of the observation that certain not unnatural specifications of m and n will cause the equation to generate some of Dirac's outsized, dimensionless physical "constants", which are about 10^{20^k}, where k is some small nonzero integer.[18] Specifically, if m is of the order of a proton mass, 1.2×10^{-52} cm, while n is of the order of Planck's length, 1.6×10^{-33} cm, then $-\bar{m}/m \approx 1 + 10^{-19}$. If instead $n \approx 2.8 \times 10^{-13}$ cm, the classical electron radius, then $-\bar{m}/m \approx 1 + 10^{-39}$. A speculative extrapolation from the asymmetry between m and $\bar{m}$ is that the universe expands because it contains more negative mass than positive, each half-particle of positive mass m being slightly overbalanced by a half-particle of negative mass $\bar{m}$ such that $-\bar{m} > m$.

The case where $m = 0$ is particularly interesting. The ether is not flowing, because $f = 0$. However, the drainhole remains open, because $r(\rho) = (\rho^2 + n^2)^{1/2} \geq n > 0$. The manifold is symmetric with respect to reflection through the drainhole. The catenoid of Fig. 1 is the cross section of $\mathfrak{M}_{0,n}$ on which $t = 0$ and $\vartheta = \pi/2$. Although massless, the particle modeled interacts with test particles, as the study of its geodesics will show.

The scalar field ϕ that holds the drainhole open satisfies the scalar wave equation. If in the flowless case some disturbance were to cause the drainhole to pinch in two, there would be left on each side a central bump in a topologically and asymptotically Euclidean 3-space. These bumps, being directly associated with ϕ via the field equations (28), would radiate away with the fundamental speed of wave propagation. The particle would have dematerialized from a drainhole to a ϕ-wave. When $m \neq 0$, the same thing presumably could happen, but in addition there should arise a traveling gradient in the ether flow, identifiable, one imagines, as gravitational radiation. Such a picture of changing topology and geometry provides a graspable basis for attempts at understanding the wave-particle duality of matter.

VIII. GEODESICS AROUND, ABOUT, AND THROUGH THE DRAINHOLE

The starting point for the study of the geodesics of the manifold $\mathfrak{M}$ bearing the line element (1) is the earlier equation

$$\dot{p} = \dot{t}e_0(p) + [\dot{\rho} - f(\rho)\dot{t}]e_1(p) + r(\rho)\dot{\vartheta}e_2(p) + r(\rho)(\sin\vartheta)\dot{\varphi}e_3(p), \tag{12}$$

which holds for every path p in $\mathfrak{M}$. From Eqs. (12), (A3), (A1), (7), and (5) it follows that

$$\begin{aligned}\ddot{p} = &\left[\ddot{t} + f'(\dot{\rho} - f\dot{t})^2 + f\left(\frac{r^2}{2}\right)'\dot{\Omega}^2\right]\frac{\partial}{\partial t}(p)\\ &+\left[\ddot{\rho} - \left(\frac{f^2}{2}\right)'[\dot{t}^2 - (\dot{\rho} - f\dot{t})^2]\right.\\ &\left.- (1 - f^2)\left(\frac{r^2}{2}\right)'\dot{\Omega}^2\right]\frac{\partial}{\partial \rho}(p)\\ &+\left[\ddot{\vartheta} + 2\frac{r'}{r}\dot{\rho}\dot{\vartheta} - (\sin\vartheta)(\cos\vartheta)\dot{\varphi}^2\right]\frac{\partial}{\partial \vartheta}(p)\\ &+\left(\ddot{\varphi} + 2\frac{r'}{r}\dot{\rho}\dot{\varphi} + 2(\operatorname{ctn}\vartheta)\dot{\vartheta}\dot{\varphi}\right)\frac{\partial}{\partial \varphi}(p),\end{aligned} \tag{47}$$

where

$$\dot{\Omega}^2 = \dot{\vartheta}^2 + (\sin\vartheta)^2\dot{\varphi}^2. \tag{48}$$

Now let p be a maximally extended geodesic path, affinely parametrized, so that $\dot{p} = 0$. This equation is equivalent to the four scalar equations that say that the components of $\dot{p}$ in Eq. (47) are 0. For reference call these the t-, ρ-, ϑ-, and φ-equations.

Reflecting the spherical symmetry of the metric, the ϑ- and φ-equations entail that the orbit of the path p lies in one of the great-circle cross sections of $\mathfrak{M}$, which are those hyperspaces typified by the equatorial cross section, defined by $\vartheta = \pi/2$. The angular-momentum first integral of the ϑ- and φ-equations is

$$r^2\dot{\Omega} = h. \tag{49}$$

The t- and ρ-equations have the first integral

$$(1 - f^2)\dot{t} + f\dot{\rho} = k. \tag{50}$$

Suppose next that the parameter on p is the proper time along p if p is timelike, the proper distance along p if p is spacelike. Then the t-, ρ-, ϑ-, and φ-equations have the first integral

$$\epsilon = |\dot{p}|^2 = \dot{t}^2 - (\dot{\rho} - f\dot{t})^2 - r^2\dot{\Omega}^2, \tag{51}$$

where ϵ, the indicator of p, is 1, 0, or -1, according as p is timelike, null, or spacelike. A consequence of Eqs. (49), (50), and (51) is that

$$\dot{\rho}^2 = k^2 - (1 - f^2)(\epsilon + h^2/r^2). \tag{52}$$

When the first integral (51) is used in the ρ-equation, there results

$$\ddot{\rho} = \epsilon(\tfrac{1}{2}f^2)' + \tfrac{1}{2}[(r^2)'(1 - f^2) - (r^2)(1 - f^2)']\dot{\Omega}^2. \tag{53}$$

If we utilize the integrals (33) and (34) of the field equations, then Eq. (53) becomes

$$\ddot{\rho} = \epsilon(-m/r^2) + (\rho - 2m)\dot{\Omega}^2. \tag{54}$$

This equation applies in each of Cases I, II, and III of Sec. VII. It implies that, when $m > 0$, test particles on radial paths are always accelerated in the direction of decreasing ρ. In Case III this means that the drainhole attracts matter on the side identified by asymptotic comparison to a Schwarzschild manifold as having positive mass, and repels it on the side to which negative mass has been ascribed.

Completeness

For a null, radial geodesic, $\epsilon = h = 0$, and Eq. (52) implies that $\dot{\rho} = \pm k$. If $k = 0$, then ρ is constant, and Eq. (50) allows two possibilities. One is that $\dot{t} = 0$, in which case t, also, is constant; the geodesic is degenerate, frozen at one point of space–time. The other possibility is that $f^2(\rho) = 1$; in this case the light signal whose path is p is stuck on a horizon, but not frozen in time. If on the other hand $k \neq 0$, then ρ is a nonconstant, linear function of the affine parameter. From this it follows that if $\mathfrak{M}$ is any one of the nonflat space–time manifolds discussed under Cases I and II of Sec. VII, then $\mathfrak{M}$ has null radial geodesics that come up to an edge where there are infinite curvatures without exhausting their affine parameters in the process. As was remarked in Sec. VII, this implies that none of those manifolds has a geodesically complete extension.

Turning now to Case III, let us see whether the space–time manifold $\mathfrak{M}_{m,n}$ is geodesically complete. Denote by p^+ that portion of the path p on which the parameter is nonnegative. If p^+ is confined to a compact region of the manifold, then p^+ includes all nonnegative numbers in its parameter interval, for p is by hypothesis maximally extended. If p^+ is not so confined, then either $\rho(p^+)$ or $t(p^+)$ is unbounded. But $f^2(\rho)$ and $r^2(\rho)$ are defined for all values of ρ, and both $1 - f^2$ and $1/r^2$ are bounded. Hence Eq. (52) implies that $\dot{\rho}$ is bounded. On the other hand, $1/(1 - f^2)$ is bounded, and, therefore, in view of Eq. (50), $\dot{t}$ is bounded. No unbounded function with bounded derivative is restricted to a bounded interval, so that again the parameter of p consumes all the nonnegative numbers. In the same fashion p's parameter uses up all the nonpositive numbers. Therefore, $\mathfrak{M}_{m,n}$ is indeed geodesically complete.

It is interesting to note that completeness depends only upon these properties of f^2 and r^2 in addition to the smoothness that they possess: (a) Each of f^2 and r^2 is defined on the interval $(-\infty, \infty)$; (b) r^2 is bounded away from 0, so that there is in fact a hole in the manifold that is bigger than a point; (c) f^2 is bounded; (d) f^2 is bounded away from 1, which means that there is no horizon, not even an asymptotic one at an edge of the manifold.

Goedesics of $\mathfrak{M}_{0,n}$

In describing the geodesics of the manifolds $\mathfrak{M}_{m,n}$ of Case III it will be easiest to treat $\mathfrak{M}_{0,n}$ separately. The condition $m = 0$ is equivalent to $f = 0$; the first part of the discussion will apply merely if $f = 0$, irrespective of whether any field equations are satisfied. The line element (1) decomposes into a purely temporal part and a purely spatial part; this shows up in Eq. (51), which now reads

$$\epsilon = \dot{t}^2 - \dot{\sigma}^2, \tag{55}$$

where

$$\dot{\sigma}^2 = \dot{\rho}^2 + r^2(\rho)\dot{\Omega}^2. \tag{56}$$

Because of this decomposition the Killing vector field $\partial/\partial t$ is orthogonal to the spatial cross sections Σ_t, and the projection of the geodesic path p on any one Σ_t via translation of its points along the t lines is a (perhaps degenerate) geodesic curve of Σ_t. This curve is also a spacelike (or else degenerate) geodesic curve of the full space–time manifold $\mathfrak{M}$, and σ measures proper distance along it.

From Eqs. (50) and (55) it follows that $\dot{\sigma}^2 = k^2 - \epsilon$, hence that $\ddot{\sigma} = 0$. Thus test particles undergo no accelerations of the classical Newtonian kind that are associated with forces. In this sense the manifold $\mathfrak{M}$ produces no gravitational effects on test particles (or on light rays, for that matter), and $\mathfrak{M}$ can therefore be said to be devoid of gravitating mass. This, however, is not to say that $\mathfrak{M}$ is free of all matter. The reason for not ruling out massless matter is that in $\mathfrak{M}_{0,n}$ all nonradial test particle or light ray paths bend toward the drainhole, even to the extent that many of them loop around it again and again. This will become apparent as next the geodesics of $\mathfrak{M}_{0,n}$ are described in detail.

It is sufficient to consider in $\mathfrak{M}_{0,n}$ those spacelike geodesic paths p for which $\dot{t} = k = 0$, inasmuch as all geodesics project onto them in the manner described above; these are just the geodesic paths of Σ_t with respect to the inherited Riemannian line element (3), parametrized by arc length. It is further sufficient to consider the case where $\vartheta = \pi/2$, and then p will lie on the catenoid depicted in Fig. 1. On some of these geodesics $\rho \equiv 0$ and $\dot{\varphi}^2 = 1/n^2$. On all others any zero that $\dot{\rho}$ has must be isolated, and for these Eqs. (49) and (52) can be combined into the orbital equation

$$\left(\frac{d\varphi}{d\rho}\right)^2 \equiv \left(\frac{\dot{\varphi}}{\dot{\rho}}\right)^2 = \frac{h^2}{r^2(r^2 - h^2)} = \frac{h^2}{(\rho^2 + n^2)(\rho^2 + n^2 - h^2)}, \tag{57}$$

valid except at isolated points of the path p.

The geodesics fall naturally into three classes, corresponding to (a) $h^2 > n^2$, (b) $h^2 = n^2$, and (c) $h^2 < n^2$. Typical and atypical geodesics in these classes are shown in Fig. 6. Each of them reflects through the drainhole onto a geodesic of the same class.

A typical geodesic satisfying $h^2 > n^2$ spirals in from infinity to a minimum distance $(h^2 - n^2)^{1/2}$ from the neck of the drainhole (where $\rho = 0$), and then spirals out to infinity again. The smaller the distance of closest approach to the neck of the drainhole, the greater the

number of revolutions around the drainhole. A test particle on such an orbit can be trapped for any length of time (whether coordinate time t or proper time τ), but ultimately it will escape. There are no atypical geodesics in this class.

If $h^2 = n^2$, a typical geodesic orbit starts from infinity and spirals in asymptotically to the center circle, which itself is the lone atypical geodesic orbit for this case. A test particle on one of these orbits will be trapped forever, or, if it follows the orbit in reverse, has been trapped forever but is gradually escaping.

In case $h^2 < n^2$, a typical geodesic spirals in from infinity, passes through the drainhole, and spirals out to infinity on the other side. The atypical geodesics trace out the ρ lines, which pass through the hole but do not spiral. Test particles following these orbits are lost forever to observers on the initial side, who would be able, however, upon looking toward the drainhole, to see them slowly fading away, like scintillations in a crystal ball.

The capturing of test particles and of light rays by the flowless drainhole for various lengths of time ranging upward to infinity would seem to warrant thinking of the manifold $\mathfrak{M}_{0,n}$ as at least a rudimentary model of what a massless nuclear binding particle might be like. One's inclination in this direction is reinforced by the observation that the capture effect is of short range. For example, if the distance of closest approach is 10 times the drainhole radius n, then the total bending of the geodesic amounts to less than 0.5°. For the total bending to be 180° (half a loop), the distance of closest approach must be about $0.2n$, which puts the point of closest approach on a sphere of symmetry whose radius is about $1.02n$; for a full loop the corresponding numbers are about $0.03n$ and $1.0006n$.

Geodesics of $\mathfrak{M}_{m,n}$ $(m > 0)$

The discussion will proceed mainly from Eq. (52), rewritten as

$$\dot{\rho}^2 = 2E + F_\epsilon(h^2, \rho), \tag{58}$$

where $E = \frac{1}{2}(k^2 - \epsilon)$ and

$$F_\epsilon(h^2,\rho) = \epsilon f^2 - h^2(1-f^2)/r^2$$
$$= \epsilon[1 - e^{-(2m/n)\alpha(\rho)}] - h^2/e^{(4m/n)\alpha(\rho)}(\rho^2 + a^2)^1. \tag{59}$$

An adequate qualitative description of the geodesics can be easily read off from the graphs, for $\epsilon = 0, 1, -1$, of the functions of the family $F_\epsilon(h^2,\rho)$. Because $\ddot{\rho} = \frac{1}{2}F_\epsilon'(h^2,\rho)$, the turning points of orbits will occur where $F_\epsilon(h^2,\rho) = -2E$ and $F_\epsilon'(h^2,\rho) \neq 0$, and circular orbits will occur where $F_\epsilon(h^2,\rho) = -2E$ and $F_\epsilon'(h^2,\rho) = 0$. The circular orbits will be stable if $F_\epsilon''(h^2,\rho) < 0$, unstable if $F_\epsilon''(h^2,\rho) > 0$.

Null geodesics

Here $2E = k^2 \geq 0$. The graphs of the functions $F_0(h^2,\rho)$ appear in Fig. 7. Adding $2E$ to $F_0(h^2,\rho)$ to get $\dot{\rho}^2$ shifts the graphs upward (unless $2E = 0$); only those points of the graphs that are shifted to the upper closed half-plane correspond to points of geodesics. For various ranges of E and h^2 the possibilities can be summarized as follows:

(i) $E = 0$:

(a) $h^2 = 0$; degenerate geodesic at each point of $\mathfrak{M}_{m,n}$.

(b) $h^2 > 0$; no geodesic.

(ii) $E > 0$:

(a) $0 \leq h^2 < 2Er^2(2m)[1 - f^2(2m)]^{-1}$; geodesics beginning at ∞, passing through the drainhole, ending at $-\infty$, and vice versa.

(b) $h^2 = 2Er^2(2m)[1 - f^2(2m)]^{-1}$; geodesics with unstable circular orbit at $2m$; geodesics beginning at ∞

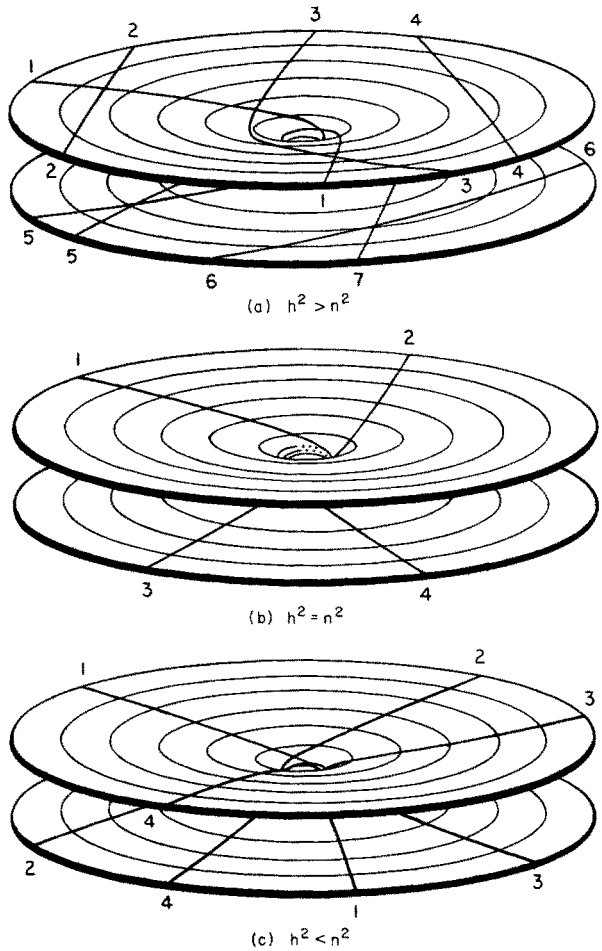

FIG. 6. Typical and atypical orbits of test particles (a) around, (b) about, and (c) through the drainhole of Case III ($n^2 < m^2$) when $m = 0$. The surface to which the orbits are confined is the catenoid of Fig. 1. It is isometric to every great-circle cross section of the spherically symmetric space surrounding the drainhole. The orbits fall into the three classes according to the amount h of angular momentum. The only atypical orbits are the central circle in (b) and the radial lines in (c). Every reflection of an orbit in the drainhole is again an orbit of the same class.

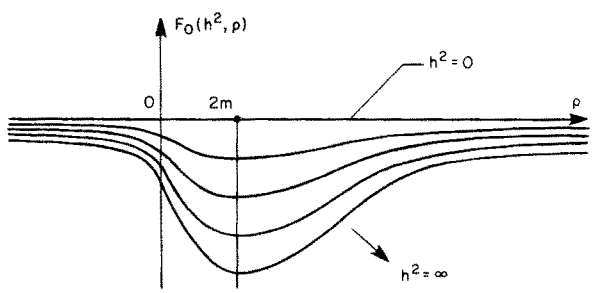

FIG. 7. The graphs of the functions $F_0(h^2,\rho)$ for various values of h^2. Each function has a minimum at $2m$. From the equation $\dot{\rho}^2 = 2E + F_0(h^2,\rho)$ one can find the turning points of null geodesics of $\mathfrak{M}_{m,n}$ by referring to this picture.

and ending by spiraling down to the circular orbit, and vice versa, spiraling up from the circular orbit to ∞; geodesics beginning at $-\infty$, passing through the drainhole, and ending by spiraling up to the circular orbit, and vice versa.

(c) $h^2 > 2Er^2(2m)[1-f^2(2m)]^{-1}$; geodesics beginning and ending at ∞, reaching lowest points (ρ a minimum) which move up from just above $2m$ to ∞ as E decreases or h^2 increases; geodesics beginning and ending at $-\infty$, reaching highest points (ρ a maximum) which move down from just below $2m$ to $-\infty$ as E decreases or h^2 increases.

Timelike geodesics

In this case $2E = k^2 - 1 \geq -1$. Figure 8 exhibits the graphs of the functions $F_1(h^2,\rho)$. Their critical points occur where $h^2 = mr^2(\rho)(\rho - 2m)^{-1} \equiv \gamma(\rho)$, which is on the upper side of $2m$. The locus of critical points has a maximum where $\rho = 3m + (4m^2+n^2)^{1/2} \equiv \rho_0$, and a zero, ρ_1, between $2m$ and ρ_0. The catalog of timelike geodesics reads as follows:

(i) $-1 \leq 2E \leq 1 + e^{-(2m\pi/a)}$: no geodesic.

(ii) $-1 + e^{-(2m\pi/a)} < 2E < -F_1(\gamma(\rho_0),\rho_0)$:

(a) $0 \leq h^2 < \gamma(\rho_0)$; geodesics beginning and ending at $-\infty$, reaching highest points (ρ a maximum) which move from above ρ_0 down to $-\infty$ as E decreases.

(b) $h^2 \geq \gamma(\rho_0)$; geodesics beginning and ending at $-\infty$, reaching highest points below ρ_0 which move down to $-\infty$ as E decreases or h^2 increases.

(iii) $2E = -F_1(\gamma(\rho_0),\rho_0)$:

(a) $0 \leq h^2 < \gamma(\rho_0)$; geodesics beginning and ending at $-\infty$, reaching highest points above ρ_0.

(b) $h^2 = \gamma(\rho_0)$; geodesics with semistable circular orbit at ρ_0 [small perturbations satisfying $2(E+\Delta E) + F_1(h^2+\Delta h^2,\bar\rho) \leq 0$, where $\bar\rho$ is the lesser root of $\gamma(\rho) = h^2 + \Delta h^2$, change the orbit but little, and all other small perturbations result in orbital decay to $-\infty$]; geodesics beginning at $-\infty$ and ending by spiraling up to the circular orbit, and vice versa, spiraling down from the circular orbit to $-\infty$.

(c) $h^2 > \gamma(\rho_0)$; geodesics beginning and ending at $-\infty$, reaching highest points which move from just below ρ_0 down to $-\infty$ as E decreases or h^2 increases.

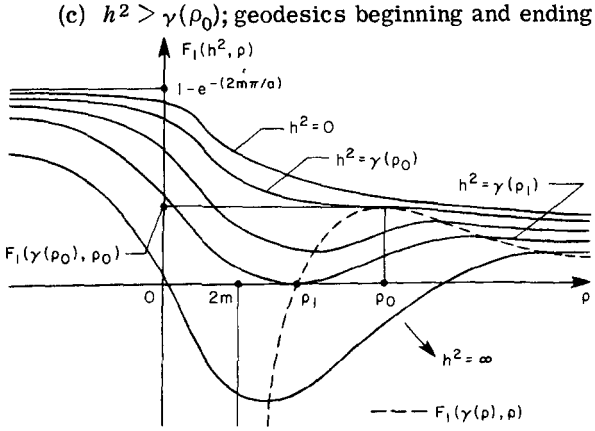

FIG. 8. The graphs of the functions $F_1(h^2,\rho)$ for various values of h^2. The dashed curve is the locus of critical points of the functions, which are minima to the left of, maxima to the right of $\rho_0[= 3m + (4m^2+n^2)^{1/2}]$. $F_1(\gamma(\rho_0),\rho)$ has a horizontal inflection point at ρ_0. As $\rho \to -\infty$, $F_1(h^2,\rho) \to 1 - e^{-(2m\pi/a)}$. As $\rho \to 2m+$, $F_1(\gamma(\rho),\rho) \to -\infty$. As $\rho \to \infty$, $F_1(h^2,\rho) \to 0$ and $F_1(\gamma(\rho),\rho) \to 0$. From the equation $\dot\rho^2 = 2E + F_1(h^2,\rho)$ one can find the turning points of timelike geodesics of $\mathfrak{M}_{m,n}$ by referring to this picture.

(iv) $-F_1(\gamma(\rho_0),\rho_0) < 2E < 0$. Let ρ^* and ρ^{**} denote the two roots of $2E + F_1(\gamma(\rho),\rho) = 0$, with $\rho^* < \rho^{**}$; let $\bar\rho^*$ denote the root of $2E + F_1(\gamma(\rho^*),\rho) = 0$ distinct from ρ^*, and define $\bar\rho^{**}$ analogously. Then $\bar\rho^{**} < \rho^* < \rho_0 < \rho^{**} < \bar\rho^*$, and, as E increases, ρ^* moves down from ρ_0- to ρ_1+, while ρ^{**} moves up from ρ_0+ to ∞.

(a) $0 \leq h^2 < \gamma(\rho^*)$; geodesics beginning and ending at $-\infty$, reaching highest points above $\bar\rho^*$.

(b) $h^2 = \gamma(\rho^*)$; geodesics with unstable circular orbit at ρ^*; geodesics beginning at $-\infty$, passing through the drainhole, and ending by spiraling up to the circular orbit, and vice versa; geodesics beginning by spiraling up from the circular orbit to highest points at $\bar\rho^*$, ending by spiraling back down to the circular orbit.

(c) $\gamma(\rho^*) < h^2 < \gamma(\rho^{**})$; geodesics beginning and ending at $-\infty$, reaching highest points between $\bar\rho^{**}$ and ρ^*; geodesics having stable bound orbits and periodic radial motions, with lowest points between ρ^* and ρ^{**} and highest points between ρ^{**} and $\bar\rho^*$.

(d) $h^2 = \gamma(\rho^{**})$; geodesics beginning and ending at $-\infty$, reaching highest points at $\bar\rho^{**}$; geodesics with stable circular orbit at ρ^{**}.

(e) $h^2 > \gamma(\rho^{**})$; geodesics beginning and ending at $-\infty$, reaching highest points below $\bar\rho^{**}$ which move down to $-\infty$ as h^2 increases.

(v) $2E \geq 0$. Let ρ^* be the root of $2E + F_1(\gamma(\rho),\rho) = 0$. As E increases, ρ^* moves down from ρ_1 to $2m+$.

(a) $0 \leq h^2 < \gamma(\rho^*)$; geodesics beginning at ∞, passing through the drainhole, ending at $-\infty$, and vice versa.

(b) $h^2 = \gamma(\rho^*)$; geodesics with unstable circular orbit at ρ^*; geodesics beginning at ∞, ending by spiraling down to the circular orbit, and vice versa; geodesics beginning at $-\infty$, passing through the drainhole, ending by spiraling up to the circular orbit, and vice versa.

(c) $h^2 > \gamma(\rho^*)$; geodesics beginning and ending at ∞, reaching lowest points which move up from just above ρ^* to ∞ as h^2 increases; geodesics beginning and ending at $-\infty$, reaching highest points which move down from just below ρ^* to $-\infty$ as h^2 increases.

Spacelike geodesics

When one examines the graphs (not presented here) of the functions $F_{-1}(h^2,\rho)$, taking into account that $2E = k^2 + 1 \geq 1$, he sees that the spacelike geodesics fall into three classes analogous to the three classes of timelike geodesics on which $2E \geq 0$. The principle observation of interest is that, as E increases, the circular orbits move up from m, where the drainhole is narrowest, to just below $2m$.

Capture of light rays and test particles

With good enough starts both light rays and test particles can coast upstream all the way to ∞, even if they begin as far down as $-\infty$ and procrastinate by spiraling as they go. The drainhole, then, is no "black hole" like the Schwarzschild singularity, surrounded by its one-way horizon. On the other hand, the drainhole does absorb many of the light rays and test particles that approach it from the upper side, by either capturing them or letting them pass through to the lower side. Perhaps the drainhole would qualify as a "gray hole."

For the Schwarzschild model with positive mass m' Darwin has established that no test particle orbit can have its pericenter as low as $3m'$.[19] The analogous proposition is true for the drainhole model: No such orbit has its lowest point or points as low as $2m$. Another aspect of the drainhole geodesics is that, although there are unstable, bound (actually circular), test particle orbits at ρ_0 and below, every such orbit that is stable must have its highest points above ρ_0 and its energy E greater than $-\frac{1}{2}F_1(\gamma(\rho_0),\rho_0)$. This property, reminiscent of a salient feature of quantum mechanical models of the hydrogen atom, also finds an analog in the Schwarzschild model.[19] It is worth noting that neither of these common properties depends upon the presence of a horizon, as the Schwarzschild manifold suggests it might.

In the Schwarzschild model the spatial cross sections Σ_t are flat, and the capture effects can be attributed to the gravitational field alone. In the drainhole model, however, some of the credit must go to the curvature of space around the drainhole, for, as we have seen, the effects persist, at short range, even when the gravitational field vanishes. Thus the drainhole with the flowing ether can be thought of as a first approximation to a geometrical model of a massive nuclear binding particle. On the other hand, one can use it in place of the Schwarzschild manifold to model the gravitational field of, for example, the sun. In this connection one can calculate that at large distances from the drainhole the bending of orbits caused by the curvature of space results in an increase in the precessions of orbital perihelia that is of higher infinitesimal order than the precessions themselves. This correction to the precessions differs both in order of magnitude and in sense from the corresponding correction in the Brans–Dicke scalar–tensor theory.[20]

IX. DISCUSSION

In the field equation (26), which the ether-flow, drainhole, particle model satisfies, the polarity of the coupling between the geometry of space–time and the scalar field is reversed from that which most physicists accept. I shall therefore review here some arguments in support of it, as well as one argument against it.

Justification of the coupling must rest ultimately on the reasonableness and usefulness that the space–time manifolds derivable from it possess as models of the physical world. The ether-flow, drainhole model derived from it has in common with the Schwarzschild manifold the useful ability to reproduce to within current observational tolerances the external gravitational field of a massive, nonrotating, spherically symmetric body. It does not have the Schwarzschild manifold's useless point singularity or the associated and equally useless incompleteness of geodesics. It also, reasonably if not usefully, has no horizon. In place of these dubious endowments it has several novelties of its own, whose reasonableness or unreasonableness, usefulness or unusefulness are yet to be determined. It ties together as two aspects of one entity the concept of negative (active) gravitational mass and that of positive,[21] at the same time hinting at a universal excess of the negative over the positive, in a ratio involving Dirac's outsized numbers. It stands as a clear indicator that within geometrodynamics, to use Wheeler's descriptive term for general relativity theory,[5] there is room at least for classical models of nuclear binding particles, with mass and without, if one will but relax the field equations enough to allow static negative curvatures of space. Finally, the drainhole suggests a dynamically topological mechanism for the dematerialization of such particles into traveling ripples in the fabric of space and also, because of time reversal symmetry, for their materialization out of these ripples.

Historically, Einstein took the coupling constant K in his field equations

$$R_{\mu\nu} - \tfrac{1}{2}Rg_{\mu\nu} = KT_{\mu\nu} \tag{60}$$

to be negative in order to satisfy the requirement that in the quasistatic, weak-field approximation these equations should approximate the content of the Poisson equation for the Newtonian gravitational potential V, an equation which reads $\nabla^2 V = 4\pi(\rho_M + \rho_E)$, where ρ_M is the density of mass and ρ_E the density of any other forms of energy that are thought to cause gravitational phenomena.[22] Einstein carried through his argument, however, only for the case in which the energy–momentum tensor components $T_{\mu\nu}$ arise solely from slowly moving dust of small but nonzero proper density, for which case $\rho_M \neq 0$ and $\rho_E = 0$. If in the other extreme ($\rho_M = 0, \rho_E \neq 0$) the only energy present is embodied in a scalar field ϕ of rest mass zero, associated with the Lagrangian density $(-g)^{1/2}\phi^{,\kappa}\phi_{,\kappa}$, then

$$T_{\mu\nu} = \phi_{,\mu}\phi_{,\nu} - \tfrac{1}{2}\phi^{,\kappa}\phi_{,\kappa}g_{\mu\nu}, \tag{61}$$

and Eq. (60) is equivalent to

$$R_{\mu\nu} = K\phi_{,\mu}\phi_{,\nu}. \tag{62}$$

In the quasistatic, weak-field approximation $V \approx \frac{1}{2}(g_{00} - 1)$, $R_{00} \approx -\nabla^2 V$, and $\phi_{,0}\phi_{,0} \approx 0$. Thus the Poisson equation whose content is approximated by $R_{00} = K\phi_{,0}\phi_{,0}$ is actually the Laplace equation $\nabla^2 V = -K \cdot 0$. The other field equations approximate to $0 = K \cdot 0$. Therefore, the requirement of correspondence with Newtonian theory yields in this case no information about K.

The failure of the correspondence requirement to fix the polarity of the scalar-field coupling leaves one free to apply other criteria to the task. It has seemed to me quite reasonable to eschew singularities and aim at a theory that will provide as a model for a mass particle at rest and alone in the universe a static space–time manifold that is geodesically complete and is asymptotic to a Schwarzschild manifold with nonzero mass parameter.[23] This criterion forces K to be positive in the variational principle (25), by way of the following argument.

Let us first take notice that every static and spherically symmetric line element is a special case of the line element (1). Indeed, every such line element can by a coordinate transformation be brought locally into the form

$$d\tau^2 = A(R)dT^2 - B(R)dR^2 - C^2(R)d\Omega^2, \tag{63}$$

with A, B, and C positive. Then a further transformation, changing only the radial coordinate, will take it to the form (1_S). The latter transformation is obtained by solving the differential equation $dR/d\rho = [A(R)B(R)]^{1/2}$ for R as an increasing, therefore invertible function of the new coordinate ρ. Finally, by using Eq. (21) in reverse, we can arrive at the form in Eq. (1).

Now let us recall that the discussion in Secs. VII and VIII established that if $\phi = \alpha(\rho)$, then the Euler equations associated with the variational principle (25) have the drainhole manifold as their only solution manifold that has a line element of the form (1), is geodesically complete, and is asymptotic to a Schwarzschild manifold with nonzero mass parameter, and, further, the drainhole is a solution only if $K > 0$. Finally, if $\phi = \alpha(t, \rho, \vartheta, \varphi)$, then one can without great difficulty see that the Euler equations imply that in fact ϕ depends only upon ρ, hence that the foregoing conclusion applies also in this case. To summarize, then, if and only if $K > 0$ does there exist a static, geodesically complete, and spherically symmetric space–time manifold that is asymptotic to a Schwarzschild manifold with nonzero mass parameter and that satisfies the variational principle (25) for some choice of the scalar field ϕ, and the drainhole manifold, with its numerous interesting and useful features, is the one.

Against the advantages that I have set forth for the nonstandard choice of coupling polarity one must array whatever implications it has that seem to be in disagreement with established theory. The only such implication that I have met is this: According to conventional interpretation, the scalar field, when coupled with nonstandard polarity to the geometry of space–time, must be accounted as having negative energy, contrary to the usual requirement of general relativity theory.[24] Specifically, with $K > 0$ one would say, following the usual convention, that the energy density of the scalar field is $-T_{00}$ as given in a physically significant reference frame by Eq. (61). Because T_{00} is positive definite in physically significant reference frames, such as local Lorentz frames, the energy density $-T_{00}$, hence also the total energy of the scalar field, would be negative definite. Perhaps this interpretation is correct. I have to confess that I have been unable to conclude or to be persuaded that the polarity of the coupling between a nonmaterial field and the geometry of space–time should determine or be determined by the positiveness or negativeness of the energy of that field. I prefer to postpone the question, looking forward to the day when we shall have a satisfactory, nonphenomenological unified field theory in which there appear no coupling constants whose polarity has to be assigned.

It is instructive to compare the scalar-field energy E_S in the drainhole model, be it positive or be it negative, with the energy E_G of the gravitational field. I take E_G to be the mass m, thereby remaining consistent with the view expressed in Sec. III and again in Sec. VIII that true gravity is generated only by internal motions of the ether and therefore vanishes when $f = 0$ (equivalently, when $m = 0$). For definiteness let us assume that $E_S \geq 0$. Then, after normalization by the conventional factor $(4\pi)^{-1}$,

$$E_S = (1/4\pi)\int_\Sigma T_{00}(-\det g_{mn})^{1/2} d^3x, \tag{64}$$

where Σ is a hyperspace orthogonal to the timelike Killing vector field $\partial/\partial T$, and $[x^\mu] = [T, \rho, \vartheta, \varphi]$, the form of the line element being therefore that in Eq. (1_S). Upon calculation of T_{00} and subsequent application of Eqs. (43), we have

$$E_S = \frac{1}{4\pi}\int_{-\pi}^{\pi}\int_0^{\pi}\int_{-\infty}^{\infty} \tfrac{1}{2}[(1-f^2)\alpha']^2 \times r^2(\sin\vartheta)(1-f^2)^{1/2} d\rho d\vartheta d\varphi$$

$$= \begin{cases} \dfrac{n\pi}{2} & \text{if } m = 0, \\ \dfrac{n^2}{2m}\left[1 - \exp\left(\dfrac{-m\pi}{(n^2-m^2)^{1/2}}\right)\right] & \text{if } m > 0. \end{cases} \tag{65}$$

One sees that, as m increases from 0 to n, E_S decreases continuously from $n\pi/2$ to $n/2$. Hence the amount of energy in the scalar field is essentially proportional to n, regardless of the amount m of gravitational energy, and it actually varies inversely with m. If $m \ll n$, then $E_S/E_G \approx n\pi/2m$. In the case of the numbers mentioned in Sec. VII, where m was approximately the mass of a proton, $E_S/E_G \approx 10^{19}$ if n is of the order of Planck's length, and $E_S/E_G \approx 10^{39}$ if n is near the classical electron radius. Here are two more occurrences of the ubiquitous Dirac numbers.[18] The large sizes of these ratios demonstrate that the scalar field (more generally, the curvature of space) is a promising agent for representing within general relativity theory natural phenomena much more energetic than gravity and having to do with particles of subatomic size.

ACKNOWLEDGMENTS

I express my fullest appreciation to István Ozsváth and to the referee for their aforementioned interest and comments,[24] and to Charles Misner for a useful critique and discussion of an early draft of this paper.

APPENDIX

This is a brief outline of the computational framework used in the body of the paper. The approach is that of Cartan.[25]

On the differentiable manifold $\mathfrak{M}$ the tangent vectors at the point P are thought of as those local differential operators on the scalar fields differentiable at P that for some coordinate system $[x^\mu]$ at P are linear combinations of the operators $(\partial/\partial x^\mu)(P)$. The tangent space at P is denoted by $\mathcal{T}^P$ and its dual, the space of tangent covectors at P, or cotangent space at P, by $\mathcal{T}_P$. The basis of $\mathcal{T}_P$ dual to the basis $\{(\partial/\partial x^\mu)(P)\}$ of $\mathcal{T}^P$ is denoted by $\{dx^\mu(P)\}$. After an obvious pattern the elements of the various tensor product spaces, such as $\mathcal{T}_P \otimes \mathcal{T}_P, \mathcal{T}_P \otimes \mathcal{T}^P, \mathcal{T}_P \otimes \mathcal{T}_P \otimes \mathcal{T}^P$, are distinguished among by use of the names cocotensor, cocontensor, cococontensor, and so on. The elements of $\mathcal{T}_P, \mathcal{T}_P \wedge \mathcal{T}_P, \mathcal{T}_P \wedge \mathcal{T}_P \wedge \mathcal{T}_P, \cdots$ are the 1-, 2-, 3-, $\cdots$ forms.

The connection forms of the covariant differentiation (affine connection) $\mathbf{d}$ on $\mathfrak{M}$, with respect to the frame system $\{e_\mu\}$ and its dual $\{\omega^\mu\}$, are the 1-forms $\{\omega_\mu{}^\kappa\}$ determined by either of the equations

$$de_\mu = \omega_\mu{}^\kappa \otimes e_\kappa \tag{A1}$$

and

$$\mathbf{d}\omega^\mu = -\omega_\kappa{}^\mu \otimes \omega^\kappa. \tag{A2}$$

If $p: I \to \mathfrak{M}$ is a path in $\mathfrak{M}$, I being its parameter interval, and u is a vector field on p (that is to say, u is a function on I, and $u(t) \in \mathcal{T}^{p(t)}$ for each t), then $\dot{u}$, the covariant derivative of u, is computed from

$$\dot{u} = [u^\mu e_\mu(p)]^\cdot = \dot{u}^\mu e_\mu(p) + u^\mu \mathbf{d}e_\mu(p)\dot{p} = [\dot{u}^\kappa + u^\mu \omega_\mu{}^\kappa(p)\dot{p}]e_\kappa(p). \tag{A3}$$

Let $\mathbf{d}_e$ be the exterior covariant differentiation based on $\mathbf{d}$, defined by saying that, for every co...co- or co...

cocontensor field V on $\mathfrak{M}$, $\mathbf{d}_e V$ is the totally skew-symmetric part of $\mathbf{d}V$. Then the torsion of $\mathbf{d}$ is the skew-symmetric cococontensor field T uniquely determined by the requirement that if v is a covector field, then

$$dv - \mathbf{d}_e v = vT, \tag{A4}$$

where d stands for (noncovariant) exterior differentiation and where juxtaposition means composition, e.g., $(vT)(u_1, u_2) = v(T(u_1, u_2))$. The curvature tensor field is the unique cococontensor field Θ that is skew-symmetric in the second and third slots and satisfies

$$\mathbf{d}_e^2 u = \Theta u - (\mathbf{d}u)T \tag{A5}$$

for every vector field u. In terms of $\{e_\mu\}$ and $\{\omega^\mu\}$,

$$\Theta = \omega^\kappa \otimes \Theta_\kappa{}^\mu \otimes e_\mu, \tag{A6}$$

the curvature 2-forms $\Theta_\kappa{}^\mu$ being given by

$$\begin{aligned}\Theta_\kappa{}^\mu &= d\omega_\kappa{}^\mu - \omega_\kappa{}^\lambda \wedge \omega_\lambda{}^\mu \\ &= -\tfrac{1}{2} R_\kappa{}^\mu{}_{\lambda\nu}(\omega^\nu \wedge \omega^\lambda),\end{aligned} \tag{A7}$$

where the $R_\kappa{}^\mu{}_{\lambda\nu}$ are the components of the Riemann-Christoffel curvature tensor field ($\equiv -2\Theta$). If $d\omega^\mu = C_\kappa{}^\mu{}_\lambda(\omega^\lambda \wedge \omega^\kappa)$, and $\omega_\kappa{}^\mu = \Gamma_\kappa{}^\mu{}_\lambda \omega^\lambda$, then

$$R_\kappa{}^\mu{}_{\lambda\nu} = 2(\Gamma_\kappa{}^\mu{}_{[\nu,\lambda]} + \Gamma_\kappa{}^\rho{}_{[\nu}\Gamma_{|\rho|}{}^\mu{}_{\lambda]} + \Gamma_\kappa{}^\mu{}_\rho C_{[\nu}{}^\rho{}_{\lambda]}). \tag{A8}$$

Here $C_{[\kappa}{}^\mu{}_{\lambda]} = \frac{1}{2}(C_\kappa{}^\mu{}_\lambda - C_\lambda{}^\mu{}_\kappa)$, and similarly for other square-bracketed pairs of indices. Contracting Θ in the second and fourth slots produces the contracted curvature tensor field Φ:

$$\Phi = \omega^\kappa \otimes \Theta_\kappa{}^\mu e_\mu = \omega^\kappa \otimes (-\tfrac{1}{2} R_{\kappa\lambda}\omega^\lambda), \tag{A9}$$

where $R_{\kappa\lambda} \equiv R_\kappa{}^\mu{}_{\lambda\mu}$, the components of the Ricci curvature tensor field ($\equiv -2\Phi$).

If $\mathbf{d}$ is required to be consistent with a metric G (any global, nondegenerate, symmetric cocotensor field on $\mathfrak{M}$), in the sense that $\mathbf{d}G = 0$, and to have torsion T (any global, skew-symmetric cococontensor field on $\mathfrak{M}$, given *a priori*), then $\mathbf{d}$ is uniquely determined. With respect to an orthonormal frame system $\{e_\mu\}$ and its dual $\{\omega^\mu\}$, the connection forms of $\mathbf{d}$ are easy to calculate using an algorithm of Misner.[26] It consists in solving for $[\omega_\kappa{}^\mu]$ the matrix equation

$$[\omega^\kappa] \wedge [\omega_\kappa{}^\mu] = [(C_\kappa{}^\mu{}_\lambda - T_\kappa{}^\mu{}_\lambda)(\omega^\lambda \wedge \omega^\kappa)], \tag{A10}$$

where $T = T_\kappa{}^\mu{}_\lambda(\omega^\lambda \wedge \omega^\kappa \otimes e_\mu)$, utilizing the symmetries and antisymmetries implied by $\omega_\mu{}^\kappa g_{\kappa\nu} + \omega_\nu{}^\kappa g_{\kappa\mu} = dg_{\mu\nu} = 0$. It is easiest to do this individually for each nonzero term on the right and then add the results.

---

[1]K. Schwarzschild, S.-B. Preuss, Akad. Wiss, Phys.-Math. Kl. 7, 189 (1916).

[2]A. Einstein and N. Rosen, Phys. Rev. **48**, 73 (1935).

[3]M. D. Kruskal, Phys. Rev. **119**, 1743 (1960).

[4]C. Fronsdal, Phys. Rev. **116**, 778 (1959).

[5]J. A. Wheeler, *Geometrodynamics* (Academic, New York, 1962).

[6]G. D. Birkhoff, *Relativity and Modern Physics* (Harvard U.P., Cambridge, Mass., 1923), p. 253.

[7]The symbol t is now overloaded, representing both coordinate function and constant. Such ambiguities are to be resolved by appeal to context.

[8]Here the argument ρ is suppressed for convenience's sake. It will be suppressed again on occasion.

[9]Einstein was convinced that the luminiferous ether, driven out of physicists' thoughts by the special theory of relativity, had returned as an essential feature of the general theory, intimately involved with gravity and only secondarily if at all connected with electromagnetism. He did not, however, recognize in it any degree of substantiality or of motility. In his essay "Relativity and the Ether" he wrote: "According to the general theory of relativity space is endowed with physical qualities; in this sense, therefore, an ether exists. In accordance with the general theory of relativity space without an ether is inconceivable.... But this ether must not be thought of as endowed with the properties characteristic of ponderable media, as composed of particles the motion of which can be followed; nor may the concept of motion be applied to it." [A. Einstein, *Essays in Science*, translated by A. Harris (Philosophical Library, New York, 1934).]

[10]Anyone willing to consider the ether-flow hypothesis as plausible might wish to ask whether, if the earth is conceivably a conglomeration of ether sinks and sources, the conventional interpretation of the Michelson–Morley experiment ought not to be revised.

[11]W. de Sitter, Proc. K. Ned. Akad. Wet. **19**, 1217 (1917); Proc. K. Ned. Akad. Wet. **20**, 229 (1917).

[12]A. Trautman has found before now that the Schwarzschild and the de Sitter line elements can be cast in the form (1). He has also applied the term "the ether" in these instances, but in a way that precludes any attribution of substantiality to the ether. [*Perspectives in Geometry and Relativity, Essays in Honor of Václav Hlavatý*, edited by B. Hoffman (Indiana U.P., Bloomington, 1966), p. 413.]

[13]One could raise the point that the Schwarzschild manifold, which satisfies Eq. (24) no matter what the coupling constant K (just let ϕ be a constant), already has a central hole, the "wormhole" of the Kruskal–Fronsdal extension. This hole, however, is not a permanent feature of the spatial cross sections; it develops into the Schwarzschild singularity when pursued in either temporal direction. Although at this point I have not said it in so many words, I have in mind that the particle models I seek shall be static, which alone would rule out the Schwarzschild manifold, even without the completeness requirement. One might also wish to say that a Schwarzschild interior solution, properly matched to an exterior solution, would provide a model of just the kind I want, without even introducing the scalar field ϕ. I would have to reply that such a manifold can only represent a mass *body*, not a particle. To electromagnetic geons as particle models (J. A. Wheeler, Ref. 5) my foremost objection would be that they are not static.

[14]A. I. Janis, E. T. Newman, and J. Winicour, Phys. Rev. Lett. **20**, 878 (1968).

[15]R. Penney, Phys. Rev. **174**, 1578 (1968).

[16]O. Bergmann and R. Leipnik, Phys. Rev. **107**, 1157 (1957). In fact these authors formulated the full set of solutions under discussion here. However, they ruled out for "physical reasons" the very case to which I attach the greatest physical significance, Case III. Handicapped by a restrictive coordinate system, they nevertheless were able to identify most parts of the solution manifolds in Cases I and II.

[17]H. Yilmaz, *Introduction to the Theory of Relativity and the Principles of Modern Physics* (Blaisdell, New York, 1965), p. 176. The $R_{\kappa\lambda}$ used by Yilmaz are the negatives of those appearing here. This has the consequence that his Eq. (18.2), although seemingly equivalent to the present Eq. (26), actually involves the coupling of opposite polarity and therefore is not satisfied by his line element (18.3) unless ϕ is constant. See also H. Yilmaz, Phys. Rev. **111**, 1417 (1958).

[18]P. A. M. Dirac, Proc. R. Soc. Lond. **165**, 199 (1938); Nature (Lond.) **139**, 323 (1937).

[19]C. Darwin, Proc. R. Soc. A **249**, 180 (1959).

[20]C. Brans and R. H. Dicke, Phys. Rev. **124**, 925 (1961).

[21]These concepts are also combined, in a somewhat analogous fashion, in the Kerr solution of the Einstein vacuum field

equations. [B. Carter, Phys. Rev. **174**, 1559 (1968).]

[22]A. Einstein, Ann. Phys. (Leipz.) **49**, 769 (1916), or see *The Principle of Relativity* (Dover, New York, 1923), p. 157.

[23]There is an earlier instance in which to meet a new criterion the polarity of the coupling in Eq. (60) was partially reversed. Einstein and Rosen reversed it for the coupling of space–time geometry to the electromagnetic field in order to represent an elementary charged mass particle as one of their "bridges" (A. Einstein and N. Rosen, Ref. 2).

[24]At the Relativity Conference in the Midwest, Cincinnati, Ohio, June 2–6, 1969, István Ozsváth was good enough to raise and to discuss with me this issue. Also, the referee has pressed it upon me in a friendly manner. The referee's argument, based upon energy conservation, is persuasive but not, to my mind, conclusive. For this reason I shall not recapitulate it here; neither shall I at the end profess to have resolved the issue

[25]É. Cartan, *Leçons sur la géométrie des espaces de Riemann* (Gauthier-Villars, Paris, 1946), 2nd ed. See also H. Flanders, *Differential Forms with Applications to the Physical Sciences* (Academic, New York, 1963), and R. L. Bishop and S. I. Goldberg, *Tensor Analysis on Manifolds* (Macmillan, New York, 1968).

[26]C. W. Misner, J. Math. Phys. **4**, 924 (1963).

Spectral properties of the linearized Balescu–Lenard operator

Allan H. Merchant and Richard L. Liboff

Department of Physics and Schools of Applied and Engineering Physics and Electrical Engineering, Cornell University, Ithaca, New York 14850

(Received 23 June 1972; revised manuscript received 21 August 1972)

The spectrum of the linearized Balescu–Lenard operator is studied in detail. It is found to be continuous, to range from zero to minus infinity, and to have no point spectrum. Analytic expressions are obtained for the $l=1$ spherical harmonic eigenfunctions in the velocity domain $\gtrsim 2.5$, where x is microscopic speed, nondimensionalized through the thermal speed. Sketches showing the typical behavior for all x of this l-mode eigenfunction are also given.

I. INTRODUCTION

In this paper we study properties of the linearized Balescu–Lenard equation. Previous similar works have concentrated on the Boltzmann equation,[1–3] reactor transport equation,[4] and the Fokker Planck (FP) equation.[5,6] Wu[7] has studied the spectrum of relaxation times of the linearized Balescu–Lenard (BL) equation by approximating the linearized BL operator with a finite-dimensional operator through the use of Laguerre polynomials. The mathematical properties of the linearized BL equation are quite similar to those of the linearized FP equation, their differences presiding primarily in the dynamic screening contribution to the differential and integral coefficients present in the BL equation.[8] The present work follows closely the work of Lewis.[5] Both Lewis[5] and Su[6] find the spectrum of the FP operator to be continuous from zero to minus infinity. In addition, Su refers to the existence of a point spectrum. In our analysis we find no such point spectrum save for the fivefold conservation degeneracy at the origin. It is shown below that the BL operator also contains a continuous spectrum from zero to minus infinity. This as well as the form of the expansion of the BL operator in terms of its own eigenfunctions is established in Sec. II.

In Sec. III, analytic expressions for the eigenfunctions of the BL operator are constructed in terms of velocity x, nondimensionalized through the thermal speed. For $x \gtrsim 2.5$, these eigenfunctions are uniformly valid for all eigenvalue λ. Numerical procedures for continuation of these eigenfunctions into the domain $x < 2.5$ are described. Properties of these matching formulas are derived for small λ. The related eigenfunctions are relevant to times much greater than the plasma relaxation time.

II. THE COMPONENTS OF THE LINEARIZED KINETIC OPERATOR

The kinetic equation which determines the time development of the velocity distribution function of a high temperature, dilute, one-component spacially homogeneous plasma is the Balescu–Lenard equation

$$\frac{\partial F}{\partial t} = \frac{8\pi^4 e^4}{m^2}\frac{\partial}{\partial v_r} \times \iint d^3u d^3k \frac{k_r k_s |\Phi(\mathbf{k})|^2 \delta(\mathbf{k}\cdot(\mathbf{v}-\mathbf{u}))}{|D^+(|k|, \hat{\mathbf{k}}\cdot\mathbf{v})|^2}\left(\frac{\partial}{\partial v_s} - \frac{\partial}{\partial u_s}\right) F(\mathbf{v})F(\mathbf{u}), \tag{1}$$

where e and m are, respectively, the charge and mass of the particles; $F(\mathbf{v})$ is the velocity distribution normalized so that $\int d^3v F(\mathbf{v}) = n_0 \equiv$ number density; $\Phi(k)$ is the Fourier transform of the interaction potential, $\hat{\mathbf{k}} = \mathbf{k}/|\mathbf{k}|$; and

$$D^+(k, \hat{\mathbf{k}}\cdot\mathbf{v}) = 1 + \frac{8\pi^3 e^2}{m}\Phi(k)\left(P\colon \int d^3u \frac{\hat{\mathbf{k}}\cdot\nabla_u F}{\hat{\mathbf{k}}\cdot(\mathbf{v}-\mathbf{u})} - i\pi\int d^3u \delta(\hat{\mathbf{k}}\cdot(\mathbf{v}-\mathbf{u}))\hat{\mathbf{k}}\cdot\nabla_u F\right), \tag{1'}$$

where P: denotes that the principal value is to be taken. This equation applies to a plasma satisfying the following conditions: (a) $n_0 d^3 \gg 1$, where d is the Debye distance, $d^2 = k_B T/4\pi n_0 e^2$, while k_B is Boltzmann's constant and T is the temperature; (b) the dielectric function $D^+(k, Z)$ has no zeros in the upper half Z plane.

Consider the following linearization of $F(\mathbf{v}, t)$:

$$F(\mathbf{v}, t) = (n_0 C^{-3}/\pi^{3/2})e^{-v^2/C^2}[1 + f(\mathbf{v})], \tag{2}$$

where

$$C^2 = 2k_B T/m.$$

Introduce the dimensionless velocity $\mathbf{x} = \mathbf{v}/C$, insert (2) into Eq. (1), and then drop terms quadratic in f. We obtain

$$\frac{\partial f}{\partial t} = \frac{1}{\tau}e^{x^2}\frac{\partial}{\partial x_r}\left(e^{-x^2}P_{rs}(\mathbf{x})\frac{\partial f}{\partial x_s} - e^{-x^2}\iint d^3x' d\Omega_k \hat{k}_r \hat{k}_s J(\hat{\mathbf{k}}\cdot\mathbf{x}) \times \delta(\hat{\mathbf{k}}\cdot(\mathbf{x}'-\mathbf{x}))e^{-x'^2}\frac{\partial f}{\partial x_s'}\right), \tag{3}$$

where the tensor

$$P_{rs}(\mathbf{x}) = \delta_{rs}Q(x) + [p(x) - Q(x)]x_r x_s/x^2, \tag{3'a}$$

$$p(x) = (4\sqrt{\pi}/x^3)\int_0^x s^2 e^{-s^2} J(s)ds, \tag{3'b}$$

$$Q(x) = (2\sqrt{\pi}/x)\int_0^x e^{-s^2}J(s)ds - \tfrac{1}{2}p(x),$$

$$\tau = 8\pi^2 n_0 C^3/\omega_p^4, \quad \omega_p^2 = 4\pi n_0 e^2/m \tag{3'c}$$

and where the function

$$J(s) \equiv 4\pi^4 \int_0^\infty \frac{k^3|\Phi(k)|^2 dk}{|D_0^+(k, Cs)|^2}. \tag{4}$$

Here $D_0^+(k, Cs)$ is the dielectric function as determined by the Maxwellian

$$D_0^+(k, Cs) = 1 + \frac{k_D^2}{k^2}\left(1 - 2se^{-s^2}\int_0^s e^{t^2}dt + i\sqrt{\pi}\, se^{-s^2}\right) \equiv 1 + \frac{k_D^2}{k^2}[M_1(s) + iM_2(s)], \tag{5}$$

and

$$k_D^2 = 4\pi n_0 e^2/k_B T.$$

Thus Eq. (3) can be written

$$\frac{\partial f}{\partial t} = \frac{1}{\tau}\hat{O}_{BL} f, \tag{6}$$

where $\hat{O}_{BL}$ is a linear integro-differential operator. Equation (6) can be studied by means of the formal tools of functional analysis if we make certain restrictions on f. These restrictions are that:

(a) f belong to the Hilbert space H in which the scalar product is defined by

$$(f|g) \equiv \int \bar{g}(\mathbf{x}) f(\mathbf{x}) e^{-x^2} d^3x;$$

(b) f satisfy

$$\lim_{x\to 0} \sqrt{x} f(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} \frac{e^{-x^2/2}}{x} \frac{\partial f}{\partial \mathbf{x}} = 0.$$

These two conditions are more restrictive than those imposed by physics, because f need not be in H in order that the fluid dynamic variables be finite. Furthermore, the conservation laws are satisfied if only

$$\lim_{x\to 0} x f(x) = 0, \quad \lim_{x\to\infty} x e^{-x^2} \frac{\partial f}{\partial \mathbf{x}} = 0.$$

If conditions (a) and (b) are satisfied, then it is easy to show that $\hat{O}_{BL}$ is symmetric and nonpositive definite. These statements hold for any $J(x)$ that obeys

$$e^{-x^2} J(x) \leq O(x^{-3}) \quad \text{as } x \to \infty, \quad J(0) < \infty.$$

Furthermore, if the domain of definition of $\hat{O}_{BL}$ is the class of f's having absolutely continuous first partial derivatives, then $\hat{O}_{BL}$ can be shown to be self-adjoint. The evolution operator $\exp(t\hat{O}_{BL})$ will be a bounded operator, and, as a consequence, conditions (a) and (b) will be preserved in time if they hold at the initial time and if the inhomogeneous driving term (if any) in the kinetic equation can be represented by a function which is an element of H.

Since all vectors in the expression for the operator $\hat{O}_{BL}$ are integrated out except the independent variable $\mathbf{x}$, $\hat{O}_{BL}$ is rotationally invariant and hence can be reduced to the spherical harmonic subspaces. Thus we may write

$$f(\mathbf{x}) = Y_{lm}(\Omega_x) h(x),$$

where Ω_x is the polar angle of $\mathbf{x}$ referred to an arbitrary reference frame and $x = |\mathbf{x}|$. We then may write,

$$\hat{O}_{BL} Y_{lm} h = Y_{lm}(\Omega_x) \hat{\bar{O}}_{BL} h$$

where $\hat{\bar{O}}_{BL}$ is an integro-differential operator in the single variable x.

It proves convenient to make the additional change of variable

$$h(x) \equiv (e^{x^2/2}/x) y(x)$$

so that

$$f(\mathbf{x}) = Y_{lm}(\Omega_x) \frac{e^{x^2/2}}{x} y(x).$$

The kinetic equation for y becomes, upon nondimensionalizing the time through τ,

$$\frac{\partial y}{\partial t} = \frac{\partial}{\partial x}\left(p \frac{\partial y}{\partial x}\right) + \left(V(x) - \frac{l(l+1)Q(x)}{x^2}\right) y + \int_0^\infty K_l(x,s) y(s) ds, \tag{7}$$

or

$$\frac{\partial y}{\partial t} = \hat{O}_l y. \tag{8}$$

In Eq. (7) p and Q are given by Eqs. (3) and (3'), and

$$V(x) \equiv (3 - x^2)p + (x - x^{-1}) \frac{dp}{dx} + 8\sqrt{\pi}\, e^{-x^2} J(x). \tag{9}$$

The kernel $K_l(x,s)$ can be written

$$K_l(x,s) = 8\sqrt{\pi}\, e^{-(x^2+s^2)/2} \times \sum_{n=n_{\min}}^{l/2} \left(\frac{G_{ln}(s)}{x^{2n}} \Theta(x-s) + \frac{G_{ln}(x)}{s^{2n}} \Theta(s-x)\right), \tag{10}$$

where

$$\Theta(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases}, \qquad n_{\min} = \begin{cases} 0, & \text{for even } l \\ \frac{1}{2}, & \text{for odd } l \end{cases}$$

and

$$G_{ln}(s) = 4\left(-\frac{A_{ln}}{2}\left(s + \frac{n}{s}\right) s^{2n} J(s) + \sum_{k'=n_{\min}}^{l/2} \frac{A_{ln} A_{lk'}}{s^{2k'}} [nk'h(2n + 2k' - 2, s) + (n + k')h(2n + 2k', s) + h(2n + 2k' + 2, s)]\right). \tag{11}$$

Here

$$h(\nu, s) = \int_0^s z^\nu J(z) dz, \tag{11'}$$

and the A_{lk} are the coefficients in the Legendre polynomial

$$P_l(x) = \sum_{k=n_{\min}}^{l/2} x^{2k} A_{lk}.$$

With the change of dependent variable as given above, the function y is in L^2 space on $(0, \infty)$. The boundary conditions on y become

$$\lim_{x\to\infty} x^{-1}\left(y + x^{-1}\frac{dy}{dx}\right) = 0 \quad \text{and} \quad \lim_{x\to 0} \frac{y}{\sqrt{x}} = 0. \tag{12}$$

In order to proceed, we must evaluate the function $J(x)$ given by Eq. (4). With $\Phi(k) = (2\pi^2 k^2)^{-1}$ the integral in J diverges at the upper limit. To obviate the (logarithmic) singularity,[9] the integral is cut off at $k_m = k_B T/e^2$, whose inverse is the distance of closest approach of a particle with energy $k_B T$ and zero impact parameter to a potential e^2/r.

The integral in J is evaluated by choosing the appropriate contour in the complex k^2/k_D^2 plane to obtain

$$J(x) = \tfrac{1}{4}\ln\left(\frac{(M_1 + \Gamma^2)^2 + M_2^2}{M_1^2 + M_2^2}\right) - \frac{M_1}{2|M_2|}\left[\arg\left(\frac{M_1}{|M_2|} + i\right) - \arg\left(\frac{M_1 + \Gamma^2}{|M_2|} + i\right)\right], \tag{13}$$

where

$$\arg(re^{i\theta}) \equiv \theta, \quad 0 \leq \arg Z < 2\pi,$$

and

$$\Gamma = \frac{k_m}{k_D} = 4\pi n_0 d^3 \gg 1.$$

We see from Eq. (5) that

$$|M_1| \leq 0(1), \quad |M_2| \lesssim 0(1),$$

and for $x \gtrsim 2$

$$|M_1|/|M_2| \gg 1.$$

Thus, since $\Gamma \gg 1$,

$$J(x) \sim \begin{cases} \ln\Gamma - \frac{1}{2}, & x \to 0, \\ \ln\Gamma + \ln x + \frac{1}{4}\sqrt{\pi}\, e^{x^2}/x^3, & x \gtrsim 2. \end{cases} \tag{14}$$

Redefining $\ln\Gamma$ slightly so that $\ln\Gamma - \frac{1}{2} \to \ln\Gamma$, Eq. (14) can be approximated by the tractable formula

$$J(x) = \ln\Gamma + \tfrac{1}{4}\sqrt{\pi}\, x e^{x^2}/(1 + x^4). \tag{15}$$

The specific nature of the interparticle interaction is contained entirely in the J function of Eq. (4). If we were to put $D_0^+ = 1$, the integral would then have to be cut off at $k = k_D$ as well as the upper limit, and we would get $J = \ln\Gamma$. Equation (3) with this J would then be just the linearized Landau equation. If we were to set $D_0^+ = 1 + k_D^2/k^2$, we would obtain the Landau equation corresponding to the Debye potential of interaction $r^{-1}e^{-k_D r}$ between particles. This D_0^+ would give $J \simeq \ln\Gamma$. The Balescu–Lenard equation includes effects of the finite response time of the screening between particles. Using Eq. (5) for D_0^+ thus gives (approximately) Eq. (15) for J.

It can readily be shown that the number, momentum, and energy densities are conserved (for a spacially homogeneous system) by Eq. (3) for any $J(x)$ function which has the property,

$$e^{-x^2}J(x) \leq O(x^{-3}) \quad \text{as } x \to \infty.$$

Using Eq. (15), we can now evaluate the functions p, Q, and V defined in Eqs. (3):

$$p(x) = 2\sqrt{\pi}\ln\Gamma\left(\frac{\sqrt{\pi}}{2x^3}\operatorname{erf}(x) - \frac{e^{-x^2}}{x^2}\right) + \frac{\pi}{4x^3}\ln(1 + x^4), \tag{16a}$$

and

$$Q(x) = \frac{\pi\ln\Gamma}{x}\operatorname{erf}(x) - \frac{p}{2} + \frac{\pi}{4x}\arctan x^2. \tag{16b}$$

As $x \to \infty$, we have

$$p(x) \sim \frac{\pi\ln\Gamma x}{x^3}, \quad V(x) \sim -\frac{\pi\ln\Gamma x}{x}, \quad \text{and} \quad Q(x) \sim \frac{\pi\ln\Gamma + \pi^2/8}{x}. \tag{16c}$$

III. SPECTRAL PROPERTIES

We now study the spectral properties of the operator $\hat{O}_l$ [defined by Eqs. (7), (8)]. The limit points of a self-adjoint operator are unchanged if we add to it a completely continuous operator.[10] In order to apply this theorem to the study of $\hat{O}_l$, we must show that the kernel K_l is square integrable (and hence the integral operator is completely continuous), and then we must study the spectrum of the differential operator alone. The latter is continuous from zero to minus infinity. Consequently, the spectrum of $\hat{O}_l$ is continuous from zero to minus infinity.

Before so proceeding, let us set [see Eq. (7)]

$$\hat{O}_l y = (\hat{D}_l + \hat{K}_l)y \equiv \hat{D}_l y + \int_0^\infty ds\, K_l(x, s)y(s). \tag{16d}$$

To show that

$$\int_0^\infty \int_0^\infty |K_l(x, s)|^2 dx ds < \infty,$$

we note that the least convergent terms in this integral come from the large x behavior of $J(x)$. Near the origin ($x = 0, s = 0$) $J \to \ln\Gamma$; hence $K_l(x, s)$ for $x, s \to 0$ goes over into the "Landau equation" form, which is already known to be square integrable near the origin. With $h(\nu, s)$ given by Eq. (11′) we can see that, for large s,

$$h(\nu, s) \sim \tfrac{1}{2}s^{\nu-4}e^{s^2}.$$

Thus for large x, s the largest terms in the double sum given by Eqs. (10) and (11) are

$$K_l(x, s) \sim e^{-(x^2+s^2)/2} \sum_{n=n_{\min}}^{l/2} \times C_{ln}\left(\frac{s^{2n-2}e^{s^2}}{x^{2n}}\Theta(x - s) + \frac{x^{2n-2}e^{x^2}}{s^{2n}}\Theta(s - x)\right), \tag{17}$$

where the C_{ln} are constants. Thus, since

$$\left(\frac{s}{x}\right)^{2n}\Theta(x - s) < 1,$$

$$|K_{ls}(x, s)| \leq \left(s^{-2}e^{-x^2/2+s^2/2}\Theta(x - s) + x^{-2}e^{-s^2/2+x^2/2}\Theta(s - x)\right) \sum_{n=n_{\min}}^{l/2} |C_{ln}|, \tag{17′a}$$

and, for $a \gg 1$,

$$\int_0^\infty\int_0^\infty |K_l(x, s)|^2 dx ds = \int_0^a\int_0^a |K_l(x, s)|^2 dx ds + 2\int_a^\infty dx \int_0^a |K_l(x, s)|^2 ds + \int_a^\infty\int_a^\infty |K_l(x, s)|^2 dx ds. \tag{17′b}$$

The first term on the right in Eq. (17′b) is convergent. In the second term we have $x > s$ for the entirety of the domain of integration, so that from Eq. (10) the function

$$\int_0^a |K_l(x, s)|^2 ds \propto \frac{e^{-x^2}}{x^{4n_{\min}}}, \quad x \gg 1.$$

The last term, by Eq. (17′b), is

$$\int_a^\infty\int_a^\infty |K_l(x, s)|^2 dx ds \leq 2\left(\sum_{n=n_{\min}}^{l/2} |C_{ln}|\right)^2 \times \int_a^\infty \frac{e^{s^2}ds}{s^4}\int_s^\infty e^{-x^2}dx \leq 2\left(\sum_{n=n_{\min}}^{l/2} |C_{ln}|\right)^2 \int_a^\infty \frac{ds}{s^5} < \infty.$$

Thus the kernel $K_l(x, s)$ is square integrable over the upper right xs plane.

We now show that the spectrum of the differential operator is continuous from zero to minus infinity. The study of the spectral properties of the differential operator is made with the aid of the theory of symmetric differential operators on an infinite domain.[11] We may summarize

the relevant results of this theory in the following theorem.

Theorem: If the differential operator, defined on $0 \le x < \infty$, can be written in the form

$$\hat{\Xi}y = (py')' + qy, \quad \text{with } p(x) > 0,$$

and, if it is of a certain type (defined below), we may write the following expansion of any square integrable function on $[0, \infty]$:

$$f(x) = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \phi_j(x,\lambda) d\rho_{jk}(\lambda) \int_0^{\infty} \phi_k(s,\lambda) f(s) ds, \quad (18)$$

where

(a) The set $\{\phi_j\}$ are any two linearly independent solutions of

$$\hat{\Xi}\phi - \lambda\phi = 0. \quad (18'a)$$

(b) The matrix function $\rho_{jk}(\lambda)$ is of bounded variation and $\rho_{jk}(\lambda) - \rho_{jk}(\mu)$ is nonnegative definite if $\lambda > \mu$. ρ_{jk} satisfies $\rho_{jk}(\lambda) - \rho_{jk}(\mu) = \lim_{\epsilon \to 0^+} \pi^{-1} \int_{\mu}^{\lambda} \mathrm{Im} M_{jk}(\nu + i\epsilon) d\nu$ at points λ, μ of continuity of ρ_{jk}.

(c) The matrix $M_{jk}(\lambda)$ is given by

$$M_{11} = [m_0(\lambda) - m_\infty(\lambda)]^{-1},$$
$$M_{12} = M_{21} = \tfrac{1}{2}[m_0(\lambda) + m_\infty(\lambda)]/[m_0(\lambda) - m_\infty(\lambda)], \quad (18'b)$$
$$M_{22} = m_0(\lambda) m_\infty(\lambda)/[m_0(\lambda) - m_\infty(\lambda)],$$

where, finally,

(d) With $\mathrm{Im}\lambda \neq 0$, $m_0(\lambda)$ and $m_\infty(\lambda)$ are those unique coefficients having the property that $\int_a^{\infty} |\phi_1(x,\lambda) + m_\infty(\lambda)\phi_2(x,\lambda)|^2 dx < \infty$ and $\int_0^a |\phi_1(x,\lambda) + m_0(\lambda)\phi_2(x,\lambda)|^2 dx < \infty$ with a an arbitrary finite number.

This completes the statement of the theorem.

The type of differential operator referred to in this discussion is that having the property indicated in (d), namely, that, for $\mathrm{Im}\lambda \neq 0$, only one linear combination ϕ of ϕ_1 and ϕ_2 has

$$\int_a^{\infty} |\phi|^2 dx < \infty,$$

and only one (in general not the same) has

$$\int_0^a |\phi|^2 dx < \infty.$$

The differential operator $\hat{D}_l$ may readily be shown to be of this type by considering the asymptotic form of the solutions to $\hat{D}\phi = \lambda\phi$ near the origin and at infinity. Near the origin there are two solutions which behave as x^{-l} and x^{l+1}, and the asymptotic solutions for large x behave as

$$\left(\frac{x^3}{\ln \Gamma x}\right)^{1/4} \exp \pm i\sqrt{-\lambda} \int_{x_0}^{x} \frac{ds}{\sqrt{p(s)}}.$$

(For $l = 0$, both solutions are square integrable at the origin. In this case we impose a homogeneous boundary condition at the origin and the above discussion is replaced by the simpler theory[11] of a differential operator with a singularity only at infinity.)

As just shown, the operator $\hat{D}_l$ satisfies the hypotheses of the above theorem. In order to use this theorem to investigate the spectrum of $\hat{D}_l$, it is necessary to calculate the components of the matrix function ρ_{ij}.

The components of the related matrix $M_{ij}(\lambda)$ may readily be calculated using Eq. (18'b).[5] For each λ in the integrand of Eq. (18), the two functions ϕ_1 and ϕ_2 are any linearly independent solutions of Eq. (18'a). If we choose $\phi_2(x,\lambda)$ to be that solution which behaves at the origin as $\phi_2 \to x^{l+1}$, then $m_0(\lambda) = \infty$ and hence the matrix

$$d\rho_{jk} = \lim_{\mathrm{Im}\lambda \to 0^+} \pi^{-1} \mathrm{Im} M_{jk} = \lim_{\mathrm{Im}\lambda \to 0^+} \pi^{-1} \begin{pmatrix} 0 & 0 \\ 0 & m_\infty(\lambda) \end{pmatrix}.$$

As $x \to \infty$, a phase integral analysis shows that

$$\phi_2(x,\lambda) \to A_\lambda \left(\frac{x^3}{\ln \Gamma x}\right)^{1/4} \cos\left(\sqrt{-\lambda} \int_{x_0}^{x} \frac{ds}{\sqrt{p(s)}} + \theta(\lambda)\right).$$

One calculates that

$$\mathrm{Im}\, m_\infty(\lambda) = \begin{cases} (A_\lambda^2 \sqrt{-\lambda})^{-1}, & \mathrm{Re}\lambda < 0, \\ 0, & \mathrm{Re}\lambda > 0. \end{cases}$$

The spectral density is

$$\frac{d\rho}{d\lambda} = (\pi A_\lambda^2 \sqrt{-\lambda})^{-1}.$$

Thus Eq. (18) becomes

$$f(x) = \pi^{-1} \int_{-\infty}^{0} \frac{d\lambda}{\sqrt{-\lambda}} \frac{1}{A_\lambda^2} y(x,\lambda) \int_0^{\infty} y(s,\lambda) f(s) ds, \quad (19)$$

where $y(x,\lambda)$ is the eigenfunction of $\hat{D}$ which is regular at the origin and has the eigenvalue λ.

The function A_λ is the asymptotic amplitude of that solution which started out at the origin as x^{l+1}. Clearly A_λ is a continuous and nonvanishing function of λ on any finite interval of the negative λ axis. This means that the spectral density $d\rho/d\lambda$ neither vanishes nor is singular for $\lambda < 0$. Thus in particular the spectrum of the differential operator has no gaps (vanishing of $d\rho/d\lambda$). Therefore, the above referred-to theorem relating to the limit points of a self-adjoint operator tells us that the spectrum of the full operator has no gaps from zero to minus infinity and is empty on the positive λ axis.

A. Absence of a point spectrum in $\hat{O}_l$

In this section we will demonstrate that $\hat{O}_l$ has no point spectrum, save for the fivefold conservation degeneracy at $\lambda = 0$. An eigenvalue is an element of a point spectrum if the corresponding eigenfunction is in L^2.[12]

The argument is as follows. Let y be an eigenfunction of $\hat{O}_l$; that is, suppose

$$\hat{O}_l y - \lambda y = 0, \quad \lim_{x \to 0} (y/\sqrt{x}) = 0.$$

Because of the nature of the integral operator, the equation

$$\hat{O}_l y - \lambda y = 0$$

can be converted into a pure linear differential equation given by $\hat{\Delta}^{(N)} y = 0$, where N is the order of this equation. However, an arbitrary solution of $\hat{\Delta}^{(N)}\Psi = 0$ is not a solution of $\hat{O}_l\Psi - \lambda\Psi = 0$. We will show that if $y(x)$ is

required to be square integrable, this generates more constraints than the number of linearly independent solutions of $\hat{\Delta}^{(N)}\Psi = 0$.

Consider first the case $l = 0$ or $l = 1$. Then $n_{\min} = l/2$, and $\hat{O}_l y$ may be written in the form

$$\hat{O}_l y = \hat{D}_l y + 8\pi^2 f_l(x) \int_0^x g_l(s)y(s)ds + 8\pi^2 g_l(x) \times \int_x^\infty f_l(s)y(s)ds, \quad (20a)$$

with g_l and f_l implied by Eqs (10) and (11).
The pure fourth-order differential operator obtained from Eq. (20a) is

$$\hat{\Delta}^{(4)} = \hat{E}(\hat{O}_l - \lambda), \quad (20b)$$

where, for arbitrary F,

$$\hat{E}F = g\{[(g/f)']^{-1}(F/f)'\}'.$$

Since $\hat{E}(af + bg) = 0$ (subscript l has been dropped from f_l and g_l), it is clear that if χ is an arbitrary solution of $\hat{\Delta}^{(4)}\chi = 0$, then $(\hat{O}_l - \lambda)\chi = \alpha f + \beta g$, where α and β are constants.

If one sets

$$y = \sum_{i=1}^{4} A_i \Psi_i,$$

where the set $\{\Psi_i\}$ are four linearly independent solutions of $\Delta^{(4)}\Psi = 0$, then, for y, so written, to be an eigenfunction of $\hat{O}_l$, it is necessary that

$$\sum_{i=1}^{4} \alpha_i A_i = 0, \quad (21a)$$

and

$$\sum_{i=1}^{4} \beta_i A_i = 0, \quad (21b)$$

where α_i and β_i and defined by

$$(\hat{O}_l - \lambda)\Psi_i = \alpha_i f + \beta_i g.$$

Equations (21a), (21b) impose two constraints on the A_i coefficients. Two more constraints result if we demand the eigenfunction y to be square integrable at the upper limit. As will be seen in Sec. IV, for large x, two solutions of $\Delta^{(4)}\Psi = 0$ behave as

$$\Psi \simeq \frac{x^{3/4}}{(\ln\Gamma x)^{1/4}} \exp\left(\pm i\sqrt{-\lambda} \int_{x_0}^{x} \frac{ds}{\sqrt{p(s)}}\right).$$

Both of these solutions must be eliminated if the eigenfunction is to be square integrable. Finally there is a fifth condition, related to the symmetry of $\hat{O}_l$, which must be imposed on the A_i, namely, the boundary condition [Eq. (12)]

$$\lim_{x\to 0} (y/\sqrt{x}) = 0.$$

There are five homogeneous conditions on the four unknowns A_i, and the A_i are overdetermined with no remaining free parameters. This completes the demonstration that for $l = 0$ and $l = 1$ there is no point component to the spectrum for finite λ.

The proof maintains for arbitrary l because the integral operator is always of the form of a finite sum (from Eqs. (10) and (11),

$$\int_0^\infty K_l(x,s)y(s)ds = 8\pi^2 \sum_{j=1}^{N_l} \times \left(f_{jl}(x) \int_0^x g_{jl}(s)y(s)ds + g_{jl}(x) \int_x^\infty f_{jl}(s)y(s)ds\right).$$

Each differentiation executed to obtain the pure $(2 + 2N_l)$th-order equation carries with it a condition which is the generalization of Eqs. (21a), (21b). With the boundary condition at the origin, Eq. (12), one always obtains one more homogeneous equation for the A_i coefficients than the number of unknowns. We note that the square integrable function which Su[6] constructs, while a solution $\hat{\Delta}^{(4)}\Psi = 0$, is not further constrained to be an eigenstate of $\hat{O}_l$.

B. Expansion theorem for $\hat{O}_l$

Now we will prove the expansion theorem for the integrodifferential operator $\hat{O}_l$; that is, we will show that an equation of the same form as Eq. (19) holds for the operator $\hat{O}_l$

Consider a sequence of operators $\hat{O}_{ln}$ which converge as $n \to \infty$ to $\hat{O}_l$ and which have the same differential operator as $\hat{O}_l$ but whose integral operators are truncated:

$$\hat{O}_{ln} y = \hat{D}_l y + 8\pi^2\Theta(x - \alpha_n) \sum_{j=1}^{N_l} f_{jl}(x) \times \int_{\alpha_n}^{x} g_{jl}(s)y(s)ds + 8\pi^2\Theta(\beta_n - x) \sum_{j=1}^{N_l} g_{jl}(x) \times \int_x^{\beta_n} f_{jl}(s)y(s)ds. \quad (22)$$

As $n \to \infty$, we suppose $\alpha_n \to 0, \beta_n \to \infty$. Then the sequence $\hat{O}_{ln}$ and the operator $\hat{O}_l$ all have the same domain of definition (since each $\hat{O}_{ln}$ differs from $\hat{O}_l$ by a bounded operator) and $\hat{O}_{ln} \to \hat{O}_l$. Thus the sequence $\hat{O}_{ln}$ and the operator $\hat{O}_l$ satisfy a theorem[13] according to which the projection operators $P_n^{(1,2)}$ which project onto the portion of the spectrum of the self-adjoint operator $\hat{O}_{ln}$ lying between λ_1 and λ_2 converge to $P^{(1,2)}$, which is correspondingly defined with respect to $\hat{O}_l$.

The expansion theorem, Eq. (19), can also be proven for $\hat{O}_{ln}$. This is because Green's formula, here written as

$$\int_{x_1}^{x_2} dx(\chi\hat{O}_{ln}\Psi - \Psi\hat{O}_{ln}\chi) = [(\chi'\Psi - \Psi'\chi)p]\Big|_{x_1}^{x_2},$$

holds for $\hat{O}_{ln}$ as long as the end points of integration (x_1, x_2) both lie outside the interval $[\alpha_n, \beta_n]$. (In practice,[11] the end points would straddle the interval $[\alpha_n, \beta_n]$). By the theorem referred to in the preceding paragraph, one therefore can assert that the sequence

$$\langle P_n^{(1,2)}f | P_n^{(1,2)}f\rangle = \int_{\mu_1}^{\mu_2} \frac{d\lambda}{\pi\sqrt{-\lambda}\, A_{n\lambda}^2} \left| \int_0^\infty y_n(s,\lambda)f(s)ds \right|^2 \quad (23)$$

converges. Choose f so that it is square integrable and vanishes outside some interval $[a, b]$. Let n be large enough so that $\alpha_n < a, \beta_n > b$. Then the eigenfunction y_n of $\hat{O}_{ln}$ can be written, for $a < x < b$,

$$y_n(x,\lambda) = \sum_{i=1}^{4} B_{ni}(\lambda)\Psi_i(x,\lambda), \quad (24)$$

where $\Psi_i(x,\lambda)$ are the four linearly independent solutions of $\Delta^{(4)}\Psi = 0$. The amplitude $A_{n\lambda}$ is determined from the B_{ni} by matching the zeroth and first derivatives of y at the point $x = \beta_n$. It follows that the integrand in Eq. (23) is thus a continuous function of the four variables B_{ni}.

The sequence of integrals in Eq. (23) converges. By considering the system of equations which determine the coefficients B_{ni}, these latter functions (of λ) converge also. If the sequence of integrands in Eq. (23) converges uniformly on $[\lambda_1, \lambda_2]$, then the limit integral will be given by the integral of the limit function. The sequence of integrands will converge uniformly on $[\lambda_1, \lambda_2]$ if the sequences B_{ni} do.[14] But these latter coefficients are the solutions of a linear inhomogeneous system of equations whose matrix of coefficients converge uniformly on $[\lambda_1, \lambda_2]$ (see Appendix). Therefore, the integrands in Eq. (23) converge uniformly on $[\lambda_1, \lambda_2]$ to the function

$$\frac{1}{\sqrt{-\lambda}}\frac{1}{A_\lambda^2}\left|\int_0^\infty y(s,\lambda)f(s)ds\right|^2,$$

whose integral from λ_1 to λ_2 equals the limit of Eq. (23). Thus it has been shown that the projection operator onto the portion of the spectrum of $\hat{O}_l$ lying between λ_1 and λ_2 can be represented

$$\langle P^{(1,2)}f|P^{(1,2)}f\rangle = \frac{1}{\pi}\int_{\lambda_1}^{\lambda_2}\frac{d\lambda}{\sqrt{-\lambda}\,A_\lambda^2}\left|\int_0^\infty y(s,\lambda)f(s)ds\right|^2.$$

Putting $\lambda_2 = 0^-$ and $\lambda_1 = -\infty$, the above equation states that the norm of f_e, where f_e is the projection of f onto the continuous spectrum, can be written

$$\langle f_e|f_e\rangle = \frac{1}{\pi}\int_{-\infty}^{0}\frac{d\lambda}{\sqrt{-\lambda}\,A_\lambda^2}\left|\int_0^\infty y(s,\lambda)f(s)ds\right|^2.$$

From this it can readily be shown[15] that the expansion of any square integrable function f into "eigenfunctions" $y(x,\lambda)$ of $\hat{O}_l$ can be written

$$f(x) = \sum_i y_i(x,\lambda=0)\langle f|y_i(\lambda=0)\rangle + \frac{1}{\pi}\int_{-\infty}^{0}\frac{d\lambda}{\sqrt{-\lambda}}\frac{y(x,\lambda)}{A_\lambda^2}\int_0^\infty dx f(x)y(x,\lambda), \quad (25)$$

where A_λ is the amplitude of the asymptotic expression for large x of the eigenfunction which goes as x^{l+1} as $x \to 0$. This completes the proof of the expansion theorem for the operator $\hat{O}_l$.

IV. The Eigenfunctions of $\hat{O}_l$, $l = 1$

Having uncovered the nature of the spectrum of $\hat{O}_l$, we now turn to its eigenfunctions. These eigenfunctions are germane to the initial value problem related to the equation

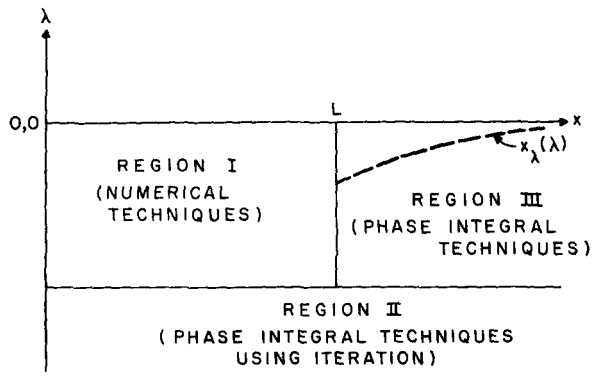

FIG. 1. The regions of the $x\lambda$ plane used in the numerical and analytic studies of the eigenfunction.

$$\frac{\partial y}{\partial t} = \hat{O}_l y.$$

We consider the special case $l = 1$.

For x and λ small, one must employ numerical techniques to construct $y(x,\lambda)$. Phase integral techniques come into play in the complimentary region of the $x\lambda$ plane. (See Fig. 1.)

The boundary between Region I and Region III is denoted by L. In Region I the following numerical procedure may be employed. On a certain set of values of λ lying between 0 and λ_M, the range $[0, L]$ is divided into N panels. In the ith panel the eigenfunction is approximated by a quadratic

$$A_i(x - x_i)^2 + B_i(x - x_i) + F_i,$$

where x_i is the left boundary of the ith panel. The A_i, B_i, F_i are constrained so that the zeroth and first derivatives match at the panel boundaries. Further equations for these coefficients are obtained by inserting the ordered set of quadratics representing the solution into $\hat{O}_l y - \lambda y$ and then requiring the result to be zero at one point in each of the N panels. Because of the nonlocal nature of the integral operator, the coefficients which represent the function in Region III will enter into the system of linear equations. These coefficients (in Region III) are in turn determined by the zeroth and first derivatives of the eigenfunction at $x = L$. The linear inhomogeneous system of equations for A_i, B_i, F_i and the coefficients in Region III are a closed system. It is inhomogeneous because we impose the condition[16] $y \to x^2$ as $x \to 0$ so that $A_1 = 1, B_1 = 0, F_1 = 0$.

In Region III, phase integral techniques may be used to obtain expressions for the eigenfunctions. The validity of this technique limits L to being no smaller than about 2.5.

First we will rewrite some of the pertinent equations and define some new functions. It is readily shown from Eqs. (10) and (11) that for $l = 1$ the function g_1 and f_1, introduced in Eq. (20a), can be written

$$f_1 = \pi^{-3/2}e^{-x^2/2}/x, \quad (26)$$

$$g_1 = e^{-x^2/2}\left(-(1+2x^2)J(x) + \frac{4}{x}\int_0^x (t^2+\tfrac{1}{2})^2 J(t)\,dt\right). \quad (26')$$

Just as the function $J(x)$ [Eq. (15)] can be written

$$J = \ln\Gamma + J_c(x),$$

where J_c is due to the dynamic screening contained in the dielectric function, so can $g_1(x)$ be written

$$g_1 = g_L + g_c,$$

where

$$g_L = e^{-x^2/2}(\tfrac{4}{5}x^4 - \tfrac{2}{3}x^2)\ln\Gamma, \quad (27a)$$

$$g_c \sim (\sqrt{\pi}/4x^3)e^{x^2/2}. \quad (27b)$$

The functions g_L and g_c result from using $\ln\Gamma$ and J_c respectively in place of J in Eq. (26'). Equation (27a) is exact, whereas Eq. (27b) is an approximate expression valid for $x \gtrsim 2$.

Rewriting Eq. (20a), with its right-hand side set equal to λy, in the form

$$\frac{d}{dx}\left(p\frac{dy}{dx}\right)+\left(V-\lambda-\frac{2Q}{x^2}\right)y+8\pi^2 f_1(x)\int_L^x g(s)y(s)ds$$
$$+8\pi^2 g_1(x)\int_x^\infty f_1(s)y(s)ds$$
$$=-8\pi^2 f_1(x)\int_0^L g_1(s)y(s)ds, \tag{28}$$

and then dividing this through by p, we get

$$y''+\frac{P'}{P}y'+\zeta y+\frac{8\pi}{\ln\Gamma}\frac{f_1}{P}\int_L^x g_1 y ds+\frac{8\pi}{\ln\Gamma}\frac{g_1}{P}\int_x^\infty f_1 y ds$$
$$=-\frac{8\pi f_1}{(\ln\Gamma)P}\int_0^L g_1 y ds, \tag{29}$$

where

$$P(x)\equiv(1/\pi\ln\Gamma)p(x),\quad \zeta(x,\lambda)\equiv(1/p)(V-\lambda-2Q/x^2).$$

By using Eq. (16c) and neglecting $g_L(x)$ in comparison to $g_c(x)$ (for $x>L$), Eq. (29) becomes approximately

$$y''-x^{-1}\left(3-\frac{\epsilon(x)}{2}\right)y'+\left(\frac{\nu x^3}{\sigma}-x^2-\frac{2}{\sigma}-\frac{\pi\epsilon(x)}{8}\right)y$$
$$+\epsilon(x)x^2e^{-x^2/2}\int_L^x\frac{ds}{s^3}e^{s^2/2}y+\epsilon(x)e^{x^2/2}\int_x^\infty\frac{ds}{s}e^{-s^2/2}y$$
$$=-\frac{4\epsilon(x)}{\sqrt{\pi}}x^2e^{-x^2/2}\int_0^L g_1(s)y(s)ds, \tag{30}$$

where

$$\epsilon(x)=2/\ln\Gamma x,\quad \sigma(x)=(\ln\Gamma x)/\ln\Gamma,\quad \nu=-\lambda/\pi\ln\Gamma.$$

Let

$$\hat{\tilde{D}}=\frac{d^2}{dx^2}+\frac{P'}{P}\frac{d}{dx}+\zeta(x,\lambda) \tag{31a}$$

and

$$\hat{I}y=\frac{8\pi}{\ln\Gamma}\frac{f_1}{P}\int_L^x g_1 y ds+\frac{8\pi}{\ln\Gamma}\frac{g_1}{P}\int_x^\infty f_1 y ds. \tag{31b}$$

The operator $\hat{I}$ here is related to $\hat{K}_1$, defined in Eq. (16d):

$$\hat{I}=(1/p)\hat{K}_1. \tag{32}$$

With these relations and definitions at hand it is convenient to indicate the process by which the solutions of $\Delta^{(4)}\Psi=0$ [see Eq. (20b)] may be found approximately. The four solutions to $\Delta^{(4)}\Psi=0$ come into play in matching the solution at $x=L$. One assumes solutions of $\Delta^{(4)}\Psi=0$, of the form e^{ϕ}, where $d\phi/dx$ behaves essentially as x to a positive power. Reference to Eq. (30) indicates that in Region III the integral operator $\hat{I}$ carries with it the small quantity $\epsilon(x)$. Therefore, the integral operator $\hat{K}$ acts as a small perturbation to the form of the eigenfunction in Region III. Because of the perturbative nature of the integral operator $\hat{I}$, two solutions of $\Delta^{(4)}y=0$ are only slightly displaced from the two linearly independent solutions of $\hat{D}y-\lambda y=0$. The correction is found quantitatively as follows: (a) Insert the function $y\sim e^{\phi}$ into

$$(\hat{\tilde{D}}+\hat{I})y=0. \tag{33}$$

(b) the one integral in Eq. (30) in which the derivatives of both ϕ and $\pm s^2/2$ have the same sign are approximated as follows,

$$\int_{x_0}^x S(s)e^{\phi(s)\pm s^2/2}ds\simeq\frac{e^{\pm s^2/2+\phi}S}{\phi_0'\pm s}\Big|_{x_0}^x, \tag{34}$$

where $S(x)$ is smoothly varying and where ϕ_0 is the phase in the corresponding solution of $\hat{D}y-\lambda y=0$. (c) The alternate integral to that referred to in (b) is "eliminated" by one propitious differentiation of Eq. (33). This, in combination with the results of (b), gives a pure third-order nonlinear differential equation for the phase $\phi(s)$. (d) This equation is solved approximately by writing $\phi(s)=\phi_0(s)+\eta(s)$, linearizing the resulting equation in $\eta'(s)$, and solving the resulting (algebraic) equation for $\eta'(s)$. Change of independent variable to

$$\tau(x)=\int_{x_0}^x\sqrt{-\zeta(s)}ds \tag{35}$$

renders this process a bit clearer.

The results of this analysis[17] are that two solutions of $\Delta^{(4)}\Psi=0$ are approximately

$$\Psi_{1,2}=\frac{1}{\sqrt{P}}\left(-\zeta-\frac{\epsilon}{x(\sqrt{-\zeta}+x)}-F_{\frac{1}{2}}(x)\right)^{-1/4}$$
$$\times\exp\left\{\mp\int_{x_\lambda}^x ds\left[-\left(\zeta+\frac{\epsilon}{s(\sqrt{-\zeta}+s)}+F_{\frac{1}{2}}(s)\right)\right]^{1/2}\right\}, \tag{36}$$

where

$$F_1(x)\simeq\frac{x^2e^{-x^2/2}}{\Psi_{10}(x)}\left(-\int_{x_0}^x\frac{\epsilon(s)ds}{s^3}e^{s^2/2}\Psi_{10}(s)+\frac{e^{x_0^2/2}\Psi_{10}(x_0)F_0}{x_0^2\sigma(x_0)}\right), \tag{36'a}$$

$$F_2(x)\simeq\begin{cases}\dfrac{e^{x^2/2}}{\Psi_{20}(x)}\displaystyle\int_x^\infty\frac{\epsilon(s)ds}{s}e^{-s^2/2}\Psi_{20}(s), & \nu x^3\gtrsim 2,\\ -4-\frac{1}{2}\epsilon, & \nu x^3\ll 2 \text{ and } \nu\ll L^{-1},\end{cases} \tag{36'b}$$

and

$$x_\lambda=\begin{cases}\text{root of } \zeta(x,\lambda)=0, & \text{if } \nu<L^{-1},\\ L, & \text{if } \nu>L^{-1},\end{cases} \tag{36'c}$$

and

$$\Psi_{\substack{10\\20}}(x)=\frac{1}{\sqrt{P}}\zeta^{-1/4}\exp\left(\mp\int_{x_\lambda}^x ds\sqrt{-\zeta}\right).$$

It can be shown that F_1 and F_2 both satisfy first-order nonlinear differential equations. In order to define a solution to either of these first order equations, it is necessary to specify the value of the solution at one point. For the case of F_2 we specify $F_2(\infty)=0$. For the case of F_1 we specify $F_1(x_0)=F_0$. The point x_0 and the value F_0 are in principle arbitrary but analysis shows that we may be guaranteed that F_1 will not diverge if we choose $x_0=(4/\lambda)^{1/3}$ and $|F_0|\le\sqrt{2\epsilon}$. In fact, if we choose $x_0=(4/\lambda)^{1/3}$ and $F_0=0$, then the solution F_1 will be of order ϵ in magnitude for all x.

For $\nu x^3\gg 2$, F_1 and F_2 become equal,

$$F_1\to F_2\to-\epsilon/x(\sqrt{-\zeta}-x),$$

so that, for $\nu x^3\gg 2$,

$$\Psi_{\frac{1}{2}}\to\frac{1}{\sqrt{P}}\left[-\left(\zeta+\frac{2\epsilon}{(\zeta+x^2)}\right)\right]^{-1/4}\exp\left\{\mp\int_{x_\lambda}^x ds\times\left[-\left(\zeta+\frac{2\epsilon}{\zeta+s^2}\right)\right]^{1/2}\right\}. \tag{36'd}$$

The significance of the number L^{-1} in Eq. (36'c) is that the root of $\zeta(x,\lambda)=0$ goes roughly as ν^{-1} as $\nu\to 0$, so that for $\nu \gtrsim L^{-1}$ there are in fact no roots of $\zeta(x,\lambda)$ greater than L. We thus have in Eqs. (36) expressions for two of the four solutions of $\Delta^{(4)}\Psi=0$ on one side of the turning point x_λ and two solutions (not the same) on the other.

We now discuss the connection of the phase integral solutions about the turning point. On each side of x_λ, Eq. (36) represents two solutions of $\Delta^{(4)}\Psi=0$. These two solutions together with the remaining two solutions on one side of the turning point connect in some manner to the corresponding set of four solutions on the other side.

It is possible to obtain expressions for the remaining two solutions of $\Delta^{(4)}\Psi=0$ by a process similar to that used to obtain Eqs. (36).[18] In this method, one makes the change of independent variable,

$$T(x)=\int_{x_0}^{x}(s^2)^{1/2}ds=\tfrac{1}{2}(x^2-x_0^2). \tag{37}$$

It turns out that two solutions sought are of the form

$$S(x)e^{\pm x^2/2}, \tag{38}$$

where S is a slowly varying function of x. In addition, with the change of variable given in Eq. (37), the said technique of solution gives no divergence at point x_λ, in contrast to the situation with the two solutions in Eq. (36).

On each side of the turning point x_λ, let the four solutions of $\Delta^{(4)}\Psi=0$ be denoted by $\Psi_i(x)$, $i=1,\ldots,4$, with Ψ_1 and Ψ_2 given in Eq. (36) and with Ψ_3 and Ψ_4 the exact solutions of $\Delta^{(4)}\Psi=0$ which behave as $e^{-x^2/2}$ and $e^{+x^2/2}$, respectively. An eigenfunction $y(x)$ will be expressed in Region III as a sum

$$y(x)=\gamma_1(x)\Psi_1(x)+\gamma_2(x)\Psi_2(x)+\gamma_3\Psi_3(x)+\gamma_4\Psi_4(x).$$

The parameters $\gamma_1(x)$ and $\gamma_2(x)$ are constant on either side of the turning point. The coefficients γ_3 and γ_4 are constant for all x since Ψ_3 and Ψ_4 are exact solutions of $\Delta^{(4)}\Psi=0$. It can be shown[19] that the relation between the values of γ_1 and γ_2 on either side of the turning point are such that two solutions, y_1 and y_2 of the equation

$$\hat{O}_l y-\lambda y=0,$$

are given by

$$\begin{aligned}&\Psi_1+b\Psi_2 \underset{x\ll x_\lambda}{\leftarrow} y_1-\gamma_3^{(1)}\Psi_3-\gamma_4^{(1)}\Psi_4 \underset{x\gg x_\lambda}{\rightarrow} \tfrac{1}{2}(\Psi_1-i\Psi_2),\\ &\Psi_2\leftarrow y_2-\gamma_3^{(2)}\Psi_3-\gamma_4^{(2)}\Psi_4\rightarrow -i\Psi_1+\Psi_2,\end{aligned} \tag{39}$$

where b is undetermined and $\gamma_3^{(1)},\gamma_4^{(1)},\gamma_3^{(2)}$, and $\gamma_4^{(2)}$ are determined by Eqs. (21a) and (21b). The solution to the connection problem is thus given by Eq. (39).

In order to construct the eigenfunction on its complete range $[0,\infty]$, it is necessary to match the solution and its first derivative at the boundary $x=L$. In this calculation we choose the four linearly independent solutions of $\Delta^{(4)}\Psi=0$ to be $W_i(x)$, $i=1,\ldots,4$, where

$$\begin{aligned}&W_3=\Psi_3,\quad W_4=\Psi_4,\\ &\text{and}\\ &\Psi_1+b\Psi_2 \underset{x\ll x_\lambda}{\leftarrow} W_1 \underset{x\gg x_\lambda}{\rightarrow} \tfrac{1}{2}(\Psi_1-i\Psi_2),\\ &\Psi_2\leftarrow W_2\rightarrow -i\Psi_1+\Psi_2.\end{aligned} \tag{40}$$

The eigenfunction is then written as

$$y(x,\lambda)=\sum_{i=1}^{4}A_i(\lambda)W_i(x,\lambda). \tag{41}$$

We can show that $A_4=0$. From Eq. (21b) we have

$$A_4=-\beta_4^{-1}\sum_{i=1}^{3}A_i\beta_1. \tag{42}$$

Next we note that $\beta_i=0$ for $i=1,2,3$, as can be seen by taking the limit $x\gg x_\lambda$ in Eq. (29), inserting $y=W_1, W_2$ or W_3, and noting that the results do not contain terms which grow as $e^{x^2/2}$ as $x\to\infty$. On the other hand, β_4 is in general not zero. Thus Eq. (41) reduces to

$$y=A_1W_1+A_2W_2+A_3W_3. \tag{45}$$

Let y_0 and w_0 be the value of the function and its first derivative, respectively, as determined numerically in Region I. Then, with $A_4=0$, the three remaining coefficients are determined by the three equations

$$A_1W_1(L)+A_2W_2(L)+A_3W_3(L)=y_0, \tag{46a}$$

$$A_1W_1'(L)+A_2W_2'(L)+A_3W_3'(L)=w_0, \tag{46b}$$

$$A_1c_1+A_2c_2+A_3c_3=-8\pi^2f_1(L)\int_0^L g_1(s)y(s)ds, \tag{46c}$$

where

$$c_i=\hat{D}_1W_i|_{x=L}+8\pi^2g_1(L)\int_L^\infty f_1(s)W_i(s)ds-\lambda W_i|_{x=L}. \tag{46d}$$

From Eq. (36) we see that W_1 and W_2 are respectively WKB solutions of

$$\left(\frac{1}{p}\hat{D}_1-\frac{\lambda}{p}+F_{(\frac12)}+\frac{\epsilon}{x(\sqrt{-\zeta}+x)}\right)W_{(\frac12)}=0,$$

so that c_1 and c_2 are

$$\begin{aligned}c_{(\frac12)}=&-p(L)\left(F_{(\frac12)}(L)+\frac{\epsilon}{L(\sqrt{-\zeta(L)}+L)}\right)W_{(\frac12)}(L)\\ &+8\pi^2g_1(L)\int_L^\infty f_1(s)W_{(\frac12)}(s)ds.\end{aligned} \tag{46e}$$

If $\nu L^3\gg 2$, then

$$c_{(\frac12)}\simeq-\frac{4\pi}{L^4}\frac{W_{(\frac12)}(L)}{(L\mp\sqrt{-\zeta(L)})} \tag{46f}$$

Equation (46c) is equivalent to Eq. (21a) in which

$$\alpha_i=8\pi^2\int_0^L dsg_1(s)W_i(s)+c_i/f_1(L),$$

where $W_i(x)$, for $x<L$, is the extension back into Region I of that solution of $\Delta^{(4)}W=0$ which is $W_i(x)$ in Region III. Equations (46a), (46b), (46c), together with the definition in Eq. (46d), imply a complete expression [Eq. (45)] for the eigenfunction in Region III.

These formulas may be used to obtain a more explicit expression for the eigenfunction y for $\nu<L^{-1}$ and $x\gg x_\lambda$. From Eqs. (36) and (40), we see that, except for the small range in ν where x_λ is close to L,

$$W_1(L)/W_2(L)=\exp\left(\int_L^{x_\lambda}(\sqrt{-\zeta+F_1}+\sqrt{-\zeta+F_2})ds\right)\gg 1.$$

Also, from Eqs. (46d) and (46f), we have

$$c_{(\frac{1}{2})} = 0(\epsilon)W_{(\frac{1}{2})}(L), \quad c_3 = 0(1)W_3(L).$$

It follows that for $\nu < L^{-1}$, to zeroth order in ϵ, one obtains

$$\frac{A_1}{A_2} \simeq -\frac{c_3(y_0W'_{2L} - w_0W_{2L}) + K(W_{2L}W'_{3L} - W_{3L}W'_{2L})}{c_3(y_0W'_{1L} - w_0W_{1L}) + K(W_{1L}W'_{3L} - W_{3L}W'_{1L})} \tag{47}$$

where

$$K \equiv -8\pi^2 f_1(L)\int_0^L g_1(s)y(s)ds.$$

From Eq. (47) we see that

$$\frac{A_1}{A_2} = O\left(\frac{W_2(L)}{W_1(L)}\right) \ll 1$$

in the domain $\nu < L^{-1}$. Therefore, the asymptotic form of the eigenfunction is

$$y(x,\nu) \underset{x \gg x_\lambda}{\sim} 2A_2 \frac{x^{3/4}}{[\nu\sigma(x)]^{1/4}} \times \cos\left[\int_{x_\lambda}^x ds\left(-\zeta + \frac{2\epsilon\sigma}{\nu s^3 - 2}\right)^{1/2} - \frac{\pi}{4}\right]. \tag{48}$$

From this we find that the function A_λ appearing in Eq. (25) is just twice $A_2(\nu)/\nu^{1/4}$ which is determined by Eqs. (46).

In the extreme limit $\nu \to 0$, these relations give a more explicit expression for A_2. The qualitative behavior is given by the zeroth order in ϵ expression:

$$A_2 \simeq \frac{P(L)}{2}W_1(L)\left[w_0 - \frac{d}{dx}\left(y_0 \ln W_1 + \frac{K}{k_3}\ln\frac{W_3}{W_1}\right)_{x=L}\right], \tag{49}$$

where

$$c_3 \equiv k_3 W_3(L).$$

Equation (49) shows that A_2 diverges as $\nu \to 0$,

$$A_2(\nu) \sim \exp\left(\int_L^{\nu^{-1}} \sqrt{-\zeta + F_1}\, ds\right).$$

With reference to the expansion theorem, this means that for a fixed but arbitrary square integrable function χ the relative contribution to the eigenfunction synthesis of χ between 0^+ and $\nu = \mu$ goes to zero proportionally to

$$\exp\left(-2\int_L^{\mu^{-1}} \sqrt{-\zeta}\, ds\right)$$

as $\mu \to 0$. In conclusion, we note that for $\nu > L^{-1}$, the functions W_1 and W_2 are complex, but naturally the solutions to Eqs. (46) are such that the function

$$y = A_1W_1 + A_2W_2 + A_3W_3$$

is real. The function A_ν^2 in Eq. (25) for the case $\nu > L^{-1}$ is given by

$$A_\nu^2 = \frac{1}{\sqrt{\nu}}(|A_1|^2 + |A_2|^2 + A_2\bar{A}_1 + A_1\bar{A}_2).$$

This completes the discussion of the matching of the analytically determined solution to the numerically determined one at $x = L$.

Next we consider the construction of the eigenfunctions of $\hat{O}_l$ in Region II. In Region II, the eigenvalue $\nu = -\lambda/\pi\ln\Gamma$ is large enough so that

$$(py')' + \left(V + \nu\ln\Gamma - \frac{2Q}{x^2}\right)y + 8\pi^2 f_1(x)\int_0^x g_1 y ds + 8\pi^2 g_1(x)\int_x^\infty f_1 y ds = 0 \tag{50}$$

can be solved by iteration. In this method the action of the integral operator on the solution is considered small compared to that of the differential operator. Therefore, the zeroth-order approximation is a combination of the two solutions of

$$\hat{D}y = (py')' + (V + \pi\nu\ln\Gamma - 2Q/x^2)y = 0. \tag{51}$$

Denote these by W_{01} and W_{02}. We write

$$y = A_{01}W_{01} + A_{02}W_{02} + y_1$$

and anticipate $|y_1| \ll |A_{01}W_{01} + A_{02}W_{02}|$.

The basis of the iteration procedure is that for large enough ν the functions W_{01} and W_{02} oscillate rapidly enough so that the effect of the differential operator in Eq. (51) on these functions is much larger than that of the integral operator. The function $A_{01}W_{01} + A_{02}W_{02}$ is inserted into the integral operator. Denote the resulting function by $H(x)$. Then the eigenfunction $y(x,\lambda)$ is taken to be the solution to the inhomogeneous equation

$$\hat{D}y - \lambda y = H$$

subject to the condition $y(x) \to x^2$ as $x \to 0$. This will determine the constants A_{01} and A_{02}.

An analysis for $l = 1$ based on the criterion that there should be at least one "wavelength" in the oscillation of W_{01} or W_{02} in the region $0.5 < x < 2$ shows that ν should be larger than about 30 for this iteration to be valid.

For $x > L$ this procedure gives the same result as the phase integral plus matching method used for Region III. This is because for large ν the difference in phase between the arguments of the circular functions in W_{01} and W_{02} and those in Eq. (36) amounts to

$$\int_L^x \frac{\sigma\epsilon}{\nu s^3}\frac{ds}{\sqrt{\nu s^3}} \propto \frac{2\epsilon\sigma}{7\nu^{3/2}L^{7/2}} \ll 1.$$

The conclusions concerning the qualitative behavior of the eigenfunction for two different values of λ are illustrated in Fig. 2. In Fig. 2(a), λ is small enough so that not only is the point (x,λ) in Region I or III (depending on whether x is less than or greater than L), but the turning point is also in Region III. In Fig. 2(b), λ is large enough so that iterative procedures relevant to Region II are employed.

CONCLUSIONS

In this paper properties of the linearized Balescu–Lenard equation which form the foundation for further physical application have been developed. The spectrum has been demonstrated to be continuous from zero to

minus infinity, and to contain no point components aside from the fivefold conservation degeneracy at the origin. The expansion theorem for $\hat{O}_l$ in terms of its own eigenfunctions was proved. Finally, the analytic and numerical procedures for approximating these eigenfunctions for the special case $l = 1$ was discussed.

ACKNOWLEDGMENTS

The authors gratefully acknowledge fruitful discussions on these topics with Professor J. Wilkins and Professor N. Rostoker. This research was supported in part by the Physics Branch of the Office of Naval Research Contract No. N 00014-67-A-0077-0015.

APPENDIX: PROOF OF UNIFORM CONVERGENCE OF ELEMENTS OF THE MATRIX WHICH DETERMINES THE COEFFICIENTS OF $\Psi_i(x)$

The solution $y_n(x, \lambda)$ given by Eq. (24) is a certain linear combination of the solutions of the pure Nth-order differential equation $\Delta^{(N)}y = 0$. The coefficients of these solutions are determined by a linear inhomogeneous system of equations, whose origin is described in the text. We will show that the elements of the matrix of coefficients in this system converge uniformly on $[\lambda_1, \lambda_2]$.

Let $\Psi_i(x)$ be the N independent solutions of $\Delta^{(N)}\Psi = 0$. An eigenfunction y_n of $\hat{O}_{nl}$ [see Eq. (22)] has the form

$$y_n(x) = \begin{cases} C_{nl}\phi_1(x) + C_{n2}\phi_2(x) & x < \alpha_n \\ \sum_{i=1}^{N} B_{ni}\Psi_i(x) & \alpha_n < x < \beta_n \\ C_{n3}\phi_1(x) + C_{n4}\phi_2(x) & \beta_n < x \end{cases} . \quad \text{(A1)}$$

The functions ϕ_1 and ϕ_2 are those solutions of $\hat{D}_l\phi = \lambda\phi$ which behave at the origin as x^{-l} and x^{l+1}, respectively.

The N coefficients B_{ni} are determined as follows: (a) The zeroth and first derivatives of y_n match at $x = \alpha_n$, and C_{nl} is set equal to zero. (b) The eigenfunction y_n is assigned a definite value at a given point x_0, where $\alpha_n < x_0 < \beta_n$. (c) These two constraints together with Eqs. (21a) and (21b), or their $N - 2$ analogs for $l \geq 2$, determine the coefficients B_{ni}.

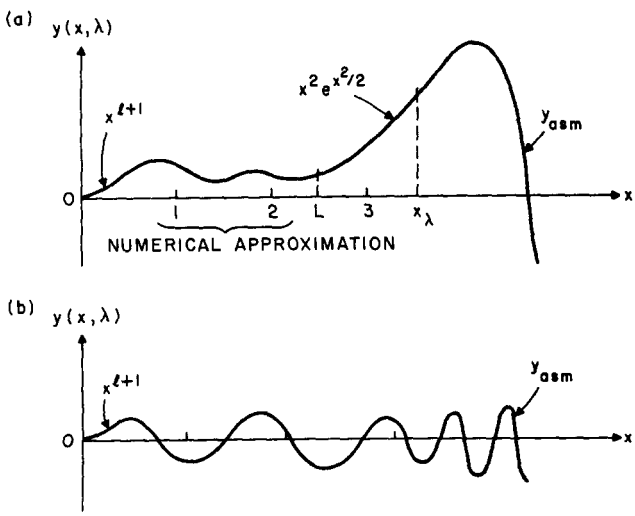

FIG. 2. Sketches of the $l = 1$ mode eigenfunction of the Linearized Balescu–Lenard operator in Region II and in Region I–III. Here $y_{asm} \sim \cos(\sqrt{\nu x^5/\sigma(x)} + \theta_\lambda)$, where θ_λ is a constant phase factor. (a) In Region I the solution is obtained by numerical techniques. This solution and its first derivative are matched to those of the phase integral solution in Region III. The turning point x_λ in the case depicted here is in Region III. (b) In Region II, the eigenfunctions are obtained by using a phase integral and an iterative technique.

We will show that the elements of the matrix which comprise the above inhomogeneous system of equations converge uniformly as $n \to \infty$ on any finite interval of the λ axis which does not include the origin $\lambda = 0$.

The equation $\Delta^{(N)}\Psi = 0$ has a regular singularity at the origin $x = 0$. Consequently, the N solutions $\Psi_i(x)$ of $\Delta^{(N)}\Psi = 0$ can be expanded near the origin:

$$\Psi_i = \sum_{j=0}^{\infty} x^{q_i+j}\gamma_{ij}(\ln x), \quad \text{(A2)}$$

where $\gamma_{ij}(z)$ are polynomials in z.

Also the two solutions to $\hat{D}\phi = \lambda\phi$ are expanded about the origin:

$$\phi_i(x) = \sum_{j=0}^{\infty} d_{ij}x^{p_i+j}, \quad \text{(A3)}$$

where in this case the d_{ij} are constants. Here $p_1 = -l$ and $p_2 = l + 1$.

Matching the function and its first derivative at $x = \alpha_n$ and setting $C_{nl} = 0$ gives

$$\sum_{j=1}^{N} B_{nj}S_j = 0, \quad \text{(A4)}$$

where

$$S_j = \sum_{k=0}^{\infty} \sigma_{jk}\alpha_n^{k+l+q_j} \quad \text{(A5)}$$

and

$$\sigma_{jk} = \sum_{i=0}^{k} \{(l + 1 + i)d_{2i}\gamma_{j,k-i} - [(q_j + i)\gamma_{ji} + \gamma'_{ji}]d_{2,k-i}\}. \quad \text{(A6)}$$

We rewrite Eq. (A5)

$$S_j = \alpha_n^{l+q_j}\left(\sigma_{j0} + \sum_{k=1}^{\infty}\sigma_{jk}\alpha_n^k\right). \quad \text{(A7)}$$

Thus, as $n \to \infty$ and $\alpha_n \to 0$,

$$S_j \to \alpha_n^{l+q_j}\sigma_{j0}, \quad \text{(A8)}$$

where

$$\sigma_{j0} = (l + 1)\gamma_{j0} - (q_j\gamma_{j0} + \gamma'_{j0}). \quad \text{(A9)}$$

We now make two assertions whose truth can be established. These are (a) if q_1, q_2, etc., satisfy $q_i < q_j$ if $i < j$, then $q_1 = -l$ and (b) the polynomial γ_{10} is just a constant, which we may set equal to unity.

It follows that, as $n \to \infty$, $S_1 \to 2l + 1$ and $S_j \to 0$ for $j > 1$. That this convergence is uniform on a finite interval $\lambda_1 \leq \lambda \leq \lambda_2$ can be seen as follows. The quantity S_j is an analytic function of both α_n and λ, since it is obtained by matching two functions each of which are analytic functions of α and λ. This means that $S_j(\alpha_n, \lambda)$ is a continuous function of α_n, λ except at points where S_j diverges. The same is true of the function $\alpha_n^{-r}S_j(\alpha_n, \lambda)$, where $r > 0$. We consider first the case $j > 1$, and then $j = 1$. For $j > 1$, the exponent $l_j \equiv l + q_j > 0$. We write Eq. (A7) in the form

$$S_j = \alpha_n^{l_j/2}\,\tilde{S}_j(\alpha_n, \lambda), \tag{A10}$$

where

$$\tilde{S}_j(\alpha_n, \lambda) = \alpha_n^{l_j/2}\left(\sum_{k=0}^{\infty} \sigma_{jk}\alpha_n^k\right).$$

From the above discussion, $\tilde{S}_j(\alpha_n, \lambda)$ is a continuous function of α_n, λ except at points where $\tilde{S}_j$ diverges. However, since $l_j/2 > 0$ and the coefficients σ_{jk} are at most powers of $\ln\alpha_n$, $\tilde{S}_j(0, \lambda) = 0$. There is therefore a rectangle, $\lambda_1 \leq \lambda \leq \lambda_2$, $0 \leq \alpha_n < a$, on which $\tilde{S}_j(\alpha_n, \lambda)$, and hence $|\tilde{S}_j(\alpha_n, \lambda)|$, are continuous and consequently possesses a maximum M. In this rectangle we thus have

$$|S_j(\alpha_n, \lambda)| \leq \alpha_n^{l_j/2} M, \tag{A11}$$

which establishes the uniform convergence of S_j to zero on $\lambda_1 < \lambda < \lambda_2$. For $j = 1$ we have, from Eq. (A7),

$$S_j(\alpha_n, \lambda) = 1 + \sum_{k=1}^{\infty} \sigma_{jk}(\ln\alpha_n, \lambda)\alpha_n^k, \tag{A12}$$

where $\sigma_{jk}(Z, \lambda)$ are polynomials in Z with coefficients depending on λ. The function $S_j - 1$ is a continuous function of σ_n, λ and vanishes as $\alpha_n \to 0$ like $\alpha_n(\ln\alpha_n)^w$, $w > 0$. Therefore it converges to zero uniformly on $\lambda_1 \leq \lambda \leq \lambda_2$ by the same type of argument by which S_j, $j > 1$, was shown to converge uniformly to zero.

The remaining conditions which the B_{ni} must satisfy, those corresponding to Eqs. (21a), (21b), or their analogs, can be shown to be equivalent to the requirement that the function $y_n(x)$ satisfy $\hat{O}_{ln}y_n = \lambda y_n$ at some point $\bar{x}$ as well as satisfying at $\bar{x}$ the equations obtained in each step in the sequence of differentiations performed to obtain $\hat{\Delta}^{(N)}y_n = 0$. The dependence on n in the coefficients of B_{ni} in the set of homogeneous equations so obtained comes from terms of the form $\int_{\bar{x}}^{\beta_n} f_{lk}(s)\Psi_i(s)\,ds$ and $\int_{\alpha_n}^{\bar{x}} g_{lk}(s)\Psi_i(s)\,ds$. The latter can be shown to converge uniformly on $\lambda_1 < \lambda < \lambda_2$ by arguments similar to those employed concerning the S_j; that is, upon writing $\int_{\alpha_n}^{\bar{x}} = \int_0^{\bar{x}} - \int_0^{\alpha_n}$, the analytic and continuity properties of $\int_0^{\alpha_n}$ are considered. The former converge uniformly on $\lambda_1 \leq \lambda \leq \lambda_2$ since, as can be shown by phase integral analysis, the Ψ_i satisfy

$$|\Psi_i(x)| < G(\lambda)e^{-x^2/2},$$

where $G(\lambda)$ is a continuous function of λ.

Finally, the determinant of the matrix which determines the B_i may vanish for one or more values of λ. The only inhomogeneous equation in the system of equations is the one in which the eigenfunction (A1) is given a definite value at a definite point x_0. The vanishing of the determinant for a certain λ merely means that for such a value of λ the eigenfunction has a node at x_0. It is therefore possible to break the interval $[\lambda_1, \lambda_2]$ into a finite number of subintervals, choosing for each subinterval an x_0 and a value for the eigenfunction (for $n = \infty$) such that the determinant of the matrix never vanishes and the B_i are continuous functions of λ from one subinterval to the next.

[1] C. S. Wang-Chang and G. E. Uhlenbeck, "On the propagation of sound in monatomic gases," Engineering Research Institute Report (University of Michigan, 1952).

[2] H. Grad, *Third symposium on rarefied gas* (Academic, New York, 1963).

[3] L. Waldmann, *Handbuck der Physik*, edited by S. Flügge (Springer-Verlag, Berlin, 1958), Vol. 12.

[4] E. Inönü and P. F. Zweifel, *Developments in transport theory* (Academic, New York, 1967).

[5] Jordan D. Lewis, J. Math. Phys. **8**, 791 (1967).

[6] C. H. Su, J. Math. Phys. **8**, 248 (1967).

[7] Ta-You Wu, *Kinetic equations of gases and plasmas* (Addison-Wesley, Reading, Mass., 1966), p. 198.

[8] (a) R. Balescu, *Statistical mechanics of charged particles* (Wiley, New York, 1963), p. 229. (b) D. C. Montgomery and D. A. Tidman, *Plasma kinetic theory* (McGraw-Hill, New York, 1964), Chap. 7. (c) R. L. Liboff, *Introduction to the theory of kinetic equations* (Wiley, New York, 1969), pp. 252–277. (d) R. L. Liboff, Phys. Fluids **2**, 40 (1959).

[9] Osamu Aono, Phys. Fluids **11**, 341 (1968).

[10] F. Riesz and B. Sz-Nagy, *Functional analysis* (Ungar, New York, 1953), p. 367.

[11] E. Coddington and N. Levinson, *Theory of ordinary differential equations* (McGraw-Hill, New York, 1955), Chap. 9.

[12] Reference 10, p. 361.

[13] Reference 10, p. 369.

[14] The quantity $I_n(\lambda) \equiv A_{n\lambda}^{-2}|\int_0^\infty Y_n(s,\lambda)\,f(s)ds|^2$ is a ratio of two quantities each of which is quadratic in the B_{ni}. The coefficients in the quadratic expression for $A_{n\lambda}^2$ as a function of the B_{ni} can be shown by an analysis of the matching process at β_n, to converge uniformly on $\lambda_1 \leq \lambda \leq \lambda_2$. Also the limit function A_λ^2 does not vanish on $\lambda_1 \leq \lambda \leq \lambda_2$ since, if it did, the eigenfunction would be square integrable.

[15] The set of square integrable functions which vanish outside a finite interval is dense in L^2. Therefore Parseval's equality holds for any square integrable function f: $\langle f|f\rangle = \Sigma_i |\langle f|y_i(\lambda = 0)\rangle|^2 + (1/\pi) \times \int_{-\infty}^{0} (d\lambda/\sqrt{-\lambda}\,A_\lambda^2)|\int_0^\infty y(s,\lambda)\,f(s)\,ds|^2$, where $y_i(\lambda = 0)$ are the square integrable eigenfunctions at $\lambda = 0$ (if any). From Parseval's equality, Eq. (25) may be established (see Ref. 11, pp. 237–38).

[16] It can be shown that the solutions to $\hat{O}_l y - \lambda y = 0$ behave at the origin as x^{-l}, x^{l+1}; that is, the same as the solutions of $\hat{D}_l\, y - \lambda y = 0$.

[17] A. Merchant, Ph. D. thesis (University Microfilms, 1971), pp. 74–86.

[18] Reference 17, pp. 86–87.

[19] Reference 17, pp. 87–96.

A model of field theory treated in the Fock–Cook formalism

Franklin E. Schroeck Jr.
Department of Mathematics, Florida Atlantic University, Boca Raton, Florida 33432
(Received 10 July 1972)

We present an alternative to the usual formalism for quantum field theory by generalizing the Fock–Cook formalism. We illustrate the method by applying it to a generalization of a familiar model of quantum field theory.

I. INTRODUCTION

In the present paper we shall develop a mathematical structure for quantum field theory, free of generalized operators, and in the spirit of the work of Cook.[1] Here, the fields are defined as *bona fide* operators acting in a Hilbert space in accordance with the notion that the formalism should only describe the creation or annihilation of particles corresponding to physically realizable wavefunctions. In this way we avoid the usual difficulties of multiplication of generalized operators.[2]

The method is illustrated by the rigorous treatment of a generalization of a familiar model of field theory—the "scalar field model."[3] The generalization (i) includes the possibility of spins and charges, (ii) allows for 1, 2, or 3 space dimensions, and (iii) includes treatment of the "heavy" particle in the recoilless, Galilean recoil and relativistic recoil cases. Since we do not include antiparticles, only a single infinite renormalization remains. Inclusion of antiparticles would make the Yukawa interaction a special case of the model; but it would also increase the difficulties connected with the renormalizations. For the present, we shall restrict ourselves to the simpler model to illustrate the formalism.

In Sec. II we review the basic techniques involved in the Fock–Cook formalism.[1] In Sec. III we derive a formal expression for the interaction. Physical principles lead us to writing this interaction as sums over the *bona fide* fields rather than as the usual integrals over operator-valued distributions. In Sec. IV we introduce cutoffs both of the ordinary momentum space or configuration space varieties, and of the form of restrictions on certain sums. The self-adjointness of the Hamiltonian with cutoffs is discussed in Sec. V. The removal of the cutoffs after renormalization will be treated in a subsequent paper.

II. THE HILBERT SPACE AND THE NOTATION

In this section all spaces and operators will have precisely the meaning attributed to them in Cook's work, henceforth denoted (C), to which we refer the reader for details.

We will be concened with two types of particle: one a boson and the other a fermion.

The bosons are to be of mass $\mu > 0$ and, in general, are to have spin and charge states labeled by the subscripts α_1 and α_2, respectively. The number of spin states, respectively, charge states, will be denoted by $(2n_{\alpha_1} + 1)$, respectively, $(2n_{\alpha_2} + 1)$. The letter "a" as a subscript on operators will label them as acting only on the bosons. The bosons will be treated as "light" particles and accordingly will always be treated relativistically. From the theory of induced representations of the Poincaré group, or otherwise,[4] we know that the one-particle Hilbert space for particles of spin zero and positive mass is the space of all measurable square integrable functions on Minkowski space, with points labeled (p_1, p_2, p_3, p_4), with respect to the measure $\delta(p^2 - \mu^2)\theta(p_4)^{-1}dp_1dp_2dp_3$, where $p^2 = p_1^2 + p_2^2 + p_3^2 - p_4^2$. Since this measure is concentrated on the hyperboloid $p^2 = \mu^2$, it is equivalent to consider the space $\mathcal{L}_a^2(\mathbb{R}^3)$ with measure $d\mu(p)^1 = \omega(p)^{-1}dp_1dp_2dp_3 = \omega(p)^{-1}d^3p$, and scalar product $(f,g)_a = \int f(p)^\dagger g(p)\omega(p)^{-1}d^3p$, where we have used the notation $\dagger$ = complex conjugate, $p = (p_1, p_2, p_3)$, and $\omega(p) = (\mu^2 + p_1^2 + p_2^2 + p_3^2)^{1/2}$. For s space dimensions we take $p = (p_1, \ldots, p_s)$, $\omega(p) = [\mu^2 + p_1^2 + \cdots + p_s^2]^{1/2}$, and $(f,g)_a = \int f(p)^\dagger g(p)\omega(p)^{-1}d^sp$. To include the spin and charge states of the bosons, we take the single particle space to be the direct sum of $\mathcal{L}_a^2(\mathbb{R}^s)$ taken $(2n_{\alpha_1} + 1)(2n_{\alpha_2} + 1)$ times: $\mathcal{K}_a^{(1)} = \mathcal{L}_a^2(\mathbb{R}^s) \oplus \cdots \oplus \mathcal{L}_a^2(\mathbb{R}^s)$. We will abbreviate $n_a = (2n_{\alpha_1} + 1)(2n_{\alpha_2} + 1)$. Elements of this space will be labeled $\vec{f} = \{f_\alpha\}, f_\alpha \in \mathcal{L}_\alpha^2(\mathbb{R}^s)$, $\alpha = (\alpha_1, \alpha_2)$, $\alpha_i = n_{\alpha_i}, n_{\alpha_i} - 1, \ldots, -n_{\alpha_i}$, $i = 1, 2$ and have scalar product $(\vec{f}, \vec{g})_a = \sum_\alpha (f_\alpha, g_\alpha)_a$. Henceforward, the arrow vector will be used exclusively to distinguish elements of the single particle Hilbert spaces from elements of the $\mathcal{L}^2$ spaces.

The Fock–Cook space constructed from $\mathcal{K}_a^{(1)}$ is denoted by $\mathcal{F}_a$; i.e., $\mathcal{F}_a = \sum_{n=0}^\infty \oplus \mathcal{K}_a^{(n)}$ where $\mathcal{K}_a^{(n)}$ designates the subspace of symmetric tensors in $\mathcal{K}^{(n)}$, the n-fold tensor product of $\mathcal{K}_a^{(1)}$ with itself. $\mathcal{K}_a^{(0)}$ is the space of complex numbers. The corresponding creation and annihilation operators are denoted by $a^+(\vec{f})$ and $a(\vec{f})$, respectively, and obey the canonical equal-time commutation relations $[a(\vec{f}), a^+(\vec{g})]^\sim = (\vec{f}, \vec{g})_a I_a$, etc., where I_a denotes the identity operator in $\mathcal{F}_a$; $(\sim)$ denotes the smallest closed extension, and the $\vec{f}, \vec{g}$ are by definition taken from $\mathcal{K}_a^{(1)}$.

The one particle free Hamiltonian for the bosons is taken to be the self-adjoint operator

$$H_{0a}^{(1)} = \sum_{i=1}^{n_a} [h_a^{\delta(i,1)} \oplus h_a^{\delta(i,2)} \oplus \cdots \oplus h_a^{\delta(i,n_a)}]^\sim,$$

where $\delta(i,j)$ is the Kronecker delta function, $h_a^0 = I$, and h_a is defined by

$$D(h_a) = \{f \in \mathcal{L}_a^2(\mathbb{R}^s) \text{ such that } \int |f(p)|^2 \omega(p) d^sp < \infty\}$$

and for such f

$$(h_a f)(p) = \omega(p) f(p).$$

The second quantized free Hamiltonian is taken to be the self-adjoint operator $\Omega(H_{0a}^{(1)})$ restricted to $\mathcal{F}_a$ and will be denoted H_{0a}. [The Ω notation is that defined in (C).] The number operator N_a is defined as the restriction of $\Omega(I)$ to $\mathcal{F}_a$, and from the inequality $\omega(p) \geq \mu$, it follows that $H_{0a} \geq \mu N_a$. From (C) we obtain the usual

commutation relations between H_{0a}, N_a, and the fields, as well as the estimates

$$\|a^+(\vec{f})\chi\| \le \|\vec{f}\|_a \|(N_a + I_a)^{1/2}\chi\|,$$
$$\|a(\vec{f})\chi\| \le \|\vec{f}\|_a \|N_a{}^{1/2}\chi\|$$

for all $\chi \in D(N_a{}^{1/2})$, $\|\vec{f}\|_a = (\vec{f},\vec{f})_a{}^{1/2}$.

We now turn to the fermions which are the "heavy" particles in this model. Because of this "heaviness" some nonrelativistic approximations of the kinetic energy are natural. We shall treat the cases of static, Galilean, and relativistic recoil, precise definitions of which will be given below. For simplicity we shall disregard the corresponding anti-fermions. The fermions are considered to have spin with spin state indexed by the subscript $\beta_1 = n_{\beta_1}, n_{\beta_1} - 1, \ldots, -n_{\beta_1}$, where n_{β_1} is a half-odd integer. They may also exist in charged states labeled by the charge index $\beta_2 = n_{\beta_2}, n_{\beta_2} - 1, \ldots, -n_{\beta_2}$. β will denote (β_1,β_2). The subscript ψ on operators, spaces, etc., will indicate that the concerned object is relevant to the bosons only. The single particle Hilbert space $\mathcal{K}_\psi{}^{(1)}$ is taken to be the $(2n_{\beta_1} + 1) \times (2n_{\beta_2} + 1)$-fold direct sum of $\mathcal{L}^2(\mathbb{R}^s)$, with Lebesgue measure, with itself. We abbreviate $n_\psi \equiv (2n_{\beta_1} + 1) \times (2n_{\beta_2} + 1)$. Vectors in $\mathcal{K}_\psi{}^{(1)}$ will be designated by $\vec{g} = \{g_\beta\}$, $g_\beta \in \mathcal{L}^2(\mathbb{R}^s)$. The Fock–Cook space for the fermions is $\mathfrak{F}_\psi = \sum_{n=0}^{\infty} \oplus \mathcal{K}_\psi{}^{(n)}$, where $\mathcal{K}_\psi{}^{(n)}$ is the subspace of antisymmetric tensors in the n-fold tensor product of $\mathcal{K}_\psi{}^{(1)}$ with itself. The corresponding creation and annihilation operators are denoted by $\psi^+(\vec{g})$ and $\psi(\vec{g})$, respectively, defined for all $\vec{g} \in \mathcal{K}_\psi{}^{(1)}$, and obey the canonical equal-time anticommutation relations $\{\psi(\vec{g}), \psi^+(\vec{h})\}_+ = (\vec{g},\vec{h})_\psi I_\psi$, where I_ψ is the identity operator on $\mathfrak{F}_\psi$ and $(\vec{g},\vec{h})_\psi = \sum_\beta (g_\beta, h_\beta) = \sum_\beta \int g_\beta(k)^\dagger h_\beta(k)\, d^s k$. The Fermi fields are bounded operators on $\mathfrak{F}_\psi$ satisfying $\|\psi^+(\vec{g})\| = \|\psi(\vec{g})\| = \|\vec{g}\|_\psi$ where $\|\vec{g}\|_\psi^2 = (\vec{g},\vec{g})_\psi$. The one-particle free Hamiltonian $H_{0\psi}{}^{(1)}$ is the self-adjoint operator

$$H_{0\psi}^{(1)} = \sum_{i=1}^{n_\psi} [h_\psi{}^{\delta(i,1)} \oplus h_\psi{}^{\delta(i,2)} \oplus \cdots \oplus h_\psi{}^{\delta(i,n_\psi)}]^{\sim},$$

where h_ψ is defined by

$$D(h_\psi) = \{f \in \mathcal{L}^2(\mathbb{R}^s) \text{ such that } \int |\epsilon(k) f(k)|^2 d^s k < \infty\},$$

and for such f, $(h_\psi f)(k) = \epsilon(k) f(k)$, where $\epsilon(k)$ is given by one of the choices m, $m + (k_1{}^2 + \cdots + k_s{}^2)(2m)^{-1}$, or $(k_1^2 + k_2^2 + \cdots + k_s^2 + m^2)^{1/2}$. These choices of $\epsilon(k)$ will be called the static, the Galilean, and the relativistic recoil cases, respectively. The second-quantized free Hamiltonian is the positive self-adjoint operator $H_{0\psi} = \Omega_\psi(H_{0\psi}^{(1)})$, the restriction of $\Omega(H_{0\psi}^{(1)})$ to $\mathfrak{F}_\psi$. The number operator is $N_\psi = \Omega_\psi(I)$, and since $\epsilon(k) \ge m$, it follows from (C), that $H_{0\psi} \ge mN_\psi$.

For both the fermion and boson cases our convention on functions over $\mathbb{R}^s$ is that the variables are to be momentum space variables unless the function is embellished with a hat ($\hat{\ }$), denoting Fourier transform, in which case the variables are to be in configuration space.

The space for the combined system of bosons and fermions is $\mathfrak{F} = \mathfrak{F}_a \otimes \mathfrak{F}_\psi$. Operators which act only on $\mathfrak{F}_a$ or $\mathfrak{F}_\psi$ are always indexed by "a" or "ψ" so that superfluous tensor product notation will be dropped. For example $a(\vec{f}) \otimes I_\psi$ will be written $a(\vec{f})$. The total free Hamiltonian is defined as $H_0 = [H_{0a} + H_{0\psi}]^{\sim}$ and is self adjoint and positive on $\mathfrak{F}$.

III. THE FORMAL MODEL

For our interaction we shall allow creation and annihilation of the bosons in the presence of fermions, so that a particular interaction would be of the form $a^\#(\vec{f})\psi^+(\vec{g})\psi(\vec{h})$, for some $\vec{f} \in \mathcal{K}_a{}^{(1)}$ and $\vec{g},\vec{h} \in \mathcal{K}_\psi{}^{(1)}$ and where # on any operator will be used henceforward to indicate either the operator itself or its adjoint. We expect the relative strength of such interactions to be proportional to the overlap of the wavefunctions and to depend, in general, on the charge and spin. The overlap property will arise naturally from translation invariance and locality requirements.

We define a canonical injection from the $\mathcal{L}^2$ spaces into the single particle spaces as follows:

Let $f \in \mathcal{L}_a^2(\mathbb{R}^s)$. Then define $\vec{f}_\alpha \in \mathcal{K}_a{}^{(1)}$ by $(\vec{f_\alpha})_{\alpha'} = \delta_{\alpha,\alpha'} f$. For example let $\{f_i\}_{i=0}^\infty$ be an orthonormal basis for $\mathcal{L}_a^2(\mathbb{R}^s)$. We generate an orthonormal basis for $\mathcal{K}_a{}^{(1)}$ labeled $\{\overrightarrow{f_{i,\alpha}}\}$, $\alpha = (\alpha_1,\alpha_2)$ by taking $\overrightarrow{f_{i,\alpha}} = \{f_{i,\alpha'}\}$, $f_{i,\alpha'} = \delta_{\alpha,\alpha'} f_i$. Similarly, let $\{g_i\}_{i=0}^\infty$ be an orthonormal basis for $\mathcal{L}^2(\mathbb{R}^s)$ and generate the particular basis for $\mathcal{K}_\psi{}^{(1)}$ given by $\{\overrightarrow{g_{i,\beta}}\}$, $\beta = (\beta_1,\beta_2)$.

The interaction is taken to have the form

$$V = \sum_{\alpha,\beta,\beta'} V(\alpha,\beta,\beta'), \tag{1a}$$

$$V(\alpha,\beta,\beta') = \sum_{i,j,k=0} C_{ijk}^{\alpha\beta\beta'} a^+(\overrightarrow{f_{i,\alpha}})\psi^+(\overrightarrow{g_{j,\beta}})\psi(\overrightarrow{g_{k,\beta'}}) + \text{c.c.}$$

Temporarily neglecting spin–orbit interactions, we assume that $C_{ijk}^{\alpha\beta\beta'}$ factors into $C_{\alpha\beta\beta'}$ and C'_{ijk}, where $C_{\alpha\beta\beta'}$ depends only on the spins and charges, and C'_{ijk} depends only on the space part. Since α,β,β' have n_a, n_ψ, and n_ψ possible values, respectively, $C_{\alpha\beta\beta'}$ is a finite matrix, which we take to be real for simplicity of notation. The details of charge conservation, allowance of spin flip, charge or spin independence of the interaction, etc., may be included as further constraints on $C_{\alpha\beta\beta'}$, which we shall, however, not need in the sequel.

In order to have the interaction formally basis independent we choose C'_{ijk} to be linear in $\overrightarrow{g_{k,\beta'}}$ and antilinear in $\overrightarrow{f_{i,\alpha}}, \overrightarrow{g_{j,\beta}}$. Thus

$$C'_{ijk} = \int d^s p\, \omega(p)^{-1} d^s q\, d^s r\, f_i(p)^\dagger g_j(q)^\dagger g_k(r) K(p,q,r),$$

where the kernel is to be determined.

It is convenient to write all expressions in such a way that they depend only on the basis functions for $\mathcal{L}^2(\mathbb{R}^s)$. For this we observe that if $\{g_i\}_{i=0}^\infty$ is a basis for $\mathcal{L}^2(\mathbb{R}^s)$, and forming $\{\overrightarrow{g_{i,\alpha}}\}$ by $(\overrightarrow{g_{i,\alpha}})_{\alpha'} = \delta_{\alpha,\alpha'} g_i$, we then have the result[5]

$$\sum_{j=0}^{\infty} \sum_{\alpha} (\overrightarrow{g_{j,\alpha}}, \vec{f}) a^+(\overrightarrow{g_{i,\alpha}})\chi = a^+(\vec{f})\chi \tag{2}$$

for all $\vec{f} \in \mathcal{K}_a{}^{(1)}$ and for all χ in $D(N_a^{1/2})$. This result is not obvious since the expansion is not an expansion in a basis for $\mathcal{K}_a{}^{(1)}$, and the scalar product is not the scalar product in $\mathcal{K}_a{}^{(1)}$. In fact these two properties compensate exactly.

Let K_{jk} be defined by $K_{jk}(p) = \int d^s q d^s r\, g_j(q)^\dagger g_k(r) \times K(p,q,r)$. We assume $K_{jk} \in \mathcal{L}^2_a(\mathbb{R}^s)$. Then we have

$$\sum_{i=0}^{\infty} C'_{ijk} a^+(\overrightarrow{f_{i,\alpha}})\chi = \sum_{i=0}^{\infty} (f_i, K_{jk})_a\, a^+(\overrightarrow{f_{i,\alpha}})\chi$$

$$= \sum_{i=0}^{\infty} (\overrightarrow{f_{i,\alpha}}, \overrightarrow{K_{jk,\alpha}})_a\, a^+(\overrightarrow{f_{i,\alpha}})\chi = \sum_{i=0}^{\infty} \sum_{\alpha'} (\overrightarrow{f_{i,\alpha'}}, \overrightarrow{K_{jk,\alpha}})_a\, a^+(\overrightarrow{f_{i,\alpha'}})\chi$$

$$= a^+(\overrightarrow{K_{jk,\alpha}})\chi = \sum_{i=0}^{\infty} \sum_{\alpha'} (\overrightarrow{g_{i,\alpha'}}, \overrightarrow{K_{jk,\alpha}}) a^+(\overrightarrow{g_{i,\alpha'}})\chi$$

$$= \sum_{i=0}^{\infty} (g_i, K_{jk}) a^+(\overrightarrow{g_{i,\alpha}})\chi.$$

We, therefore, may transform $V(\alpha,\beta,\beta')$ to

$$V(\alpha,\beta,\beta') = C_{\alpha\beta\beta'} \sum_{i,j,k=0}^{\infty} D_{ijk} a^+(\overrightarrow{g_{i,\alpha}})\psi^+(\overrightarrow{g_{j,\beta}})\psi(\overrightarrow{g_{k,\beta'}}) + \text{c.c.}, \tag{1b}$$

where $D_{ijk} = (g_i, K_{jk})$

$$= \int d^s p d^s q d^s r\, g_i(p)^\dagger g_j(q)^\dagger g_k^{\vee}(r) K(p,q,r)$$

$$= \int d^s x d^s y d^s z\, \hat{g}_i(x)^\dagger \hat{g}_j(y)^\dagger\, \hat{g}_k^{\vee}(z) \hat{K}(x,y,z),$$

where $\hat{}$ denotes Fourier transform and where $^{\vee}$ denotes the second Fourier transform, i.e., for all $h \in \mathcal{L}^2(\mathbb{R}^s)$, $h^{\vee}(k) = h(-k)$. Translation invariance requires

$$\hat{K}(x,y,z) = \hat{K}(x-c, y-c, z-c) \quad \text{for all} \quad c \in \mathbb{R}^s.$$

Locality requires $\hat{K}(x,y,z) = \rho(x)\delta^{(s)}(x-y)\delta^{(s)}(x+z)$, so that both require

$$D_{ijk} = (\hat{g}_i \hat{g}_j, \hat{g}_k). \tag{1c}$$

We assume sufficient regularity on the basis so that these integrals exist. For instance, it is sufficient to assume $\hat{g}_i \in \mathcal{L}^2(\mathbb{R}^s) \cap \mathcal{L}^\infty(\mathbb{R}^s)$ for all $i = 0, 1, \cdots$. With these assumptions and recalling (2) we may perform the i sum obtaining

$$V(\alpha,\beta,\beta') = C_{\alpha\beta\beta'} \sum_{j,k=0}^{\infty} a^+(\overrightarrow{g_j^{\dagger\vee} * g_{k,\alpha}})\psi(\overrightarrow{g_{j,\beta}})\psi(\overrightarrow{g_{k,\beta'}}) + \text{c.c.}, \tag{1b'}$$

where $*$ denotes convolution.

Alternatively we may choose to perform the j,k summations. For these purposes, let us now define an auxiliary operator:

Let $\hat{f} \in \mathcal{L}^\infty(\mathbb{R}^s)$. Let $\Theta(f)$ be the bounded operator on $\mathcal{L}^2(\mathbb{R}^s)$ defined by $D(\Theta(f)) = \mathcal{L}^2(\mathbb{R}^s)$, and for $g \in \mathcal{L}^2(\mathbb{R}^s)$, $\Theta(f)g = \overbrace{\hat{f}\hat{g}}$, where $\overbrace{\ }$ denotes the inverse Fourier transformation. Then $\|\Theta(f)\| = \|\hat{f}\|_\infty$.

Let $M(f,\alpha)$ denote the bounded operator on $\mathfrak{K}_\psi^{(1)}$ such that for all $\vec{g}$ in $\mathfrak{K}_\psi^{(1)}$, $\vec{g} = \{g_\beta\}$,

$$(M(f,\alpha)\vec{g})_\beta = \sum_{\beta'} C_{\alpha\beta\beta'} \Theta(f) g_{\beta'}.$$

We may think of $M(f,\alpha)$ as an $n_\psi \times n_\psi$ matrix of operators and deduce a bound for $M(f,\alpha)$ as follows:

$$\|M(f,\alpha)\vec{g}\|_\psi^2 = \sum_\beta \|(M(f,\alpha)\vec{g})_\beta\|^2$$

$$= \sum_\beta \Big\| \sum_{\beta'} C_{\alpha\beta\beta'} \Theta(f) g_{\beta'} \Big\|^2$$

$$\le \sum_\beta \|\hat{f}\|_\infty^2 \Big\| \sum_{\beta'} C_{\alpha\beta\beta'} g_{\beta'} \Big\|^2$$

$$\le \|\hat{f}\|_\infty^2 \sum_\beta \Big(\sum_{\beta'} |C_{\alpha\beta\beta'}| \|g_{\beta'}\| \Big)^2$$

$$\le \|\hat{f}\|_\infty^2 \sum_\beta \Big(\sum_{\beta'} |C_{\alpha\beta\beta'}| \Big)^2 \|\vec{g}\|_\psi^2$$

or

$$\|M(f,\alpha)\| \le \Big[\sum_\beta \Big(\sum_{\beta'} |C_{\alpha\beta\beta'}| \Big)^2 \Big]^{1/2} \|\hat{f}\|_\infty.$$

Since $M(f,\alpha)$ is a bounded operator, any dense set of vectors in $\mathfrak{K}_\psi^{(1)}$ is a core for $M(f,\alpha)$. Furthermore $\Omega_\psi(M(f,\alpha))$ exists and $D[\Omega_\psi(M(f,\alpha))] \supseteq D(N_\psi)$. We now look for an expansion for $\Omega_\psi(M(f,\alpha))$ in terms of the fields. The results is already known if $M(f,\alpha)$ is a normal operator.[6] In general, $M(f,\alpha)$ is not normal; however, we may capitalize on the fact that $M(f,\alpha)$ is a bounded operator. Then the analysis goes through exactly as in the case for normal operators, but using the additional fact that $\Omega_\psi(M(f,\alpha))$ is defined everywhere on each $\mathfrak{K}_\psi^{(m)}$. Thus we have $\Omega_\psi(M(f,\alpha))^+ = \Omega_\psi(M(f,\alpha)^\dagger)$ and for the basis $\{\overrightarrow{g_{i,\beta}}\}$ of $\mathfrak{K}_\psi^{(1)}$, $\Omega_\psi(M(f,\alpha))$

$$= \sum_{j,k=1}^{\infty} \sum_{\beta,\beta'} (\overrightarrow{g_{j,\beta}}, M(f,\alpha)\overrightarrow{g_{k,\beta'}})\psi^+(\overrightarrow{g_{j,\beta}})\psi(\overrightarrow{g_{k,\beta'}})$$

$$= \sum_{j,k=1}^{\infty} \sum_{\beta,\beta'} C_{\alpha\beta\beta'} (\hat{g}_j, \hat{f}\hat{g}_k)\psi^+(\overrightarrow{g_{j,\beta}})\psi(\overrightarrow{g_{k,\beta'}}).$$

Therefore, choosing the basis $\{g_i\}$ such that $\hat{g}_i \in \mathcal{L}^\infty(\mathbb{R}^s)$ for all i, we have

$$V = \sum_\alpha V_\alpha,$$

$$V_\alpha = \sum_{i=0}^{\infty} a^+(\overrightarrow{g_{i,\alpha}})\Omega_\psi(M(\overbrace{g_i^{\dagger\vee}}, \alpha)) + \text{c.c.} \tag{3}$$

If we wished to include spin–orbit interactions we could easily do so by defining the appropriate M operator and using form (3).

These results could have been formally derived from the form $\int [a^+(x) + a(x)]\psi^+(x)\psi(x) d^s x$ by using the completeness property of the basis $\{\overrightarrow{g_{i,\beta}}\}$.[7] Any of these forms would require an infinite renormalization. This is indicated by the fact that V is not defined on vectors of the form $\chi \otimes \phi$ where $\chi \in \mathfrak{F}_a$, $\phi \in \mathfrak{F}_\psi$, $(\phi, N_\psi \phi) \neq 0$, as a simple calculation will show.[8] For our purpose, the summation form for V is preferable to the integral form since every term in the sum is a well-defined operator on $\mathfrak{F}$, whereas the integrand is as ill-defined as is the entire integral. We shall obtain a *bona fide* operator by restricting the sums ("mode cutoffs").

IV. THE CUTOFF INTERACTION

We now show how to introduce into the interaction the usual configuration space and momentum space cutoffs, as well as less conventional mode cutoffs obtained by restricting some of the infinite sums.

Suppose that we wished to describe the situation in which the interaction vanished outside some volume $\mathcal{U}$ of configuration space. We can introduce this effect into V by altering the coefficients D_{ijk}, which control the space behavior, as follows:

Let $\hat{\rho}$ be the characteristic function for $\mathcal{U}$,

$$\hat{\rho}(x) = \begin{cases} 1, & x \in \mathcal{U} \\ 0, & x \notin \mathcal{U}. \end{cases}$$

We then obtain the desired behavior by the change

$$D_{ijk} \to \int d^s x\, \hat{\rho}(x) \hat{g}_i(x)^\dagger \hat{g}_j(x)^\dagger \hat{g}_k(x).$$

For more generality we shall only require $\hat{\rho} \in \mathcal{L}^\infty(\mathbb{R}^s)$. This cutoff is then removed by taking the limit $\hat{\rho} \to 1$ in some appropriate sense.

Alternatively, $\hat{\rho}$ may be removed from the D_{ijk} coefficients and placed instead in the fields as follows:

Let $\hat{\rho} \in \mathcal{L}^\infty(\mathbb{R}^s)$, and let $\mathcal{O}(\rho)$ be the bounded operator on $\mathcal{L}^2(\mathbb{R}^s)$ defined as in the previous section. We then extend $\mathcal{O}(\rho)$ to an operator on $\mathcal{K}_a^{(1)}$ (using the same notation) by defining $(\mathcal{O}(\rho)\vec{f})_\alpha = \mathcal{O}(\rho)f_\alpha$ for all $\vec{f} = \{f_\alpha\}$ in $\mathcal{K}_a^{(1)}$. We then define creation and annihilation operators smeared (regularized) in configuration space by $a_{\hat{\rho}}^{\#}(\vec{f}) = a^{\#}(\mathcal{O}(\rho)\vec{f})$. Recalling Eq. (2) of Sec. III we then have the identity

$$\sum_{i=0}^{\infty} \int d^s x\, \hat{\rho}(x)\hat{g}_i(x)^\dagger \hat{g}_j(x)^\dagger \hat{g}_k(x)\, a^\dagger(\overrightarrow{g_{i,\alpha}})\chi = \sum_{i=0}^{\infty} \int d^s x\, \hat{g}_i(x)^\dagger \hat{g}_j(x)^\dagger \hat{g}_k(x)\, a_{\hat{\rho}}^\dagger(\overrightarrow{g_{i,\alpha}})\chi$$

for all $\chi \in D(N_a^{1/2})$. The coefficient is now just D_{ijk}.

Suppose that we wish instead to have a momentum space cutoff. We follow a similar procedure, but work in momentum space after an inverse Fourier transform:

Let $\rho \in \mathcal{L}^\infty(\mathbb{R}^s)$ and let J_ρ be the bounded operator given by multiplication by ρ: $(J_\rho f)(k) = \rho(k)f(k)$ for all $f \in \mathcal{L}^2(\mathbb{R}^s)$. Then we introduce the cutoff into the interaction by the change

$$D_{ijk} \to \int d^s x\, \widehat{J_\rho g_i}(x)^\dagger \hat{g}_j(x)^\dagger \hat{g}_k(x).$$

The cutoff would be removed by taking the limit $\rho \to 1$ appropriately.

At this stage, singling out g_i for special treatment is only whimsical. This choice puts the cutoff manifestly on the momenta of the bosons. We transfer the location of ρ to the fields as follows:

Keeping the same notation, define the bounded operator J_ρ on $\mathcal{K}_a^{(1)}$ by $(J_\rho\vec{f})_\alpha = J_\rho f_\alpha$ for all $\vec{f} = \{f_\alpha\}$ in $\mathcal{K}_a^{(1)}$. Define the smeared (regularized) creation and annihilation operators by $a_\rho^{\#}(f) = a^{\#}(J_\rho\vec{f})$ (no hat on the ρ). Then use Eq. (2) of Sec. III as before.

In order to introduce cutoffs by restricting the sums, we make the change $D_{ijk} \to K_{ijk}D_{ijk}$ where $\{K_{ijk}\}$ is a sequence of c numbers. For the preservation of formal symmetry we require $K_{ijk}{}^\dagger = K_{ijk}$. This cutoff is removed by taking the limit $K_{ijk} \to 1$ in an appropriate manner.

Since only a momentum-space divergence appears in the model,[3] we shall henceforth take as our general cutoff interaction either of the forms

$$V_{\rho,K} = \sum_{\alpha,\beta,\beta'} V_{\rho,K}(\alpha,\beta,\beta'),$$

$$V_{\rho,K}(\alpha,\beta,\beta') = C_{\alpha\beta\beta'} \sum_{j,k=0}^{\infty} K_{jk} a_\rho^\dagger(\overrightarrow{\bar{g}_j^\dagger * g_{k,\alpha}})\psi^\dagger(\overrightarrow{g_{j,\beta}})\psi(\overrightarrow{g_{k,\beta'}}) + \text{c.c.}, \tag{1'}$$

or

$$V_{\rho,K} = \sum_\alpha V_{\rho,K}(\alpha),$$

$$V_{\rho,K}(\alpha) = \sum_{i=0}^{\infty} K_i a_\rho^\dagger(\overrightarrow{g_{i,\alpha}})\Omega_\psi(M(\bar{g}_i^\dagger, \alpha)) + \text{c.c.} \tag{3'}$$

We shall next give a precise mathematical meaning to the formal manipulations carried out in this section.

V. SELF-ADJOINTNESS OF THE CUTOFF HAMILTONIAN

In this section we shall exhibit classes of cutoffs such that $H_0 + V_{\rho,K}$ is self-adjoint and such that additional useful properties hold. We first exhibit some general theorems on domain mappings for the fields and second-quantized operators. Then we apply these results to the particular case at hand.

Mathematical preliminaries

We will use the following multiple commutator notation: Let A, B be operators such that $BD(A^n) \subseteq D(A^{n-1})$ for $n = 1, 2, \ldots, M$ (where M may be ∞). Then we define

$$(\text{ad}A)^0(B) = B,$$

$$(\text{ad}A)^n(B) = [A, (\text{ad}A)^{n-1}(B)]^\sim, \quad n = 1, 2, \ldots, M,$$

with domain the domain of the right-hand side.

Lemma 1: Let A, B be positive self-adjoint operators in $\mathcal{K}_a^{(1)}, \mathcal{K}_\psi^{(1)}$, respectively. Then for $\vec{f} \in D(A^n)$, $\vec{g} \in D(B^n)$, $n = 0, 1/2, 1, \ldots$, we have

(a) $a^{\#}(\vec{f})D([\Omega_a(A)]^{n+1/2}) \subseteq D([\Omega_a(A)]^n)$,

(b) $\psi^{\#}(\vec{g})D([\Omega_\psi(B)]^n) \subseteq D([\Omega_\psi(B)]^n)$.

Proof: For the cases $n = 0, 1, 2, \ldots$, the result (a) follows by induction using

(i) $[\Omega_a(A), a^\dagger(\vec{f})]^\sim = a^\dagger(A\vec{f})$,

(ii) $[\Omega_a(A), a(\vec{f})]^\sim = -a(A\vec{f})$,

(iii) $D(a^{\#}) \supseteq D[(N_a + I)^{1/2}] \supseteq D[\Omega_a(A)^{1/2}]$,

(iv) for $\chi \in D([\Omega_a(A)]^{n+1/2})$, $[\Omega_a(A)]^n a^{\#}(\vec{f})\chi$

$$= \sum_{p=0}^{n} \binom{n}{p} [\text{ad}\Omega_a(A)]^{(p)}[a^{\#}(\vec{f})]\Omega_a(A)^{n-p}\chi$$

$$= \sum_{p=0}^{n} \sigma(\#)^p \binom{n}{p} a^{\#}(A^p\vec{f})[\Omega_a(A)]^{n-p}\chi,$$

where $\sigma(\#) = +1$ if $a^{\#} = a^\dagger$, and -1 otherwise. (i), and (ii) are contained in Cook's result, (iii) follows from the existence of some constant α such that $A \geq \alpha I$ which implies $\Omega_a(A) \geq \alpha N_a$, and (iv) is a known multiple commutator result.[9]

For the cases $n = \frac{1}{2}, \frac{3}{2}, \ldots$, we use the fact that if G is any closed operator, $D(G^\dagger G)$ is a core for G.[10] Thus for any $\chi \in D([\Omega_a(A)]^{n+1/2})$, there exists a sequence $\{\chi_t\}$, $\chi_t \in D([\Omega_a(A)]^{2n+1})$ for all t, with χ_t converging strongly to χ and $[\Omega_a(A)]^{n+1/2}\chi_t$ converging strongly to $[\Omega_a(A)]^{n+1/2}\chi$. We then have

$$\|[\Omega_a(A)]^n a(\vec{f})(\chi_t - \chi_s)\|^2 \leq \sum_{p=0}^{n-1/2} \sum_{p'=0}^{n+1/2} (-1)^{p+p'} \binom{n-1/2}{p} \binom{n+1/2}{p'} \times \|a^\dagger(A^{p'}\vec{f})a(A^p\vec{f})[\Omega_a(A)]^{n-(1/2)-p}(\chi_t - \chi_s)\| \times \|\Omega_a(A)^{n-p'+(1/2)}(\chi_t - \chi_s)\|$$

which converges to zero as $t, s \to \infty$, since

$$\|a^{\#}(\vec{h})\phi\| \leq \|\vec{h}\|_a \|(N_a + I)^{1/2}\phi\| \text{ and } \Omega_a(A) \geq \alpha N_a.$$

Since $[\Omega_a(A)]^n$ is self adjoint, it is closed; hence $a(\vec{f})\chi \in D([\Omega_a(A)]^n)$. The proof for case (b) is also similar, the better domain condition coming from the fact that $\psi^{\#}(\vec{g})$ is a bounded operator.

Lemma 2: Let A, B be operators in $\mathfrak{K}^{(1)}$ such that B is positive, and for all $\chi \in D(B)$, $\|A\chi\| \leq \|B\chi\|$. Then for all $\phi \in D(N^{1/2}\Omega(B))$, $\|\Omega(A)\phi\| \leq \|N^{1/2}\Omega(B)\phi\|$.

Proof: Since the number operator N commutes with $\Omega(A), \Omega(B)$, it suffices to prove the result for $\phi \in D(\Omega(B)) \cap \mathfrak{K}^{(l)}$ for each $l = 0, 1, 2, \cdots$. Consider ϕ, a finite linear combination of vectors of the form $\chi_1 \otimes \chi_2 \otimes \cdots \otimes \chi_l$ with $\chi_i \in D(B), i = 1, 2, \ldots, l$. Then, defining the j-fold tensor product of the identity with itself to be $I^{(j)}$,

$$\begin{aligned}\|\Omega(A)\phi\| &= \Big\|\sum_{i=1}^{l} I^{(i-1)} \otimes A \otimes I^{(l-i)}\phi\Big\| \\ &\leq \sum_{i=1}^{l} \|I^{(i-1)} \otimes A \otimes I^{(l-i)}\phi\| \\ &\leq \sum_{i=1}^{l} \|I^{(i-1)} \otimes B \otimes I^{(l-i)}\phi\|.\end{aligned}$$

If a_i, b_i are any real numbers, then

$$\left(\sum_{i=1}^{l} a_i b_i\right)^2 \leq \sum_{i=1}^{l} a_i^2 \sum_{j=1}^{l} b_i^2$$

so that choosing $b_i \equiv 1$, we have

$$\sum_{i=1}^{l} a_i \leq l^{1/2}\left(\sum_{i=1}^{l} a_i^2\right)^{1/2}.$$

Thus

$$\begin{aligned}\|\Omega(A)\phi\| &\leq l^{1/2}\left(\sum_{i=1}^{l} \|I^{(i-1)} \otimes B \otimes I^{(l-i)}\phi\|^2\right)^{1/2} \\ &\leq l^{1/2}\|\Omega(B)\phi\| = \|N^{1/2}\Omega(B)\phi\|,\end{aligned}$$

the last inequality coming from the positivity of B. Since vectors of the form considered for ϕ form a core for $\Omega(B)$ restricted to $\mathfrak{K}^{(l)}$, the results extend to all of $D(\Omega(B)) \cap \mathfrak{K}^{(l)}$.

Lemma 3: Let A, B be closed operators in $\mathfrak{K}^{(1)}$ such that

(a) A is strictly positive and self-adjoint ($A \geq \eta I$, for some positive real number η),
(b) B is bounded.
(c) there exists a fixed integer M such that for all $n = 0, 1, 2, \ldots, M$, $BD(A^{n+1}) \subseteq D(A^n)$, and
(d) there exist positive constants $c_n, n = 0, 1, \ldots, M$ such that $\|(\mathrm{ad}A)^n(B)\chi\| \leq c_n\|A^n\chi\|$ for all $\chi \in D(A^n)$.

Then $\Omega(B)D(\Omega(A)^{n+1}) \subseteq D(\Omega(A)^n)$ and

$$\begin{aligned}&\|\Omega(A)^n\Omega(B)\chi\| \\ &\leq \left(\sum_{p=0}^{n} c_p\right)\|N\Omega(A)^n\chi\| \leq \left(\sum_{p=0}^{n} c_p\right)\eta^{-1}\|\Omega(A)^{n+1}\chi\|,\end{aligned}$$

for all $\chi \in D(\Omega(A)^{n+1})$.

Proof: We first show the result to be true on $\mathfrak{K}^{(l)}$, l a positive integer. We again recall that the l-fold tensor product of $D(A^{n+1})$ with itself is a core for $\Omega(A)^{n+1}$ restricted to $\mathfrak{K}^{(l)}$. Therefore, consider χ a finite linear combination of vectors of the form $\chi_1 \otimes \chi_2 \otimes \cdots \otimes \chi_l$, $\chi_i \in D(A^{n+1}), i = 1, 2, \ldots, l$. We wish to show that $\|\Omega(A)^n\Omega(B)\chi\| < \infty$. We recall that for bounded operators B_1, B_2, $[\Omega(B_1), \Omega(B_2)]^{\sim} = \Omega([B_1, B_2])$ so that, in general, $[\mathrm{ad}\Omega(B_1)]^p[\Omega(B_2)] = \Omega[(\mathrm{ad}B_1)^p(B_2)]$. Because of condition (c) and the particular nature of χ, we then know

$$\begin{aligned}\|\Omega(A)^n\Omega(B)\chi\| &= \Big\|\sum_{p=0}^{n} [\mathrm{ad}\Omega(A)]^p[\Omega(B)]\Omega(A)^{n-p}\chi\Big\| \\ &= \Big\|\sum_{p=0}^{n} \Omega[(\mathrm{ad}A)^p(B)]\Omega(A)^{n-p}\chi\Big\| \\ &\leq \|\Omega(B)\Omega(A)^n\chi\| \\ &\quad + \sum_{p=1}^{n} \|\Omega[(\mathrm{ad}A)^p(B)]\Omega(A)^{n-p}\chi\|.\end{aligned}$$

Since A is strictly positive, there is a positive number η such that $A \geq \eta I$ and thus $\Omega(A) \geq \eta\Omega(I) = \eta N$. Since $\Omega(A), \eta N$ commute and are positive, it follows that for all $\phi \in D(\Omega(A))$, $\|N\phi\| \leq (1/\eta)\|\Omega(A)\phi\|$. B is bounded, and, by (d), we know that $\|\Omega(B)\psi\| \leq c_0\|N\psi\|$ for all $\psi \in D(N)$. Thus we estimate the first term by

$$\|\Omega(B)\Omega(A)^n\chi\| \leq c_0\|N\Omega(A)^n\chi\| \leq c_0\eta^{-1}\|\Omega(A)^{n+1}\chi\|.$$

The remaining terms may be estimated using Lemma 2, and conditions (a), (d), and the result $\|\Omega(A^p)\chi\| \leq \|\Omega(A)^p\chi\|$ which follows from the positivity of A (Ref. 11):

$$\begin{aligned}\|\Omega[(\mathrm{ad}A)^p(B)]\Omega(A)^{n-p}\chi\| &\leq c_p\|N^{1/2}\Omega(A^p)\Omega(A)^{n-p}\chi\| \\ &= c_p\|\Omega(A^p)N^{1/2}\Omega(A)^{n-p}\chi\| \\ &\leq c_p\|\Omega(A)^pN^{1/2}\Omega(A)^{n-p}\chi\| \\ &= c_p\|N^{1/2}\Omega(A)^n\chi\| \leq c_p\|N\Omega(A)^n\chi\| \\ &\leq c_p\eta^{-1}\|\Omega(A)^{n+1}\chi\|.\end{aligned}$$

We, therefore, have obtained

$$\|\Omega(A)^n\Omega(B)\chi\| \leq \sum_{p=0}^{n} c_p\|N\Omega(A)^n\chi\| \leq \sum_{p=0}^{n} c_p\eta^{-1}\|\Omega(A)^{n+1}\chi\|.$$

This inequality extends to all $\chi \in D(\Omega(A)^{n+1}) \cap \mathfrak{K}^{(l)}$. Since the inequality is independent of l and since the number operator N reduces $\Omega(A)$, the result extends to all $\chi \in D(\Omega(A)^{n+1})$. QED

There are several variations of the previous two lemmas. For instance, Lemma 2 can be altered to give the inequality without the factor of $N^{1/2}$ if one also knows $A \leq B$. If the factor of $N^{1/2}$ does not appear in this estimate, then proceeding to Lemma 3 one may obtain the same results with the weakened condition $\|(\mathrm{ad}A)^n(B)\chi\| \leq c_n\|A^{n+1}\chi\|$. These variations are of general interest, but will not be pursued here since they are not needed for the present model. Condition (c) for Lemma 3 can be weakened to requiring only that there exists a dense domain D such that $BD \subseteq D(A^n)$. Then the fact that A, B are closed implies (c) (Ref. 12).

The last two lemmas and variations of them, show how the behavior of operators in $\mathfrak{K}^{(1)}$ determines the behavior of the second-quantized operators and how some hard estimates may be simplified by considering the corresponding estimates in $\mathfrak{K}^{(1)}$. As an example, we note that for A, B as above, and for all $\chi \in D(\Omega(A)^{n+1})$ we have $[\mathrm{ad}\Omega(A)^{(n)}][\Omega(B)]\chi = \Omega[(\mathrm{ad}A)^n(B)]\chi$, and for $n > 0$,

$$\|[\mathrm{ad}\Omega(A)]^n[\Omega(B)]\chi\| = \|\Omega[(\mathrm{ad}A)^{(n)}(B)]\chi\| \le c_n\|N^{1/2}\Omega(A^n)\chi\| \le c_n\|N^{1/2}\Omega(A)^n\chi\|.$$

With this type of analysis in mind, we now examine a particular choice for A, B. We will use the notation $k = (k_1, \ldots, k_s) \in \mathbb{R}^s$, and $\epsilon(k), h_\psi$, and $\Theta(f)$ will denote the operators defined in Secs. II and III.

Lemma 4: Let f be such that $h_\psi{}^p f \in \mathcal{L}^1(\mathbb{R}^s)$, f real, for $p = 0, 1, 2, \ldots, n$. Then for all $\phi \in D(h_\psi{}^n)$,

$$\|(\mathrm{ad}h_\psi)^n(\Theta(f))\phi\| \le 4\theta^n \sum_{j=0}^{n} 2^j\binom{n}{j}\|h_\psi{}^{n-j}f\|_1\|h_\psi{}^j\phi\|,$$

where

$$\theta = \begin{cases} 2, \text{ if } \epsilon(k) = m + |k|^2(2m)^{-1} \text{ (Galilean)} \\ 1, \text{ if } \epsilon(k) = [|k|^2 + m^2]^{1/2} \text{ (relativistic)} \\ \tfrac{1}{2}, \text{ if } \epsilon(k) = m \text{ (static)}. \end{cases}$$

Proof: We first establish the estimate $\epsilon(k) \le \theta[\epsilon(k-p) + \epsilon(p)]$.

For the static case $\epsilon(k) = m$, the estimate is trivial with $\theta = \frac{1}{2}$.

For the Galilean case $\epsilon(k) = m + |k|^2(2m)^{-1}$, we have $|k| \le |k-p| + |p|$ so that $|k|^2 \le 2|k-p|^2 + 2|p|^2$. Therefore the estimate holds with $\theta = 2$.

For the relativistic case, $\epsilon(k) = [|k|^2 + m^2]^{1/2}$ we have

$$\begin{aligned}\epsilon(k)^2 &= |k|^2 + m^2 \le |k-p|^2 + |p|^2 + 2|k-p||p| + m^2 \\ &\le |k-p|^2 + m^2 + |p|^2 + m^2 \\ &\quad + 2[|k-p|^2 + m^2]^{1/2}[|p|^2 + m^2]^{1/2} \\ &= [\epsilon(k-p) + \epsilon(p)]^2.\end{aligned}$$

Thus $\theta = 1$.

Let $\phi \in D(h_\psi{}^n)$ and assume ϕ, f are positive. Then,

$$\begin{aligned}&\|h_\psi{}^n(f*\phi)\|^2 \\ &= \int f(k-q)^\dagger\phi(q)^\dagger\epsilon(k)^{2n}f(k-p)\phi(p)d^sp\ d^sq\ d^sk \\ &\le \theta^{2n}\int f(k-q)^\dagger\phi(q)^\dagger[\epsilon(k-q) + \epsilon(q)]^n \\ &\times [\epsilon(k-p) + \epsilon(p)]^n f(k-p)\phi(p)d^sp\ d^sq\ d^sk \\ &= \theta^{2n}\sum_{i,j=0}^{n}\binom{n}{i}\binom{n}{j}\int(h_\psi{}^{n-i}f)(k-q)^\dagger(h_\psi{}^i\phi)(q)^\dagger \\ &\times (h_\psi{}^{n-j}f)(k-p)(h_\psi{}^j\phi)(p)d^sp\ d^sq\ d^sk \\ &\le \theta^{2n}\sum_{i,j=0}^{n}\binom{n}{i}\binom{n}{j}\|(h_\psi{}^{n-i}f)*(h_\psi{}^i\phi)\|\|(h_\psi{}^{n-j}f)*(h_\psi{}^j\phi)\|.\end{aligned}$$

More simply

$$\begin{aligned}\|h_\psi{}^n(f*\phi)\| &\le \theta^n\sum_{i=0}^{n}\binom{n}{i}\|(h_\psi{}^{n-i}f)*(h_\psi{}^i\phi)\| \\ &\le \theta^n\sum_{i=0}^{n}\binom{n}{i}\|\widehat{h_\psi{}^{n-i}f}\|_\infty\|h_\psi{}^i\phi\| \\ &= \theta^n\sum_{i=0}^{n}\binom{n}{i}\|h_\psi{}^{n-i}f\|_1\|h_\psi{}^i\phi\|.\end{aligned}$$

More generally, $f = f_+ - f_-, \phi = \phi_+ - \phi_-$ where f_+, f_-, ϕ_+, ϕ_- are positive. Then by the linearity of the convolution we have

$$\|h_\psi{}^n(f*\phi)\| \le \theta^n\sum_{i=0}^{n}\binom{n}{i}\begin{cases}\|h_\psi{}^{n-i}f_+\|_1\|h_\psi{}^i\phi_+\| \\ +\|h_\psi{}^{n-i}f_-\|_1\|h_\psi{}^i\phi_+\| \\ +\|h_\psi{}^{n-i}f_+\|_1\|h_\psi{}^i\phi_-\| \\ +\|h_\psi{}^{n-i}f_-\|_1\|h_\psi{}^i\phi_-\|\end{cases}.$$

Since ϕ_+, ϕ_- have disjoint supports, we have

$$\begin{aligned}\|h_\psi{}^i\phi\|^2 &= \int\epsilon(k)^{2i}[\phi_+(k)^2 + \phi_-(k)^2]d^sk \\ &= \|h_\psi{}^i\phi_+\|^2 + \|h_\psi{}^i\phi_-\|^2\end{aligned}$$

or $\|h_\psi{}^i\phi_\pm\| \le \|h_\psi{}^i\phi\|$. Since f_+, f_- have disjoint supports, we know $\|h_\psi{}^{n-i}f_+\|_1 + \|h_\psi{}^{n-i}f_-\|_1 = \|h_\psi{}^{n-i}f\|_1$. Summarizing, we have for ϕ real,

$$\|h_\psi{}^n(f*\phi)\| \le 2\theta^n\sum_{i=0}^{n}\binom{n}{i}\|h_\psi{}^{n-i}f\|_1\|h_\psi{}^i\phi\|.$$

Similarly, if ϕ is complex, replace the 2 with a 4. This insures that $\Theta(f)D(h_\psi{}^n) \subseteq D(h_\psi{}^n)$. Recalling that

$$(\mathrm{ad}h_\psi)^n(\Theta(f))\phi = \sum_{p=0}^{n}(-1)^{n-p}\binom{n}{p}h_\psi{}^p\Theta(f)h_\psi{}^{n-p}\phi$$

and the identity $\binom{n}{p}\binom{n-p}{j-p} = \binom{n}{j}\binom{j}{p}$ we finally have

$$\begin{aligned}\|(\mathrm{ad}h_\psi)^n(\Theta(f))\phi\| &\le \sum_{p=0}^{n}\binom{n}{p}\|h_\psi{}^p\Theta(f)h_\psi{}^{n-p}\phi\| \\ &\le 4\theta^n\sum_{p=0}^{n}\sum_{i=0}^{n-p}\binom{n}{p}\binom{n-p}{i}\|h_\psi{}^{n-p-i}f\|_1\|h_\psi{}^{p+i}\phi\| \\ &= 4\theta^n\sum_{p=0}^{n}\sum_{j=p}^{n}\binom{n}{p}\binom{n-p}{j-p}\|h_\psi{}^{n-j}f\|_1\|h_\psi{}^j\phi\| \\ &= 4\theta^n\sum_{p=0}^{n}\sum_{j=p}^{n}\binom{n}{j}\binom{j}{p}\|h_\psi{}^{n-j}f\|_1\|h_\psi{}^j\phi\| \\ &= 4\theta^n\sum_{p=0}^{n}\sum_{j=0}^{n}\binom{n}{j}\binom{j}{p}\|h_\psi{}^{n-j}f\|_1\|h_\psi{}^j\phi\| \\ &= 4\theta^n\sum_{j=0}^{n}\binom{n}{j}2^j\|h_\psi{}^{n-j}f\|_1\|h_\psi{}^j\phi\|.\end{aligned}$$

Corollary 1: Since $m^{-1}h_\psi \ge I$ we know that $\|h_\psi{}^j\phi\| \le m^{j-n}\|h_\psi{}^n\phi\|$ so that there is a constant c_n depending on θ, f such that $\|(\mathrm{ad}h_\psi)^n(\Theta(f))\phi\| \le c_n\|h_\psi{}^n\phi\|$ for all $\phi \in D(h_\psi{}^n)$.

Application to the interaction

Assertion 1: Let $\{g_i\}_0^\infty$ be an orthonormal basis of $\mathcal{L}^2(\mathbb{R}^s)$ such that $g_i \in \mathcal{S}(\mathbb{R}^s)$ for all i. Let $\rho \in \mathcal{L}^\infty(\mathbb{R}^s)$. Let $L \ge 1$ be a fixed integer, and let $\{K_{j,k}\}$ be a sequence of complex numbers such that $K_{jk}^* = K_{kj}$ and such that the sums

$$F(n_1, n_2, n_3) \equiv \sum_{j,k=0}^{\infty}|K_{jk}|\|h_\psi{}^{n_1}g_j\|\|h_\psi{}^{n_2}g_k\|\|h_a{}^{n_3}J_\rho\sqrt{g_j^\dagger}*g_k\|_a$$

converge for all positive integral n_1, n_2, n_3 such that $0 \le n_1 + n_2 + n_3 \le L$. Let

$$V_{\rho,k}^{\sigma,\nu}(\alpha,\beta,\beta') = C_{\alpha,\beta,\beta'}\sum_{j=1}^{\sigma}\sum_{k=1}^{\nu}K_{jk}a_\rho^+(\overrightarrow{\sqrt{g_j^\dagger}*g_{k,\alpha}})\psi^+(\overrightarrow{g_{j,\beta}})\psi(\overrightarrow{g_{k,\beta'}}) + \text{c.c.}$$

Then $V_{\rho,k}^{\sigma,\nu}(\alpha,\beta,\beta')$ converges strongly on $D(N_a{}^{1/2})$ as $\sigma, \nu \to \infty$, can be rearranged arbitrarily, and defines a

symmetric operator $V_{\rho,k}(\alpha,\beta,\beta')$ in the limit. Defining $V_{\rho,k}=\sum_{\alpha,\beta,\beta'} V_{\rho,k}(\alpha,\beta,\beta')$, we have $H_0+V_{\rho k}$ is self-adjoint and $D(H_0^n)=D([H_0+V_{\rho k}]^n)$ for $n=0,1,\ldots,L$. Furthermore, defining $R_n \equiv (H_0+V_{\rho k})^n - H_0^n$, then for each n above, there exist constants $\rho(n), \delta(n), \gamma_i(n), i=1,2,3,4$, such that $H_0+V_{\rho k}+\delta(n)\geq 0$ and for all $\chi \in D(H_0^n)$

(i) $\|R_n\chi\| \leq \gamma_1(n)\|[H_0+V_{\rho k}+\delta(n)I]^n\chi\|$,

(ii) $\|(R_n+\delta(n))\chi\| \leq \gamma_2(n)\|[H_0+V_{\rho k}+\delta(n)I]^n\chi\|$,

(iii) $\|H_0^n\chi\| \leq \|[H_0+\rho(n)I]^n\chi\| \leq \gamma_3(n)\|H_0+V_{\rho k}+\delta(n)I]^n\chi\|$,

(iv) $\|[H_0+V_{\rho k}+\delta(n)I]^n\chi\| \leq \gamma_4(n)\|[H_0+\rho(n)I]^n\chi\|$.

Proof: Since $g_i \in \mathcal{S}(\mathbb{R}^s)$, g_j and g_k are in $\cap_{n=0}^{\infty} D(h_\psi{}^n) \equiv C^\infty(h_\psi)$. Similarly $J_\rho \sqrt{g_j}^\dagger * g_k \in \mathcal{S}(\mathbb{R}^s) \subseteq C^\infty(h_a)$. Thus, by Lemma 1, the summand of $V^{\sigma,\nu}_{\rho,k}(\alpha,\beta,\beta')$ maps $D(H_0^{n+1/2}) \to D(H_0^n)$, $n=0,1,2,\cdots$. We shall show that this mapping property carries over to the limit operator $V_{\rho,k}$ for the cases $n=0,1,2,\ldots,L$ by showing that for $q=0,1,2,\ldots,n$, $H_0{}^q V^{\sigma,\nu}_{\rho,k}(\alpha,\beta,\beta')\chi$ converges for all χ in $D((N_a+I)^{1/2}H_0^n)$. The convergence for the cases $q=0$ and $q=n$ along with the fact that H_0^n is closed gives the desired result. Furthermore, by the multiple commutator identity $A^qB=\sum_{p=0}^{q}\binom{q}{p}(\mathrm{ad}A)^p(B)A^{q-p}$ it suffices to show the convergence of

$$(\mathrm{ad}H_0)^p[V^{\sigma,\nu}_{\rho,k}(\alpha,\beta,\beta')]\chi \text{ for all } \chi \text{ in } D[(N_a+I)^{1/2}H_0^n]$$

for $p=0,1,\ldots,n$. We recall the properties

$$\|\psi^{\#}(\vec{g})\| = \|\vec{g}\|_\psi,$$

$$\|a^{\#}_\rho(\vec{f})\phi\| \leq \|J_\rho \vec{f}\|_a \|(N_a+I)^{1/2}\phi\|, \qquad \phi \in D(N_a{}^{1/2}),$$

and $D((N_a+I)^{1/2}) \supseteq D(H_0{}^{1/2})$. Now on $D((N_a+I)^{1/2}H_0^n)$ we have

$$(\mathrm{ad}H_0)^p(V^{\sigma,\nu}_{\rho,k}(\alpha,\beta,\beta'))$$

$$= C_{\alpha,\beta,\beta'} \sum_{\substack{n_1,n_2,n_3=0 \\ p=n_1+n_2+n_3}}^{p} \frac{p!(-1)^{n_3}}{n_1!n_2!n_3!} \sum_{j=1}^{\sigma}\sum_{k=0}^{\nu} K_{jk}$$

$$\times a^+(\overrightarrow{h_a^{n_1} J_\rho \sqrt{g_j}^* * g_{k,\alpha}})\psi^+(\overrightarrow{h_\psi{}^{n_2} g_{j,\beta}})\psi(\overrightarrow{h_\psi{}^{n_3} g_{k,\beta'}})$$

$$+ (-1)^p \text{ c.c.}$$

Thus

$$\|(\mathrm{ad}H_0)^p(V^{\sigma,\nu}_{\rho,k}(\alpha,\beta,\beta'))\chi - (\mathrm{ad}H_0)^p(V^{\sigma'\nu'}_{\rho,k}(\alpha,\beta,\beta'))\chi\|$$

$$\leq 2|C_{\alpha,\beta,\beta'}| \sum_{\substack{n_1,n_2,n_3=0 \\ p=n_1+n_2+n_3}}^{p} \frac{p!}{n_1!n_2!n_3!}\left(\sum_{j=0}^{\sigma}\sum_{k=0}^{\nu} - \sum_{j=0}^{\sigma'}\sum_{k=0}^{\nu'}\right)|K_{jk}|$$

$$\times \|h^{n_1}J_\rho \sqrt{g_j}^\dagger * g_k\|_a \|h_\psi{}^{n_2} g_j\| \|h_\psi{}^{n_3} g_k\| \|(N_a+I)^{1/2}\chi\|$$

which converges as $\sigma,\nu,\sigma',\nu' \to \infty$, for $p \leq L$, by the convergence of the $F(n_1,n_2,n_3)$. In particular with the choice $\sigma=\nu$, $V^{\sigma\sigma}_{\rho k}(\alpha,\beta,\beta')$ is symmetric and, hence, so is $V_{\rho k}$. We also obtain

$$\|(\mathrm{ad}H_0)^p(V_{\rho,k})\chi\|$$

$$\leq 2 \sum_{\alpha,\beta,\beta'} |C_{\alpha,\beta,\beta'}| \sum_{\substack{n_1,n_2,n_3=0 \\ n_1+n_2+n_3=p}}^{p} \frac{p!}{n_1!n_2!n_3!} F(n_1,n_2,n_3)$$

$$\times \|(N_a+I)^{1/2}\chi\|, \qquad \text{for } p < L.$$

Using the additional estimate

$$\|(N_a+I)^{1/2}\chi\| \leq \mu^{-1/2}\|H_{0a}{}^{1/2}\chi\| + \|\chi\|$$

$$\leq \mu^{-1/2}\|H_0{}^{1/2}\chi\| + \|\chi\|,$$

the rest of the assertion follows.[12]

We remark that the above assertion is also true with H_0 replaced by H_{0a} throughout, modulo some terms which no longer contribute. We also notice that any choice of K_{jk}, such that the $F(n_1,n_2,n_3)$ converge for all finite n_1,n_2,n_3, gives the result $C^\infty(H_0)=C^\infty(H_0+V_{\rho k})$. In particular, the choice K_{jk}, identically zero except for a finite number of j,k, gives this result.

Corollary 2: With g_i, ρ as before and $K_{jk} \equiv K_{kj}{}^*$ such that $F(0,0,0)$ converges, then $V_{\rho,k}$ is self-adjoint.

Proof: We have shown that

$$\|V_{\rho,k}\chi\| \leq 2 \sum_{\alpha,\beta,\beta'} |C_{\alpha\beta\beta'}|F(0,0,0)\|(N_a+I)^{1/2}\chi\|$$

so that $V_{\rho,k}(N_a+I)^{-1/2}$ is a bounded operator on $\mathfrak{F}$ with bound $b=2\sum_{\alpha,\beta,\beta'}|C_{\alpha\beta\beta'}|F(0,0,0)$.

Consider vectors of the form

$$\Phi = \phi_0 + \phi_1 + \cdots + \phi_r,$$

where $\phi_i \in \mathcal{K}_a^{(i)} \otimes \mathfrak{F}_\psi$. Vectors of this form will be said to "have at most r bosons." The set of all vectors having at most a finite number of bosons we shall call $\mathfrak{F}_a^0$. $\mathfrak{F}_a^0$ is dense in $\mathfrak{F}$, and $\mathfrak{F}_a^0 \subseteq D(N_a)$. Furthermore $V_{\rho,k}\Phi$ has at most $r+1$ bosons and, hence, is in $D(N_a)$. By induction $(V_{\rho,k})^n\Phi$ has at most $r+n$ bosons, and we have the estimate

$$\|(V_{\rho,k})^n\Phi\| = \|V_{\rho,k}(N_a+I)^{-1/2}(N_a+I)^{1/2}(V_{\rho,k})^{n-1}\Phi\|$$

$$\leq b\|(N_a+I)^{1/2}(V_{\rho,k})^{n-1}\Phi\|$$

$$\leq b(r+n-1+1)^{1/2}\|(V_{\rho,k})^{n-1}\Phi\|$$

$$\leq b^n[(r+n)(r+n-1)\cdots(r)]^{1/2}\|\Phi\|.$$

Thus $V_{\rho,k}$ is a symmetric operator defined on $\mathfrak{F}_a^0$, $\mathfrak{F}_a^0$ is stable under application of $V_{\rho,k}$, and $\sum_{n=0}^{\infty}(|t|^n/n!) \times \|(V_{\rho,k})^n\Phi\| < \infty$ for all $t \in \mathbb{R}, \Phi \in \mathfrak{F}_a^0$. Thus by a theorem of Nelson,[13] $V_{\rho,k}$ is self adjoint.

We conclude this analysis by treating the interaction with the alternate form of cutoff (3′):

Assertion II: Let $\{g_i\}_0^\infty$ be an orthonormal basis for $\mathcal{L}^2(\mathbb{R}^s)$ such that $g_i \in \mathcal{S}(\mathbb{R}^s)$ for all i. Let $\rho \in \mathcal{L}^\infty(\mathbb{R}^s)$ and L be a fixed integer. Let $\{K_i\}$ be a sequence of real numbers such that the sums

$$F(p,q) = \sum_{i=0}^{\infty} |K_i| \|h_a{}^p J_\rho g_i\|_a \|h_\psi{}^q g_i\|_1$$

converge for all positive integers p,q such that $0 \leq p+q \leq L$. Let

$$V^\sigma_{\rho,k} = \sum_{\alpha}\sum_{i=0}^{\sigma} K_i a^+_\rho(\overrightarrow{g_{i,\alpha}})\Omega_\psi(M(\sqrt{g_i}^\dagger,\alpha)) + \text{c.c.}$$

Then the sequence $\{V^\sigma_{\rho,k}\}$ converges strongly on $D(N_a{}^{1/2})$

$\otimes \mathcal{K}_\psi{}^{(n)}$ to a symmetric operator $V_{\rho,k}(n)$, so that $\{V^\sigma_{\rho,k}\}$ converges to the symmetric operator $V_{\rho,k} = \sum_{n=0}^{\infty} \oplus V_{\rho,k}(n)$ defined in $\mathfrak{F}$. Furthermore $H_0 + V_{\rho,k}$ is self adjoint and $D([H_0 + V_{\rho,k}]^p) = D(H_0{}^p)$ for $p = 1, 2, \ldots, L$. In each subspace $\mathfrak{F}_a \otimes \mathcal{K}_\psi{}^{(n)}$ and for each p above, there exist constants $\rho(n,p), \delta(n,p), \gamma_i(n,p), i = 1, 2, 3, 4$ such that (omitting the n,p subscripts), $H_0 + V_{\rho,k} + \delta I$ is positive in $\mathfrak{F}_a \otimes \mathcal{K}_\psi{}^{(n)}$. Defining the remainders $R(p) = [H_0 + V_{\rho,k}(n)]^p - H_0{}^p$ in each $\mathfrak{F}_a \otimes \mathcal{K}_\psi{}^{(n)}$, we also have, for all χ in $D(H_0) \cap \mathfrak{F}_a \otimes \mathcal{K}_\psi{}^{(n)}$,

$$\|R(p)\chi\| \le \gamma_1 \|(H_0 + V_{\rho,k} + \delta I)^p \chi\|,$$
$$\|(R(p) + \delta I)\chi\| \le \gamma_2 \|(H_0 + V_{\rho,k} + \delta I)^p \chi\|,$$
$$\|H_0{}^p \chi\| \le \|(H_0 + \rho I)^p \chi\| \le \gamma_3 \|(H_0 + V_{\rho,k} + \delta I)^p \chi\|,$$
$$\|(H_0 + V_{\rho,k} + \delta I)^p \chi\| \le \gamma_4 \|(H_0 + \rho I)^p \chi\|.$$

Proof: Since the boson number operator $N_\psi \equiv \Omega_\psi(I)$ reduces H_0 and $V^\sigma_{\rho,k}$ and since $\mathfrak{F} = \sum_{n=0}^{\infty} \oplus \mathfrak{F}_a \otimes \mathcal{K}_\psi{}^{(n)}$, we may restrict all consideration to $\mathfrak{F}_a \otimes \mathcal{K}_\psi{}^{(n)}$ for a general n. From here, the proof is similar to that of the previous assertion. We first consider the multiple commutators of $H_{0\psi}{}^{(1)}$ with $M(\check{g}_i^\dagger, \alpha)$ in order to obtain estimates on the commutators of H_0 with $V_{\rho,k}$.

We recall that, letting $\vec{f} = \{f_\beta\} \in \mathcal{K}_\psi{}^{(1)}$, we have $(H_{0\psi}{}^{(1)}\vec{f})_\beta = h_\psi f_\beta$ and

$$[M(\check{g}_i^\dagger, \alpha)\vec{f}]_\beta = \sum_{\beta'} C_{\alpha\beta\beta'} \Theta(\check{g}_i^\dagger) f_{\beta'} = \sum_{\beta'} C_{\alpha\beta\beta'} \check{g}_i^\dagger * f_{\beta'}.$$

Since $\check{g}_i^\dagger \in \mathcal{S}(\mathbb{R}^s)$, $f_{\beta'} \in \mathcal{L}^2(\mathbb{R}^s)$, then $\check{g}_i^\dagger * f_{\beta'} \in \mathcal{S}(\mathbb{R}^s)$. But $\mathcal{S}(\mathbb{R}^s) \subseteq C^\infty(h_\psi)$ so that in particular, $M(\check{g}_i^\dagger, \alpha) D(H_{0\psi}{}^p) \subseteq C^\infty(H_{0\psi}) \subseteq D(H_{0\psi}{}^p)$ and $M(\check{g}_i^\dagger, \alpha) D(H_0{}^p) \subseteq D(H_0{}^p)$. Furthermore

$$[(\mathrm{ad}H_{0\psi}{}^{(1)})^p (M(\check{g}_i^\dagger, \alpha))\vec{f}]_\beta = \sum_{\beta'} C_{\alpha\beta\beta'} (\mathrm{ad}h_\psi)^p (\Theta(\check{g}_i^\dagger)) f_{\beta'},$$

from which we obtain

$$\|(\mathrm{ad}H_{0\psi}{}^{(1)})^p (M(\check{g}_i^\dagger, \alpha))\vec{f}\|^2 = \sum_\beta \|[(\mathrm{ad}H_{0\psi}{}^{(1)})^p (M(\check{g}_i^\dagger, \alpha))\vec{f}]_\beta\|^2 = \sum_\beta \Big\| \sum_{\beta'} C_{\alpha\beta\beta'} (\mathrm{ad}h_\psi)^p (\Theta(\check{g}_i^\dagger)) f_{\beta'} \Big\|^2$$
$$\le \sum_\beta \left(\sum_{\beta'} 4 |C_{\alpha\beta\beta'}| \theta^p \sum_{j=0}^{p} 2^j \binom{p}{j} \|h_\psi{}^{p-j} \check{g}_i^\dagger\|_1 \|h_\psi{}^j f_{\beta'}\| \right)^2.$$

The last factor we overestimate by $\|h_\psi{}^j f_{\beta'}\| \le \|H_{0\psi}{}^{(1)j} \vec{f}\|$ and use the result $\check{\epsilon}(k)^\dagger = \epsilon(k)$ to obtain

$$\|(\mathrm{ad}H_{0\psi}{}^{(1)})^p (M(\check{g}_i^\dagger, \alpha))\vec{f}\| \le 4 \left[\sum_\beta \left(\sum_{\beta'} |C_{\alpha\beta\beta'}|^2 \right) \right]^{1/2} \theta^p \sum_{j=0}^{p} 2^j \binom{p}{j} \|h_\psi{}^{p-j} g_i\|_1 \|H_{0\psi}{}^{(1)j} \vec{f}\|.$$

Since $H_{0\psi}^{(1)} \ge mI$, there exists a constant c_p such that the above may be overestimated by $c_p \|H_{0\psi}{}^{(1)p} \vec{f}\|$, so that we may apply Lemmas 2 and 3. We, therefore, have, for all $\chi \in D(N_a{}^{1/2} H_0{}^q) \cap \mathfrak{F}_a \otimes \mathcal{K}_\psi{}^{(n)}$, $q > 0$,

$$\|(\mathrm{ad}H_0)^q (V_{\rho,k})\chi\| \le \sum_{p=0}^{q} \binom{q}{p} \sum_{i=0}^{\sigma} |K_i| \sum_\alpha \times \|(\mathrm{ad}H_{0\alpha})^p [a_\rho^+(\overrightarrow{g_{i,\alpha}}) + a_\rho^{\vee}(\overrightarrow{\check{g}_{i,\alpha}})] (\mathrm{ad}H_{0\psi})^{q-p} [\Omega_\psi(M(\check{g}_i^\dagger, \alpha))]\chi\|$$
$$\le \sum_{p=0}^{q} \binom{q}{p} \sum_{i=0}^{\sigma} |K_i| \sum_\alpha \|h_a{}^p J_\rho g_i\|_a \left[\sum_\beta \left(\sum_{\beta'} |C_{\alpha\beta\beta'}| \right)^2 \right]^{1/2} \theta^{q-p} 4 \times \sum_{j=0}^{q-p} 2^j \binom{q-p}{j} \|h_\psi{}^{q-p-j} g_i\|_1 m^{(p+j-q)} \|(2N_a + 1)^{1/2} N_\psi{}^{1/2} H_{0\psi}{}^{q-p} \chi\|$$
$$\le 4 \left[\sum_\beta \left(\sum_{\beta'} |C_{\alpha\beta\beta'}| \right)^2 \right]^{1/2} \sum_{p=0}^{q} \binom{q}{p} \left(\frac{m}{\theta} \right)^{q-p} \sum_{j=0}^{q-p} (2m)^j \binom{q-p}{j} \times F(p, q-p-j) \|(2N_a + I)^{1/2} N_\psi{}^{1/2} H_{0\psi}{}^{q-p} \chi\|.$$

From the next to the last inequality and the known convergence of $F(p, q-p-j)$, we deduce the existence of the strong limit of $(\mathrm{ad}H_0)^q (V^\sigma_{\rho,k})$ as $\sigma \to \infty$. From $H_{0\psi}{}^{q-p} \le m^{-p} H_{0\psi}{}^q \le m^{-p} H_0{}^q$ and letting

$$E_q = \sum_\alpha \left[\sum_\beta \left(\sum_{\beta'} |C_{\alpha\beta\beta'}|^2 \right) \right]^{1/2} \sum_{p=0}^{q} \binom{q}{p} \left(\frac{m}{\theta} \right)^{q-p} \sum_{j=0}^{q-p} (2m)^j \binom{q-p}{j} \times 4F(p, q-p-j) m^{-p},$$

we then have

$$\|(\mathrm{ad}H_0)^q (V^\sigma_{\rho,k})\chi\| \le E_q \|(2N_a + I)^{1/2} N_\psi{}^{1/2} H_0{}^q \chi\|$$

which holds for all σ including the limit as $\sigma \to \infty$. For $q = 0$, the above also holds with $N_\psi{}^{1/2}$ replaced by N_ψ. Similarly,

$$\|H_0{}^p V^\sigma_{\rho,k} \chi\| \le \sum_{q=0}^{p} \binom{p}{q} \|(\mathrm{ad}H_0)^q (V^\sigma_{\rho,k}) H_0{}^{p-q} \chi\| \le \sum_{q=1}^{p} \binom{p}{q} E_q \|(2N_a + 1)^{1/2} N_\psi{}^{1/2} H_0{}^p \chi\| + E_0 \|(2N_a + 1)^{1/2} N_\rho H_0{}^p \chi\|$$

holds for all σ including the limit $\sigma \to \infty$. Thus the limit operator exists, is symmetric, and $V_{\rho,k}$ maps $D(H_0{}^{p+1})$ into $D(H_0{}^p)$. These conditions imply all the desired results.[12]

We remark that as long as all $F(p,q)$ exist, then $C^\infty(H_0) = C^\infty(H_0 + V_{\rho,k})$. In particular this occurs if K_i is nonzero only for a finite number of i's.

Corollary 3: With the conditions of the previous assertion, $V_{\rho,k}$ is self-adjoint.

The proof of this corollary is identical to the previous one.

CONCLUSION

We have developed a different mathematical approach for quantum field theory and applied it to a model which exhibits an infinite mass renormalization. By this framework we avoided referring to ill-defined products of distributions. After introducing an unconventional type of cutoff, which we called a "mode cutoff," we proved the self-adjointness of the cutoff Hamiltonian. The removal of the cutoff and the relation between the asymptotic and interpolating fields, will be the topic of a forthcoming paper.

ACKNOWLEDGMENT

This material is an extension of the author's thesis

work. That work and the present extension could not have been performed without the encouragement, criticism, suggestions, and guidance from Professor Gérard G. Emch.

[1]J. M. Cook, Trans. Am. Math. Soc. **74,** 222 (1953).
[2]G. Svetlichny, J. Math. Phys. **11,** 3433 (1970).
[3]S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1962), pp. 339–351.
[4]V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. USA **34,** 211 (1948), and references therein. In the case of $2n_{a1}+1\equiv d$ spin components and three space dimensions, the measure should properly be $d\mu(p)=\omega(p)^{-d-1}d^3p$; however, we will neglect this for simplicity of notation only. The choice of measure only affects the details of calculation, and not the method. The analogous results for the Galilee group may be found in J.-M. Levy-Leblond, J. Math. Phys. **4,** 776 (1963).
[5]This is a slight generalization of a result in F. E. Schroeck Jr., thesis (Univ. of Rochester, 1970), pp. 106–09.
[6]F. E. Schroeck Jr., J. Math. Phys. **12,** 1849 (1971).
[7]F. E. Schroeck Jr., Ref. 5, Chap. 1.
[8]Reference 5, pp. 80–82.
[9]Reference 5, pp. 94–96.
[10]T. Kato, *Perturbation Theory for Linear Operators* (Springer-Verlag, New York, 1966), p. 275.
[11]F. E. Schroeck Jr. (1971), Ref. 6, Theorem 2g.
[12]F. E. Schroeck Jr., "A Note on Kato's Perturbation Theory," Florida Atlantic University (1971).
[13]E. Nelson, Ann. Math. **7,** 583 (1959).

The anisotropic Kepler problem in two dimensions

Martin C. Gutzwiller
IBM Thomas J. Watson Research Center, P. O. Box 218, Yorktown Heights, New York 10598
(Received 6 April 1972)

The classical trajectories are investigated for a particle with an anisotropic mass tensor in an ordinary Coulomb potential. For negative energies (bound states) these trajectories are isomorphic with the geodesics on a Riemannian surface which can be immersed in a Euclidean space and which looks like a "double snail." For vanishing energy (or near a collision) the equations of motion can be reduced to an autonomous system whose trajectories can be fully discussed. On the basis of extensive numerical computations, it has been possible to give a simple, yet complete description of all trajectories with negative energy. A binary sequence is associated with any trajectory where each term gives the sign of the position coordinate for the consecutive intersections with the "heavy" axis. If the binary sequence is represented by two real numbers, a one-to-one and continuous map from them to the initial conditions can be constructed. Thus, the Poincaré map for the trajectories is equivalent with a shift of the binary Bernoulli scheme (tossing a coin), and all the periodic orbits can be obtained systematically. A number of these are discussed to illustrate the consequences of the isomorphism with the binary sequences. Finally, the baker transformation and its use for finding the trajectories which connect any two given endpoints, is mentioned.

This paper is concerned with classifying all the classical trajectories of a particular dynamical system. The reasons for investigating this special case and for emphasizing certain features are explained in the first section. Briefly, the motion of a charged particle with an anisotropic mass tensor in an ordinary Coulomb potential is of interest when one tries to understand the relations between classical and quantum mechanics. No fruitful progress in this area seems possible unless some specific examples can be fully discussed; but these examples have to be nontrivial. In two dimensions, there should be no constant of motion besides the energy. The results are not applied to the problem of connecting quantum and classical mechanics in this report, because there is enough work to be done just to describe the classical system without relating it to the corresponding quantum system. Therefore, with the exception of the section entitled "Background," anybody interested in classical dynamical systems can follow the discussion. The mode of presentation, however, does not agree with the generally accepted rules of the trade. The latter requires a rigorously logical advance, starting from the equations of motion and ending with the precise statements of theorems, including mathematically clean proofs. I have not been able to construct such proofs for most of the results, although I am convinced of their correctness on the basis of extensive numerical calculations.

Computational exploration has become a recognized tool in the study of dynamical systems. In particular, the restricted three-body problem (two heavy and one light body, attracted to one another by an inverse-square-of-the-distance force and moving in one plane) has been examined in this manner for over fifty years.[1] The results have been essentially qualitative and, in a certain sense, incomplete. There are too many different kinds of trajectories to fit comfortably into some scheme which catches them all. This happens apparently whenever the mathematical structure of a problem is so involved that it can be approached only through numerical work.

The anisotropic Kepler problem has never been investigated in this manner, to my knowledge. There is no evidence that the equations of motion can be separated, and the limit of isotropic masses provides only very poor information about the anisotropic case. A substantial effort has gone into understanding two features which make the ordinary two-body problem so simple. The first feature is the isomorphism between the geodesics (great circles) on a sphere and the trajectories in momentum space (hodograph). A similar isomorphism can be found; but the sphere becomes a Riemannian surface which can be immersed in Euclidean space as a "double snail," having obviously one badly singular point. The second feature is the behavior near a collision where the kinetic energy is much larger than the absolute value of the total energy. There seems to be no way to regularize the equations of motion by an appropriate choice of the variables, as it is possible to do in the ordinary Kepler problem. The best one can do is to reduce the equations to an autonomous (but not Hamiltonian) system in two dimensions, and gain insight into the trajectories when the total energy vanishes.

It comes, then, as a considerable surprise to find from numerical calculations that all trajectories can be described in a very simple, yet complete fashion. The main clues are the trajectories which intersect the "heavy" axis in position space perpendicularly. If one plots the further intersections of these trajectories with the "heavy" axis in a Poincaré map (conjugate momentum vs position), he finds a set of curves which can be used to define a natural coordinate system. Each trajectory is determined by two infinite sequences of binary numbers which give the signs of the position coordinate in the consecutive intersections with the "heavy" axis, for the forward and for the backward motion. If each binary sequence is interpreted as a real number (giving the natural coordinates), there is a one-to-one, continuous mapping into the initial conditions for the trajectory.

The existence of such a map makes the anisotropic Kepler problem in two dimensions an ideal example of a dynamical system. So far, only the geodesics in a space of negative curvature have been completely described in terms of Bernoulli schemes.[2] But, their behavior is quite different. The elements in the associated sequences are positive integers, rather than simply 0's and 1's as in the present case. Also, the trajectories are without conjugate points, whereas in the anisotropic Kepler problem neighboring trajectories with the same initial coordinates cut into one another. This reflects the mainly positive curvature of the "double snail" whose geodesics are being studied.

The Poincaré map is now identical with a double shift

of the binary sequences. Periodic orbits can be found quite systematically by discussing periodic sequences. Many of them are self-retracing in position space, a fact which follows directly from their binary sequence, but might not be easily understood otherwise. Their behavior as the anisotropy vanishes is of great interest because only the circular orbit in the ordinary Kepler problem survives as a periodic orbit if the masses become different. Finally, one can get some idea about the variety of trajectories which join any two given endpoints. The well-known baker transformation gives a picturesque demonstration of the problems involved, since one has to find the points of intersection between two closed curves; one of which is being gradually distorted by the iterated baker transformation. The discontinuities which arise are directly related to collisions.

This paper has been written in such a way that somebody can understand its content who is not familiar with the theory of dynamical systems. Also, many results are explained in rather descriptive terms, and a number of drawings are presented, because the conclusions are mostly based on the observation of extensive numerical work and not so much on mathematical deduction.

BACKGROUND

In a number of papers I have tried to widen the applicability of classical mechanics to the approximate solution of quantum mechanical problems.[3] The main emphasis has been on the phase integral approximation to find bound state energies in cases where the variables cannot be separated either in Schrödinger's equation or in the corresponding classical equations of motion. The relevant ideas are presented in the previous paper, and they are applied to a simple nontrivial example.

There might be simpler examples; but it seemed important to pick a physical situation which is sufficiently close to the Kepler problem. It was shown, in the first paper of this series, that the phase integral approximation gives perfect results for the bound states of the hydrogen atom, i.e., not only the energies, but also the wavefunctions for all bound states are given correctly. Introducing some spatially anisotropic feature appeared to be the most natural next step. This can be done with the help of an external electric or magnetic field, and the resulting situation is akin to the restricted three-body problem. Another, less well-known situation arises with a donor impurity in a semiconductor. In this case the potential energy remains isotropic (and Coulombic), but the kinetic energy becomes effectively anisotropic due to the electronic band structure in the solid. It is as if the mass of the electron in one direction is much larger than in the two other directions.

The starting point for the preceding investigations has been the classical approximation $\tilde{G}(q''q'\,E)$ for the quantum mechanical Green's function $G(q''q'\,E)$, which is the probability amplitude for an electron to reach the position q'' if it started as q' and has been moving with the energy E. If only the spectrum is required, but not the eigenstates, it is sufficient to consider the integral $\int d^3q\; G(qqE) = G(E)$ which has poles at the eigenvalues E_i of the energy.

The approximation $\tilde{G}(q''q'\,E)$ can be written as a sum over the classical trajectories from q' to q'' at the energy E, where each term consists of an amplitude and a phase factor. The former measures the density of those trajectories near q'' which started at q', and the latter is given by $\exp[(i/\hbar)S(q''q'\,E) - \frac{1}{2}\nu\pi]$, where $S(q''q'\,E)$ is the integral $\int p\,dq$ from q' to q'' along the particular trajectory and ν is the number of conjugate points between q' and q''.

If the energy E is negative, corresponding to bound states, there are always many classical trajectories between any two accessible points q' and q''. If the corresponding terms in $\tilde{G}(q''\,q'\,E)$ add up "in phase" for a particular value $\tilde{E}_i$ the approximate Green's function has a singularity as function of E, and an approximate eigenstate can be found with the energy $\tilde{E}_i$ as the approximate eigenvalue. If everything goes well, $\tilde{E}_i$ can be associated with a particular eigenvalue E_i. It is important to know how $\tilde{G}$ depends on q' and q'' for E near $\tilde{E}_i$.

In most practical cases, the behavior of the classical trajectories is quite complicated. Therefore, it is reasonable to compute $\tilde{G}(E) = \int d^3q\; \tilde{G}(qqE)$ and to get the approximate eigenvalues $\tilde{E}_i$ without worrying about the approximate eigenstates. This was done in the previous paper. By a very simple argument, it was shown that the integration of $\tilde{G}(qqE)$ over q emphasizes the periodic orbits in the summation over all classical trajectories. The trajectories which are closed but not periodic, i.e., where initial and final position coordinates coincide, but not the initial and final momenta, contribute only terms of higher order in Planck's quantums to $\tilde{G}(E)$.

Thus, the leading terms in $\tilde{G}(E)$ can be written as summation over all periodic orbits of energy E. Also, each term takes on a particularly simple appearance. The phase factor contains the phase integral $S(E) = \oint p\,dq$ over the periodic orbit and the number of conjugate points. The amplitude factor can be expressed in terms of the period T and the stability exponent. The condition for a resonance in $\tilde{G}(E)$ resembles the ordinary Bohr-Sommerfeld quantization condition. The integral $\oint p\,dq$ equals an integer times Planck's constant, and there are certain corrections connected with the number of conjugate points as well as the stability angle u, i.e., the imaginary part of the stability exponent. The real part v of the stability exponent has the effect of broadening the resonance. Its width relative to the separation between resonances is given by $v/2\pi$.

With all these results in mind, the main task in any particular case is to find the periodic orbits as a function of the energy. If they are sufficiently stable, i.e., if $v \ll 2\pi$, the quantization condition gives a series of resonances in $\tilde{G}(E)$ which qualify as approximate eigenvalues of the energy.

This idea was applied to the simplest periodic orbit of the anisotropic Kepler problem in the previous paper. The particular orbit was found with the help of a Fourier expansion very much like the one used by Hill in his classic work in the motion of the moon. Its stability exponent is small compared to 2π so that the quantization rules can be applied and a series of approximate energy $\tilde{E}_i$ found. These energies are associated with certain quantum numbers which are used in the description of impurity levels; but only a small fraction is approximated in this way.

The next step in this whole investigation is, therefore, quite obvious. A complete representation of all the periodic orbits in the anisotropic Kepler problem has to be found, and their contribution to the approximate

response function $\tilde{G}(E)$, particularly the resulting resonance structure, has to be established. The present paper goes a long way in this direction by describing the full variety of periodic orbits for the two-dimensional case. At the time of writing this is the only nontrivial case where such a complete description of the orbit structure is available with the exception of the geodesics in a space of negative curvature.

CHOICE OF COORDINATES

The present treatment of the anisotropic Kepler problem is entirely different from the preceding one. The nomenclature has been changed to adapt to the new way of looking at it. Before bringing quantum mechanics into the picture, there are three physical quantities to cope with: The charge e_0 of the electron, the energy unit E_0, and the masses of the electron, m_1 for the longitudinal and m_2 for the transverse mass out of which we get the mass unit $m_0 = (m_1 m_2)^{1/2}$.

With e_0, E_0, and m_0 we can get natural units for any other physical quantity, such as $(2m_0E_0)^{1/2}$ for the linear momentum, $e_0^2/2\kappa_0 E_0$ for the Cartesian coordinates where κ_0 is the dielectric constant of the medium (11.4 for Si and 15.36 for Ge), $(m_0 e_0^4/2\kappa_0^2 E_0)^{1/2}$ for the angular momentum. Everything will be expressed in these natural units including the time for which $(m_0 e_0^4/e\kappa_0^2 E_0^3)^{1/2}$ is the scale.

If x is the Cartesian coordinate in the longitudinal direction (large mass), y and z the Cartesian coordinates in the transverse plane (small mass), and u, v, w the conjugate momenta, the Hamiltonian is given by

$$\frac{\mu^2}{2\mu} + \frac{v^2}{2\nu} + \frac{w^2}{2\nu} - (x^2 + y^2 + z^2)^{-1/2}, \tag{1}$$

where $\mu = (m_1/m_0)^{1/2}$ and $\nu = (m_2/m_0)^{1/2}$ so that $\mu > \nu$ and $\mu\nu = 1$. The Hamiltonian has a constant value along any particular trajectory which is called $-\mathcal{E}/2$ so that the energy has the value $-\mathcal{E}E_0$ in ordinary units.

Instead of the Cartesian coordinates and the linear momenta, it is sometimes helpful to work with angular coordinates and angular momenta. For reasons which will become obvious later on, it seems advantageous to use angular coordinates in momentum space. Therefore, we write

$$\begin{aligned} u &= \sqrt{\mu}\, e^{\chi} \cos\vartheta, \\ v &= \sqrt{\nu}\, e^{\chi} \sin\vartheta \cos\varphi, \\ w &= \sqrt{\nu}\, e^{\chi} \sin\vartheta \sin\varphi, \end{aligned} \tag{2}$$

where $-\infty < \chi < \infty$, $0 \le \vartheta \le \pi$, $0 \le \varphi \le 2\pi$. The angular momenta L, M, and N are related to the Cartesian coordinates by

$$\begin{aligned} \sqrt{\mu}\, x &= (-L \sin\vartheta + N \cos\vartheta)\, e^{-\chi}, \\ \sqrt{\nu}\, y &= (L \cos\vartheta \cos\varphi - M(\sin\varphi/\sin\vartheta) + N \sin\vartheta \cos\varphi)\, e^{-\chi}, \\ \sqrt{\nu}\, z &= (L \cos\vartheta \sin\varphi + M(\cos\varphi/\sin\vartheta) + N \sin\vartheta \sin\varphi)\, e^{-\chi}. \end{aligned} \tag{3}$$

The canonical equations of motion are preserved because

$$x\, du + y\, dv + z\, dw = L\, d\vartheta + M\, d\varphi + N\, d\chi, \tag{4}$$

and the new Hamiltonian is given by

$$\tfrac{1}{2} e^{2\chi} - (e^{\chi}/R), \tag{5}$$

where for $R = r\, e^{\chi}$ and $r^2 = x^2 + y^2 + z^2$ one finds the expression

$$\begin{aligned} R^2 = L^2(\mu \cos^2\vartheta + \nu \sin^2\vartheta) + 2(\mu - \nu) \sin\vartheta \cos\vartheta\, LN \\ + N^2(\nu \cos^2\vartheta + \mu \sin^2\vartheta) + \mu(M^2/\sin^2\vartheta). \end{aligned} \tag{6}$$

It is worth noting that L, M, and N are easily expressed in terms of the Cartesian coordinates and the linear momenta, namely

$$\begin{aligned} L &= \nu u(yv + zw)/(v^2 + w^2)^{1/2} - \mu x\sqrt{v^2 + w^2}, \\ M &= vz - wy, \\ N &= ux + vy + wz. \end{aligned} \tag{7}$$

Since φ does not occur in the new Hamiltonian, the angular momentum M is a constant of motion. It is tempting to reduce the problem to one degree of freedom using χ as independent variable. In order to do so we shall use the abbreviations

$$\begin{aligned} e &= \mu \cos^2\vartheta + \nu \sin^2\vartheta, \\ f &= (\mu - \nu) \sin\vartheta \cos\vartheta, \\ g &= \nu \cos^2\vartheta + \mu \sin^2\vartheta. \end{aligned} \tag{8}$$

From the conservation of energy we get

$$R^2 = eL^2 + 2fLN + gN^2 + \mu M^2/\sin^2\vartheta = [2e^{\chi}/(\mathcal{E} + e^{2\chi})]^2. \tag{9}$$

N can now be considered as the reduced Hamiltonian which depends on the two conjugate variables ϑ and L, and on the independent variable χ. The equations of motion are

$$\frac{dL}{d\chi} = \frac{\partial N}{\partial \vartheta}, \quad \frac{d\vartheta}{d\chi} = -\frac{\partial N}{\partial L}. \tag{10}$$

The corresponding Langrangian Λ as a function of $\dot{\vartheta} = d\vartheta/d\chi$, ϑ, and χ results by eliminating L in the expression

$$\Lambda = L\frac{d\vartheta}{d\chi} + N \tag{11}$$

with the help of the second equation of motion. Thus, we obtain the expression

$$\Lambda = \left[\left(\frac{2e^{\chi}}{\mathcal{E} + e^{2\chi}}\right)^2 - \frac{\mu M^2}{\sin^2\vartheta}\right]^{1/2} \cdot (e - 2f\dot{\vartheta} + g\dot{\vartheta}^2)^{1/2} \operatorname{sgn}\left(\frac{N}{e - f\dot{\vartheta}}\right). \tag{12}$$

The last factor in Λ is really of no interest because it has no effect on the equations of motion. Obviously, $\Lambda d\chi$ looks like the element of length in a Riemannian space of coordinates χ and ϑ.

The Riemannian metric $\Lambda d\chi$ is quite different from the one which is ordinarily associated with a mechanical system. In the present case, there is a direct connection between momentum space and the coordinates χ and ϑ, not between position space and the Riemannian coordinates as usual. Whenever the bound states of a mechanical system are discussed, it seems much more informative to study momentum space provided the potential has a Coulomb type singularity. The trajectories in momentum space connect any given initial

momentum with any final momentum, whereas such a proposition does not, in general, hold for position space and negative energies. Consequently, it appears intuitively easier to find an immersion (if not an imbedding) of the above Riemannian space in a three-dimensional Euclidean space. Such an immersion will be constructed in the next section. It corresponds to the well-known stereographic projection of the momentum space in the ordinary Kepler problem onto the sphere. In this manner, one gains a more direct picture of the intricacies that come with the anisotropic Kepler problem.

MOMENTUM SPACE AS RIEMANNIAN SURFACE

The Riemannian space associated with momentum space can be obtained more directly by considering the virial along some trajectory, i.e., the integral over the expression (4) between some initial and final momenta. With the equations

$$\frac{du}{dt} = -\frac{x}{r^3}, \quad \frac{dv}{dt} = -\frac{y}{r^3}, \quad \frac{dw}{dt} = -\frac{z}{r^3}, \tag{13}$$

one finds that

$$-\int (x\,du + y\,dv + z\,dw) = \int r(du^2 + dv^2 + dw^2)^{1/2}, \tag{14}$$

because the vectors (x, y, z) and (du, dv, dw) are parallel. Since the Hamiltonian (1) has the constant value $-\mathcal{E}/2$, the radius r can be expressed in terms of u, v, and w. Thus, we find for the virial the expression

$$2\int [\mathcal{E} + (u^2/\mu) + (v^2/\nu) + (w^2/\nu)]^{-1}\,(du^2 + dv^2 + dw^2)^{1/2}. \tag{15}$$

It should be noted that for a closed orbit the virial is equal to the action integral $\int (u\,dx + v\,dy + w\,dz)$. By a straightforward calculation it follows that the equations for the geodesics in the Riemannian space with metric (15) are the same as the equations of motion which result from the Hamiltonian (1). The length of a geodesic equals the value of the virial between the corresponding endpoints in the anisotropic Kepler problem. For a closed geodesic the length equals the action integral around the corresponding periodic orbit.

If the polar coordinates (2) are used, the element of length becomes

$$\left(\frac{2e^{\chi}}{\mathcal{E} + e^{2\chi}}\right)^2 (e\,d\chi^2 - 2f\,d\chi\,d\vartheta + g\,d\vartheta^2 + \nu \sin^2\vartheta\,d\varphi^2) \tag{16}$$

with the abbreviations (8). For any subspace with $d\varphi = 0$, e.g., $z = 0$, this metric coincides with (12) provided $M = 0$. Since the present report is concerned with the two-dimensional anisotropic Kepler problem, all further calculations will be restricted to $d\varphi = 0$, or, equivalently, $z = 0, M = 0$.

Instead of the metric (16) in a plane with polar coordinates χ and ϑ, one can think of it as attached to a sphere in three-dimensional Euclidean space with coordinates ξ, η, ζ. The mapping from the plane onto the sphere is given by the formulas

$$\xi = \frac{2e^{\chi}}{\mathcal{E} + e^{2\chi}} \cos\vartheta, \quad \eta = \frac{2e^{\chi}}{\mathcal{E} + e^{2\chi}} \sin\vartheta, \quad \zeta = \frac{1}{\sqrt{\mathcal{E}}} \frac{e^{2\chi} - \mathcal{E}}{e^{2\chi} + \mathcal{E}} \tag{17}$$

The radius of the sphere is $1/\sqrt{\mathcal{E}}$, and its natural metric (induced by the Euclidean metric in ξ, η, ζ) is

$$[2e^{\chi}/\mathcal{E} + e^{2\chi})]^2\,(d\chi^2 + d\vartheta^2), \tag{18}$$

corresponding to (16) when $\mu = \nu = 1$.

Consider now two pencils of planes in the Euclidean space (ξ, η, ζ). The first consists of all planes through the straight line $\eta = 0$, $\zeta = 1/\sqrt{\mathcal{E}}$; the second consists of all planes through $\xi = 0$, $\zeta = 1/\sqrt{\mathcal{E}}$. Each plane in the first pencil can be characterized by the angle α it makes with the ζ axis, where $\alpha > 0$ if the plane intersects the positive η axis. Similarly, β measures the angle between a plane in the second pencil and the ζ axis, where $\beta > 0$ if the plane intersects the positive ξ axis. The equations for the first pencil are

$$\eta \cos\alpha + \zeta \sin\alpha = \sin\alpha/\sqrt{\mathcal{E}}, \quad -\tfrac{1}{2}\pi < \alpha < \tfrac{1}{2}\pi, \tag{19}$$

and for the second pencil one has

$$\xi \cos\beta + \zeta \sin\beta = \sin\beta/\sqrt{\mathcal{E}}, \quad -\tfrac{1}{2}\pi < \beta < \tfrac{1}{2}\pi. \tag{20}$$

The angles α and β can be used as coordinates on the sphere given by (17). The transformation from the coordinates χ and ϑ to the coordinates α and β follows from the formulas

$$\tan\alpha = (\mathcal{E})^{-1/2}\, e^{\chi} \sin\vartheta, \quad \tan\beta = (\mathcal{E})^{-1/2}\, e^{\chi} \cos\vartheta. \tag{21}$$

After some straightforward calculations the element of length (16) with $d\varphi = 0$ becomes

$$\frac{1}{\mathcal{E}} \left(\frac{2}{1 + \tan^2\alpha + \tan^2\beta}\right)^2 \left(\nu \frac{d\alpha^2}{\cos^4\alpha} + \mu \frac{d\beta^2}{\cos^4\beta}\right), \tag{22}$$

and the element (18) differs only by having effectively $\mu = \nu = 1$.

The lines of constant α or constant β on the sphere are the intersections between the sphere and the corresponding plane in one of the two pencils. If one measures the distance between two planes in the same pencil by integrating the element (22), the distance between two β planes comes out larger than the distance between the two α planes with the same values of the angles, because $\mu > \nu$. The idea is, therefore, simply to open up the angles between the planes in the β pencil.

Let us, therefore, rotate the points in one of the planes of the β pencil by the angle $(\gamma - \beta)$ around the axis of the pencil, i.e., the line $\xi = 0, \zeta = 1/\sqrt{\mathcal{E}}$. The value of γ as a function of β has to be determined later. The transformation of the points in the Euclidean space is given by the formulas

$$\begin{aligned} \xi' &= \xi \cos(\gamma - \beta) + ((\mathcal{E})^{-1/2} - \zeta) \sin(\gamma - \beta), \quad \eta' = \eta, \\ (\mathcal{E})^{-1/2} - \zeta' &= -\xi \sin(\gamma - \beta) + (\mathcal{E})^{-1/2} \cos(\gamma - \beta), \end{aligned} \tag{23}$$

provided ξ and ζ satisfy Eq. (20) of the β plane. We can combine (17) and (21) to describe the points which are simultaneously on the sphere of radius $1/\sqrt{\mathcal{E}}$ and the β plane,

$$\left(\xi, \eta, \frac{1}{\sqrt{\mathcal{E}}} - \zeta\right) = \frac{2/\sqrt{\mathcal{E}}}{1 + \tan^2\alpha + \tan^2\beta} (\tan\beta, \tan\alpha, 1). \tag{24}$$

The new surface in Euclidean space is given by

$$\left(\xi', \eta', \frac{1}{\sqrt{\mathcal{E}}} - \zeta'\right) = \frac{2/\sqrt{\mathcal{E}}}{1 + \tan^2\alpha + \tan^2\beta} \left(\frac{\sin\gamma}{\cos\beta}, \tan\alpha, \frac{\cos\gamma}{\cos\beta}\right). \tag{25}$$

Its natural metric is given by

$$d\xi'^2 + d\eta'^2 + d\zeta'^2 = \frac{1}{\mathcal{E}}\left(\frac{2}{1 + \tan^2\alpha + \tan^2\beta}\right)^2 \times \left\{\frac{d\alpha^2}{\cos^4\alpha} + \left[\cos^2\beta\left(\frac{d\gamma}{d\beta}\right)^2 + \sin^2\beta\right]\frac{d\beta^2}{\cos^4\beta}\right\}. \quad (26)$$

This agrees with (22) up to a factor ν if γ is chosen such that

$$\cos^2\beta\left(\frac{d\gamma}{d\beta}\right)^2 + \sin^2\beta = \frac{\mu}{\nu} = \frac{m_1}{m_2} = \mu^2. \quad (27)$$

The solution of this first-order equation for γ as function of β with $\gamma = 0$ for $\beta = 0$ gives the required angle γ for the transformation (23) of the sphere with radius $1/\sqrt{\mathcal{E}}$.

The differential equation (27) can be integrated without difficulty, and yields

$$\gamma = \int_0^\beta (\mu^2 - \sin^2\beta)^{1/2}\frac{d\beta}{\cos\beta} = \arcsin\left(\frac{\sin\beta}{\mu}\right) + (\mu^2 - 1)^{1/2}\log\left(\frac{(\mu^2 - \sin^2\beta)^{1/2} + (\mu^2 - 1)^{1/2}\sin\beta}{\mu\cos\beta}\right). \quad (28)$$

In spite of its appearance, the last term is antisymmetric in β so that $\gamma(-\beta) = -\gamma(\beta)$.

Each line of constant β on the surface (25) in three-dimensional Euclidean space is a circle of radius $\cos\beta/\sqrt{\mathcal{E}}$ which is tangent to the line $\xi = 0$, $\zeta = 1/\sqrt{\mathcal{E}}$. Its diameter in the (ξ', ζ') plane makes an angle γ with the negative ζ axis. If one plots the endpoints of these diameters as a function of γ in a polar diagram, he gets the crossections of the surface in the (ξ', ζ') plane. The result is a double snail. As β increases from 0 to $\pi/2$, the diameter goes to zero, but the angle γ goes to infinity logarithmically as is obvious from (28). The figure is symmetric with respect to the ζ axis, and the two halves cut into each other. Thus, only one half of the Riemannian space with metric (22) can be imbedded in a three-dimensional Euclidean space. When the two halves are glued together one has only an immersion. The cross section of the logarithmic double snail is plotted in Fig. 1 for $\mu^2 = 5$. Each half resembles the Nautilus shell of New Guinea (Nautilus pompilius) of which a photograph is presented in Fig. 2.

THE TRAJECTORIES FOR ZERO ENERGY

The behavior of the trajectories in the neighborhood of the origin is obviously important for the understanding of the particle motion in the anisotropic Kepler problem. It would, of course, be desirable to regularize the equations of motion in the event of a collision or near-collision; but it is doubtful whether such a procedure is possible. On the other hand, it is reasonable to assume that the behavior of the trajectories near collision is independent of the energy, because both kinetic and potential energy are large compared to the total energy. Thus, we can study the case of zero total energy as typical for all other cases provided the trajectory is near enough to the origin. In the case of the anisotropic Kepler problem with vanishing total energy, the equations of motion can be written as those of an autonomous system with one degree of freedom, i.e., they are equivalent with a (time-independent) vector field in two dimensions. Such a system can be discussed completely.

The discussion starts with Eqs. (10) where N is defined by (9) as a function of L, ϑ, and χ with $M = 0$. Introduce the quantities

$$\theta = \frac{\mathcal{E} + e^{2\chi}}{2e^\chi}L, \qquad \Psi = \frac{\mathcal{E} + e^{2\chi}}{2e^\chi}N, \quad (29)$$

so that one has the relation

$$e\theta^2 + 2f\theta\Psi + g\Psi^2 = 1. \quad (30)$$

The quantity Ψ depends only on θ and ϑ, not on χ. The equations of motion (10) can now be written as

$$\frac{d\theta}{d\chi} = \frac{\partial\Psi}{\partial\vartheta} + \frac{e^{2\chi} - \mathcal{E}}{e^{2\chi} + \mathcal{E}}\theta, \qquad \frac{d\vartheta}{d\chi} = -\frac{\partial\Psi}{\partial\theta}. \quad (31)$$

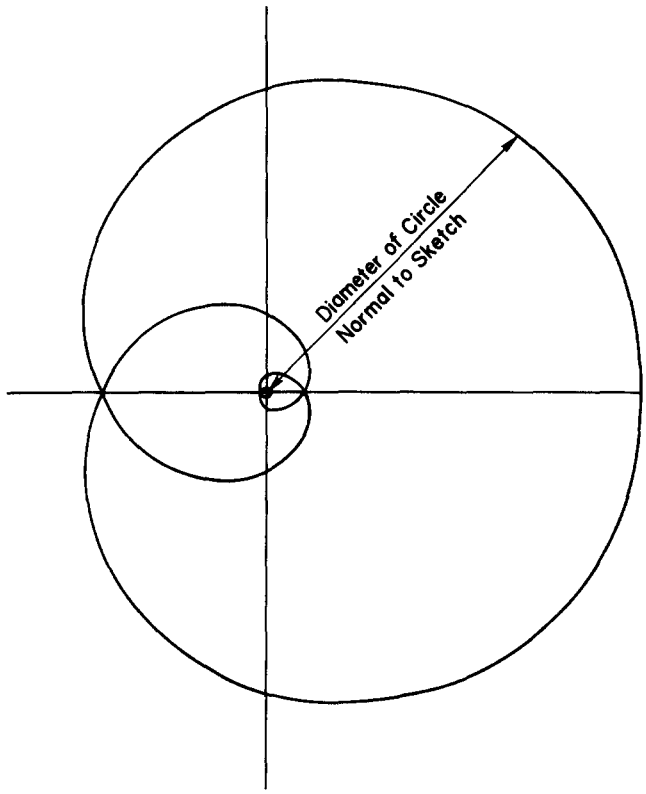

FIG. 1. Cross section through the "double snail" given by formula (28) with $\mu^2 = 5 = m_1/m_2$, i.e., silicon. Above each radius one has to draw a circle normal to the plane of the paper to get the two-dimensional surface in three-dimensional space.

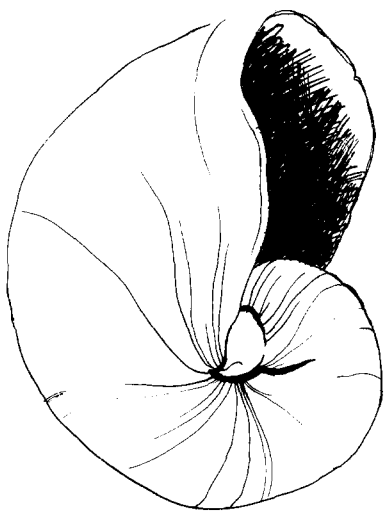

FIG. 2. Sketch of the Nautilus shell (Nautilus pompilius) from New Guinea. It can be thought as one half of the "double snail" whose cross section is shown in Fig. 1, except that the shell corresponds rather to a mass ratio $\mu^2 \cong 20 \cong m_1/m_2$, i.e., germanium.

Since the kinetic energy is given by $\frac{1}{2}e^{2\chi}$ according to (2), and the total energy by $-\mathcal{E}/2$, the equations can be directly expanded in powers of the ratio of the latter to the former. Retaining only the first term in such an expansion, the equations of motion become

$$\frac{d\theta}{d\chi} = \frac{\partial \Psi}{\partial \vartheta} + \theta, \quad \frac{d\vartheta}{d\chi} = -\frac{\partial \Psi}{\partial \theta}, \tag{32}$$

which is an autonomous system in two dimensions.

If Ψ is computed from (30), there is an ambiguity because one has to solve a quadratic equation. This is easily explained as follows. The angle ϑ determines the direction motion as can be seen from (2), whereas θ together with χ determines the angular momentum L (as always we assume $\varphi = 0, z = 0, w = 0$). But the two together do not indicate whether the particle is approaching the origin or going away from it. On the other hand, it follows from the equations of motion (13) and the conservation of energy that

$$\frac{d}{dt}(xu + yv + zw) = -\frac{\mathcal{E}}{2} + \frac{u^2}{2\mu} + \frac{v^2}{2\nu} + \frac{v^2}{2\nu}, \tag{33}$$

from which one usually derives the virial theorem. In this context one notices that $N = xu + yv + zw$ is always increasing with time if $\mathcal{E} = 0$. The sign of N, and hence of Ψ, tells whether the particle is coming in or going out.

With the help of (8) one can write the relation (30) as

$$\mu(\theta \cos\vartheta + \Psi \sin\vartheta)^2 + \nu(\theta \sin\vartheta + \Psi \cos\vartheta)^2 = 1. \tag{34}$$

For fixed ϑ, the values of θ and Ψ lie on an ellipse in the (θ, Ψ) plane with a semimajor axis of $\sqrt{\mu}$ along the direction $(-\sin\vartheta, \cos\vartheta)$ and a semiminor axis of $\sqrt{\nu}$ along $(\cos\vartheta, \sin\vartheta)$. As ϑ varies, this ellipse rotates. In a three-dimensional space of Cartesian axes with labels ϑ, θ, and Ψ, all these ellipses together form a surface which is topologically equivalent to a cylinder. However, since ϑ is an angular variable which is limited to 2π, the two ends of the cylinder at 0 and 2π have to be identified. A surface results which is topologically like a torus.

The differential equations (32) together with

$$\frac{d\Psi}{d\chi} = \frac{\partial \Psi}{\partial \theta}\frac{d\theta}{d\chi} + \frac{\partial \Psi}{\partial \vartheta}\frac{d\vartheta}{d\chi} = \theta \frac{\partial \Psi}{\partial \vartheta} \tag{35}$$

form a vector field on the torus, whereas Eqs. (32) alone are the projections of this vectorfield onto the (ϑ, θ) plane. They result from writing

$$\Psi = \frac{1}{g}[-f\theta + \lambda(g - \theta^2)^{1/2}], \tag{36}$$

where $\lambda = 1$ for the projection from above the (ϑ, θ) plane, and $\lambda = -1$ from below. The variable θ has to lie inside the strip which is defined by $\theta = \pm\sqrt{g(\vartheta)}$. The right-hand sides of (32) are finite inside, but become infinite at the boundary. The reason can be checked as follows. The independent variable χ goes through a maximum as the trajectory comes closest to the origin of the (x, y) plane; but neither the angle ϑ nor the angular momentum L, i.e., θ, has an extremum there. Thus, when θ goes to $\pm\sqrt{g}$, the vector field (32) becomes singular only because of the peculiar behavior of χ. However, in an autonomous system of differential equations the independent variable does not matter for the construction of the trajectories. As the projection of the trajectory from the torus onto the (ϑ, θ) plane reaches the boundary, one has to change the value of λ from $+1$ to -1 according as the trajectory passes from the upper to the lower part of the torus, or vice versa. As a simple rule of operation one can always use the fact that the angle ϑ changes monotonically and continuously while this happens.

The projection of the vector field from the upper part of the torus ($\lambda = 1$) is simply related to the projection from the lower part ($\lambda = -1$). Suppose the two components of the projection $\lambda = 1$ into the (ϑ, θ) plane are known for all points with $\theta^2 \leq g(\vartheta)$. The ϑ component for $\lambda = -1$ at (ϑ, θ) is the same as the ϑ component for $\lambda = 1$ at $(\vartheta, -\theta)$, whereas the θ component changes its sign. In other words, the projection $\lambda = -1$ results from the projection $\lambda = 1$ by reflecting both the points in the (ϑ, θ) plane and the vector components on each point on the ϑ-axis.

Finally, it is possible to work with only one projection, say $\lambda = 1$, and make the following rule in agreement with the above arguments: Whenever the trajectory reaches one of the boundaries $\theta = \pm\sqrt{g(\vartheta)}$, it jumps to the opposite boundary and proceeds with the opposite sign for θ, while keeping the value for ϑ.

With these technicalities out of the way, it is now relatively easy to get an explicit picture of the vector field (32) by studying its singularities. A straightforward calculation shows that they occur only at the points $\vartheta = 0, \frac{1}{2}\pi, \pi, \frac{3}{2}\pi$ with $\theta = 0$. Since e, f, and g have a period of π, the singularity at $\vartheta = 0$ is the same as the one at $\vartheta = \pi$, and the singularities at $\vartheta = \frac{1}{2}$ and at $\vartheta = \frac{3}{2}\pi$ are identical.

The linear part of Eqs. (32) near $\vartheta = 0, \theta = 0$ is given by

$$\begin{pmatrix} d\vartheta/d\chi \\ d\theta/d\chi \end{pmatrix} = \mu \begin{pmatrix} \mu - \nu & \sqrt{\mu} \\ -\sqrt{\mu}(\mu - \nu) & -\mu + 2\nu \end{pmatrix} \cdot \begin{pmatrix} \vartheta \\ \theta \end{pmatrix} \tag{37}$$

The eigenvalues of the matrix are

$$\Lambda = \tfrac{1}{2}[+1 \pm (9 - 8\mu^2)^{1/2}]. \tag{38}$$

In the typical case, such as $\mu^2 = 5$, one has complex values for Λ with the imaginary part larger than the real part. The resulting spiral is quite elongated with the long axis lying in the second and fourth quadrant of the (ϑ, θ) plane. The counterclockwise motion goes inward.

The linear part of Eqs. (32) near $\vartheta = \frac{1}{2}\pi, \theta = 0$ is given by

$$\begin{pmatrix} d\vartheta/d\chi \\ d\theta/d\chi \end{pmatrix} = \nu \begin{pmatrix} -\mu + \nu & \sqrt{\nu} \\ \sqrt{\nu}(\mu - \nu) & 2\mu - \nu \end{pmatrix} \cdot \begin{pmatrix} \vartheta - \frac{1}{2}\pi \\ \theta \end{pmatrix}, \tag{39}$$

where the eigenvalues of the matrix

$$\Lambda = \tfrac{1}{2} + 1 \pm (9 - 8\nu^2)^{1/2} \tag{40}$$

are now always real because $\nu^2 = \mu^{-2} \leq 1$. The resulting "saddle point" has one axis in the first and third quadrant almost parallel to the θ axis, while the other axis is in the second and fourth quadrant of the $(\vartheta - \pi/2, \theta)$ plane, almost parallel to the ϑ axis. The latter join the inward motion of the spirals on either side, while the former go to the boundaries $\theta = \pm\sqrt{g}$. A sketch of the resulting trajectories is shown in Fig. 3.

When $\mu^2 < \frac{9}{8}$ the spirals at $\vartheta = 0$ and π lose their twist,

and the inward trajectories become tangent to an axis in the first and third quadrant which becomes the ϑ axis as μ^2 approaches 1.

Since N always increases as time proceeds, Ψ goes through zero only once along any trajectory. The points where $\Psi = 0$ serve, therefore, as convenient starting points. They are given by

$$\theta = \lambda \operatorname{sgn} f\,[g/(1+f^2)]^{1/2} \tag{41}$$

according to (36). The projection of the vector field onto the (ϑ, θ) plane is never tangent to the line (41) except at $\vartheta = 0, \pi/2, \pi$, etc.

It is sufficient to investigate what happens for starting points in the domain $0 \leq \vartheta \leq \frac{1}{2}\pi$ and $\theta > 0$, i.e., $\lambda = 1$ (projection from above). If ϑ increases, one is led almost immediately to the boundary of the projection at $\theta = \sqrt{g}$, and from there into the projection $\lambda = -1$. The trajectories then go into the neighborhood of the saddle point, and end up in one of the spirals, at $\vartheta = 0$ or at $\vartheta = \pi$. If ϑ decreases, one always ends up at the spiral near $\vartheta = 0$. The only complication arises for small initial values of $\vartheta > 0$, because the trajectory hits the boundary of the projection $\lambda = 1$ for $\vartheta < 0$. Thus, one has to spend a little time in the projection $\lambda = -1$ before getting back into the projection $\lambda = 1$ and proceeding into the spiral near $\vartheta = 0$.

In view of formulas (2) the whole discussion in this section effectively describes the trajectory in momentum space. One has only to remember that χ goes to zero as the trajectory goes into the spiral, either in the forward direction ($\lambda = 1$) or the backward direction ($\lambda = -1$). The behavior of the trajectory in position space is more revealing, however, and it can be obtained with the help of formulas (3). If we combine (3) and (29) with $\mathcal{E} = 0$, we find that

$$\begin{aligned} \sqrt{\mu}\, x &= 2\,(-\theta \sin\vartheta + \Psi \cos\vartheta)\, e^{-2\chi}, \\ \sqrt{\nu}\, y &= 2\,(\theta \cos\vartheta + \Psi \sin\vartheta)\, e^{-2\chi}, \end{aligned} \tag{42}$$

where we can insert the results of integrating the linear differential equations (37) together with (36).

As the spirals at either end of the trajectory are approached, the absolute values of both x and y go to ∞, but the ratio y/x goes to zero. Thus, the trajectory in position space goes to infinity along the x axis by oscillating around it with increasing amplitude. According to the exact phase of this oscillation the trajectory will come in from one side of the x axis and go back out either on the same side or on the opposite. The critical phase obviously is the one where the trajectory hits the saddle point. Since the saddle point is at $\vartheta = \pi/2$ or $-\pi/2$, and since χ increases indefinitely as the saddle point is approached, the critical trajectory not only leads to collision, but the collision occurs only when the trajectory approaches the origin along the y axis.

This last discussion should make it very evident that the trajectories in the anisotropic Kepler problem are quite different from what they are in the usual isotropic situation.

TRAJECTORIES PERPENDICULAR TO THE X AXIS

The present section is entirely heuristic; but the results, if correct, are interesting enough to be discussed even on the basis of empirical rather than purely logical evidence. Also, it seems more convincing to present the evidence in the order in which it arose, with the interpretation given at the end. There are obviously many lemmas and proofs missing, and the final conclusion is not watertight.

The ultimate goal is the construction of a Poincaré map for the anisotropic Kepler problem with negative energy. To recapitulate the general idea, let q_1 and q_2 be the coordinates, p_1 and p_2 the conjugate momenta, and $H(p_1 p_2 q_1 q_2)$ the Hamiltonian. If the energy is fixed at E the initial conditions for any trajectory can be chosen as q_1 and p_1 with $q_2 = 0$ and $p_2 > 0$. The condition $H = E$ defines a domain D in the $(q_1 p_1)$ plane, each point of which defines a trajectory. This trajectory is followed forward in time until one finds again $q_2 = 0$ and $p_2 > 0$. Thus, the initial point in D is mapped into some other point of D. This map is one-to-one, continuous, and area-preserving.[4] The periodic orbits are simply fix points of some iterate of this Poincaré map. In addition, one hopes to find some relation between the values of q_1 and p_1 on one hand, and the shape of the trajectory on the other.

In the case of the anisotropic Kepler problem with negative energy, the choice of p_1 and q_1 is fairly obvious. In terms of the Hamiltonian (1) one chooses $p_1 = u, q_1 = x$, $p_2 = v > 0$, and $q_2 = y = 0$. The reason is the following lemma: Between any two crossings of the y axis, each trajectory crosses the x axis at least once.

Proof: Assume some trajectory which crosses the y axis twice without crossing the x axis. It is then possible to find a point on the trajectory between the two y axis crossings, such that the tangent to the trajectory at that point goes through the origin, and the curvature (and, therefore, the acceleration) is away from the x axis. However, since the force is directed toward the origin, and the anisotropic masses are such as to enhance the acceleration in the y direction, one has a contradiction. Thus, the trajectories cross the x axis more often than the y axis.

The domain D is defined by the inequality

$$\frac{u^2}{2\mu} - \frac{1}{|x|} = -\frac{\mathcal{E}}{2} - \frac{v^2}{2\nu} \leq -\frac{\mathcal{E}}{2}\,. \tag{43}$$

The shape of D in the (x, u) plane can be more easily visualized from the identical condition

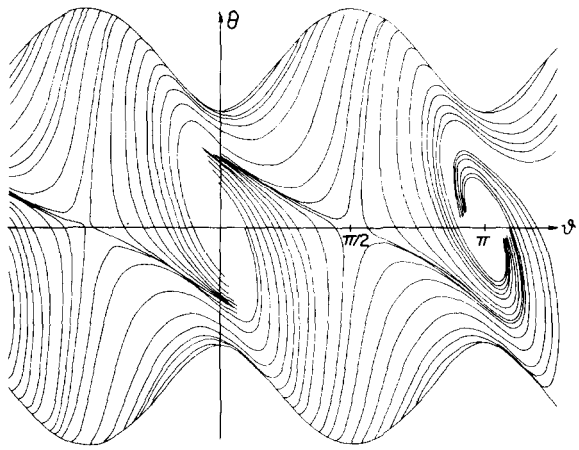

FIG. 3. Tangent curves for the vector field (31) with $\lambda = 1$. When one of the curves reaches a boundary, one has to continue on the opposite boundary with ϑ continuing monotonically.

$$|x| \leq \frac{2}{\mathcal{E} + u^2/\mu}, \tag{44}$$

where u stretches from $-\infty$ to $+\infty$. For plotting, it is better to have finite dimensions for D by mapping it into a (X, U) plane with the formulas

$$X = x[\mathcal{E} + (u^2/\mu)], \quad U = (\mu/\mathcal{E})^{1/2} \arctan[u/(\mu\mathcal{E})^{1/2}], \tag{45}$$

which preserves the area. (44) now becomes $|X| \leq 2$ with $|U| < (\mu/\mathcal{E})^{1/2}(\pi/2)$.

It is reasonable to ask first whether any orbit of the usual Kepler problem stays periodic even after introducing the anisotropic masses. According to Reeb and Moser, the perturbation part of the Hamiltonian has to be integrated around an ordinary Kepler orbit.[5] The particular orbits for which the value of this integral is stationary in the manifold of all orbits, remain periodic in the perturbed system. A simple calculation shows that only the circular orbit is of this kind. The resulting periodic orbit of the anisotropic Kepler problem is just the one which was discussed in the previous paper. It will be called the pseudo-circular orbit from now on. It looks roughly like an ellipse with the long axis in the y direction. It intersects both x and y axes perpendicularly. It was found for finite anisotropy exactly as Hill's variation orbit in the theory of moon, by Fourier expansion.

As a first step toward finding other periodic orbits, one can try to get at least the ones which intersect the x axis perpendicularly. In order to accomplish this, one chooses initial conditions with $u = 0, y = 0, v > 0$ and lets x vary between -2 and 2. (The energy $\mathcal{E}$ will be assumed henceforth to equal 1. The case $\mathcal{E} \neq 1$ can be reduced to $\mathcal{E} = 1$ by scaling the coordinates with $\mathcal{E}$ and the momenta with $1/\sqrt{\mathcal{E}}$.) If it is possible to find some initial value x such that at some later crossing of the x axis one has again $u = 0$, a periodic orbit results from mirroring the trajectory on the x axis. This is, of course, the reasoning which Hill applied originally to find a periodic variation orbit in the restricted three-body problem.

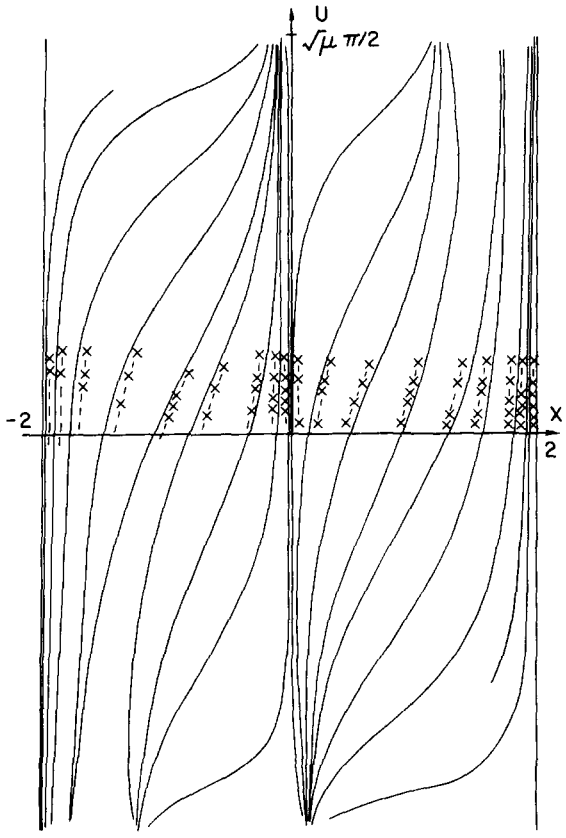

FIG. 4. Third iterate of the Poincaré map from the interval (0.25, 1.10) along the X axis into the (X, U) plane. Each curve is identified by the binary sequence corresponding to the consecutive intersections of its trajectory with the x axis. (Mass ratio $\mu^2 = 5 = m_1/m_2$).

All the following calculations were done by integrating the equations of motion in Cartesian coordinates with the help of a standard fourth-order Runge–Kutta method. The initial value of x (together with $u = 0$, $y = 0$, and $v > 0$) was changed in sufficiently small steps, starting with the value for the pseudo-circular orbit. The integration for each initial x was stopped as soon as the trajectory had crossed the x axis a certain number of times. The consecutive crossings can be numbered, with odd number when $v < 0$, and even numbers when $v > 0$. In this manner a sequence $(x_0, 0)$, $(x_1, u_1), (x_2, u_2), \cdots$ is obtained where the even numbered terms are just the consecutive iterates of the Poincaré map for $(x_0, 0)$.

As x_0 varies, the points (x_n, u_n) for fixed n in the (x, u) plane, or (X_n, U_n) in the (X, U) plane, run through a set of relatively smooth curves which will now be described in detail. Since the first task is to find initial values x_0 such that $u_n = 0$, one is looking for the intersections of these curves with the x_n axis. It is, therefore, a great relief to find the validity of the following.

Proposition 1: As x_0 increases, u_n always increases. Actually, u_n goes to $+\infty$, jumps to $-\infty$, and increases back up to $+\infty$, jumps again to $-\infty$, etc. According to (45), U_n increases to $\sqrt{\mu}(\pi/2)$, jumps to $-\sqrt{\mu}(\pi/2)$, and increases back up to $\sqrt{\mu}(\pi/2)$, jumps again to $-\sqrt{\mu}(\pi/2)$, etc. It is, therefore, quite easy to find as many values of x_0 where u_n vanishes, as there are curves in the (X, U) plane which go from $U = -\sqrt{\mu}(\pi/2)$ to $U = +\sqrt{\mu}(\pi/2)$.

These curves in the (X, U) plane do not intersect one another. Otherwise one would have some particular values X_n and U_n which lead to different values of x_0 upon integrating the corresponding trajectory backward in time, starting with $x_n, u_n, y = 0, v \gtrless 0$. Each curve stays entirely in one half-plane, either $X > 0$ or $X < 0$. It starts either with $0 < X < 1$ or with $-2 < X < -1$ at $U = -\sqrt{\mu}(\pi/2)$, and it ends either with $1 < X < 2$ or with $-1 < X < 0$ at $U = \sqrt{\mu}(\pi/2)$. X increases while U increases, over most of the domain D in the (X, U) plane. This last statement fails in the corners $(-2, -\sqrt{\mu}(\pi/2))$, $(-0, +\sqrt{\mu}(\pi/2))$, $(+0, -\sqrt{\mu}(\pi/2))$, $(2, \sqrt{\mu}(\pi/2))$, especially when μ is near 1. It is not needed for the further arguments, but it helps to visualize the curves. An example is given in Fig. 4 for $m_1 = 5m_2$, i.e., silicon.

To each curve in the (X, U) plane corresponds a (relatively short) interval on the X axis, i.e., the set of values x_0 which served as initial conditions. At first sight, it is absolutely not clear in which way the curves are ordered. Most emphatically, the curves in the (X, U) plane do not arise in the order of the intervals for x_0 out of which they come. There is, however, a surprisingly simple principle to explain the ordering of these curves and to relate it to the values of x_0. In order to understand this principle, one has to look more closely

at the shapes of the trajectories in the (x,y) plane as the initial coordinate x_0 varies. In particular, one has to examine what happens when the trajectory describes a collision of the particle with the origin.

Each trajectory which has been investigated so far $(x_0, u_0 = 0, y_0 = 0, v_0 > 0)$ can be described by a sequence of binary characteristics $(b_0 b_1 b_2 \cdots b_n)$ where $b_j = \mathrm{sgn}(x_j)$. The pseudo-circular orbit starting with $x_0 > 0$ has the alternating sequence $(+ - + \cdots (-1)^n)$. As x_0 varies there are certain critical values of x_0 where some of the characteristics change. At such a critical value of x_0, there is always a lowest index k such that b_k changes, but not b_j for $j < k$. The corresponding curve in the $(X_k\ U_k)$ plane reaches the boundary $U_k = \pm\sqrt{\mu}(\pi/2)$, but none of the U_j with $j < k$ do; on the other hand, all U_j with $j > k$ also reach the boundary $\pm\sqrt{\mu}(\pi/2)$. Obviously, the kth crossing of the x axis leads to a collision, but none of the earlier ones do. The question is: How do all the later crossings $(j > k)$ behave, as the kth crossing (but none of the earlier) sweeps through the collision with the origin? The answere is contained in the following.

Proposition 2: If the collision is swept over with increasing x_0, the characteristic b_k always goes from $-$ to $+$, and the characteristics b_j with $j > k$ simultaneously go from $+$ to $-$. In other words, as the kth crossing (but no earlier one) reaches a collision with x_0 increasing, it does so by having x_k approach 0 from below. Also, the trajectory just before the kth crossing is nearly perpendicular to the x axis and stays that way as x_0 goes beyond the critical value. Meanwhile, all the later crossings occur with $x_j > 0$ for x_0 below the critical value, and switch simultaneously to $x_j < 0$ above. No exception to this proposition has been found in extensive computations.

Before presenting the mathematical (rather than numerical) arguments in favor of the last proposition, let us look at the immediate consequences. It should be remarked that as x_0 approaches the critical value from below, all U_j with $j \geq k$ approach $\sqrt{\mu}(\pi/2)$ from below; and as x_0 approaches the critical value from above, all U_j with $j \geq k$ approach $-\sqrt{\mu}(\pi/2)$ from above. Therefore, each continuous piece of curve in the (X_n, U_n) plane has a unique sequence of binary characteristics, particularly its intersection with the X_n axis which gives rise to a periodic orbit.

Consider now for example the case $n = 6$, i.e., the third iterate of the Poincaré map. Start with the pseudo-circular orbit, i.e., the binary sequence $(+ - + - + - +)$. As x_0 increases, the first time a collision occurs is at the fifth crossing. The new sequence is $(+ - + - + + -)$. The next collision happens at the sixth crossing, giving the new sequence $(+ - + - + + +)$. The next collision comes at the third crossing, yielding the new sequence $(+ - + + - -)$. And so forth. Each new sequence gives a trajectory for which $U_6 = 0$. If we associate a binary number with each sequence by writing

$$B_n = \sum_0^n b_j (\tfrac{1}{2})^j, \tag{46}$$

B_n is obviously a monotonically increasing function of x_0. All 2^{n+1} binary rationals B_n from -2 to $+2$ actually occur, each exactly once as x_0 varies from -2 to $+2$. Thus, we find exactly as many periodic orbits which are symmetric with respect to the x axis, for each value of $n \geq 1$.

Binary sequences of identical length and their periodic orbits have been ordered so far. But it is not clear how the initial value x_{on} for a periodic orbit with binary rational B_n, and another periodic orbit with binary rational B_m and initial x_{om}, are ordered with respect to each other if $n < m$. The rule is quite simple: Since the nth crossing of the first and the mth crossing of the second orbit have zero momentum in the x direction, both orbits and their associated binary sequences can be continued without further calculation. If l is the least common multiple of n and m, the lth crossing for both has again a vanishing momentum in the x direction. Therefore, the two orbits are now described by a binary rational of identical length which tells us immediately whether x_{on} is smaller than x_{om} or vice versa.

A trajectory which starts with $u_0 = 0$ and has a later crossing, say the nth, with $u_n = 0$, should really be assigned an infinite binary number by extending the binary sequence $(b_0 \cdots b_n)$ beyond n. The procedure is simply to define $b_{n+i} = b_{n-i}$ which extends the sequence to b_{2n}. Then one defines $b_{2n+j} = b_{2n-j}$ which extends the sequence to b_{4n} and so forth. Thus, one gets the binary number

$$B = \sum_0^\infty b_j (\tfrac{1}{2})^j \tag{47}$$

to be associated with a trajectory which cuts the x axis perpendicularly at least twice. B is a monotonically increasing function of the initial value x_0 of those orbits.

Since B was obtained by expanding a finite binary sequence as explained in the previous paragraph, not all real numbers between -2 and $+2$ can be obtained in such a manner. However, the special numbers B form a dense set in the interval $(-2, 2)$. Therefore, the map from x_0 to B can be defined for all values of x_0, and remains, of course, monotonically increasing.

With the help of this map one can describe what happens as the anisotropy of the masses changes with the limited class of periodic orbits which intersect the x axis perpendicularly. The initial value x_{00} for the pseudo-circular orbit is always mapped into $B = \pm\frac{2}{3}$. In the limit of isotropic masses this particular value x_{00} goes to ± 1, whereas for $m_1 = 5m_2$ one has $x_{00} = \pm .49754$. As x_0 moves away from x_{00}, B moves away from $\pm\frac{2}{3}$. But the rate at which B moves with respect to x_0 increases very strongly as the anisotropy increases. In fact, this rate drops to zero when the anisotropy vanishes because in the ordinary Kepler problem all trajectories are periodic and have an alternating binary sequence. It is as if the periodic orbits with $B \neq \pm\frac{2}{3}$ are pushed out of the x axis when the anistropy decreases. Their initial values x_0 are crowded more and more toward the ends of the intervals $(-2, 0)$ and $(0, 2)$.

All the various statements in this section should be corroborated by mathematical deduction from the original equations of motion. But very little progress has been made along this line. Only some qualitative arguments can be made on the basis of what was shown in the previous paper and the preceding sections.

One obvious remark has to do with the stability of the pseudo-circular orbit. It was shown in the previous paper to be unstable and to have two conjugate points. From this one can conclude at once that as x_0 increases beyond x_{00}, any one intersection x_n also increases. For even n and $x_{00} > 0$, therefore, x_n moves to the right of

x_{00} and x_0, whereas for odd n it moves toward the origin. Thus, the odd intersections x_n lead to a collision before the even x_{n+1} does. But that explains only the very first change in a binary sequence starting with the alternating sequence.

The other remark refers to the preceding section, which was written mostly to provide some insight into the collision process. If one starts with a trajectory ($\mathcal{E} = 0$) and modifies it so as to make it go through one of the saddle points, firstly the approach always goes along the y axis, and secondly, the sign of all intersections with the x axis after the collision changes simultaneously. Therefore, the case $\mathcal{E} = 0$ is consistent with the cases $\mathcal{E} > 0$. But, that still does not explain all the other relatively simple features which were observed when $\mathcal{E} > 0$ and which allow such a detailed and, to a certain extent, exhaustive description of the trajectories.

NATURAL COORDINATES FOR THE POINCARÉ MAP

All those periodic orbits which cut the x axis twice with vanishing momentum in the x direction, were effectively enumerated in the preceding section. Also, the enumerating scheme is such as to determine the value x_0 of the x coordinate where the periodic orbit intersects the x axis perpendicularly. The immediate problem now is to generalize this method for the anisotropic Kepler problem so as to cover all trajectories. Or in other words, is it possible to extend the mapping from x_0 to B which was discussed in the preceding section from the x axis to the whole domain D of the (X, U) plane? The affirmative answer will be presented in this section.

The curves in the (X, U) plane which were discussed in the preceding section, and sketched in Fig. 4, give the $\frac{1}{2}n$th iteration of the Poincaré map from the X axis into the (X, U) plane. Because of the symmetry with respect to time of the equations of motion, one can also consider the X axis as the image of the curves in the (X, U) plane under the $\frac{1}{2}n$-times iterated Poincaré map. More precisely, if a trajectory starting with the initial conditions $x = x_0, u = 0, y = 0$, and $v > 0$ leads to the sequence

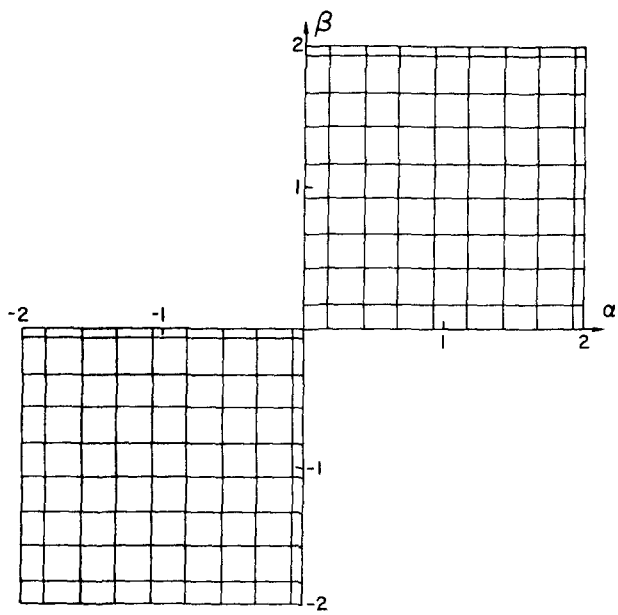

FIG. 5. Domain of the two real numbers α and β given by (48) which characterize the binary sequence $(\cdots a_{-1}a_0a_1\cdots)$ of any trajectory.

$(x_0, 0), (x_1, u_1), \ldots, (x_n, u_n)$ of intersections with the x axis, then a trajectory with the initial conditions $x = x_n$, $u = -u_n, y = 0$, and $v > 0$, leads to the sequence $(x_n, -u_n)$, $(x_{n-1}, -u_{n-1}), \ldots, (x_1, -u_1), (x_0, 0)$ of intersections with the x axis.

Each trajectory with initial conditions $x = x_0, u = u_0$, $y = 0$, and $v > 0$ can be assigned an infinite sequence of crossings with the x axis $\cdots (x_{-1}, u_{-1}), (x_0, u_0), (x_1, u_1)$, $(x_2, u_2)\cdots$ where the odd indices correspond to crossings with $v < 0$ and the even ones to those with $v > 0$. Also, positive indices correspond to crossing for $t > 0$, while negative indices correspond to $t < 0$. Again, there is an infinite sequence of binary characteristics $\cdots$, $a_{-1}, a_0, a_1, a_2, \cdots$, where $a_j = \mathrm{sgn}(x_j)$. Two real numbers, α and β, can be defined by

$$\alpha = \sum_0^\infty a_n(\tfrac{1}{2})^n, \qquad \beta = \sum_0^\infty a_{-n}(\tfrac{1}{2})^n, \tag{48}$$

whose domain is given by the union of the two squares $(0 < \alpha < 2, 0 < \beta < 2)$ and $(-2 < \alpha < 0, -2 < \beta < 0)$ shown in Fig. 5.

A trajectory with binary characteristics $(b_0\, b_1 \cdots b_n)$ in the preceding section can now be viewed as a trajectory starting with $x = x_n, u = u_n, y = 0$, and $v > 0$, whose binary characteristics are known backward in time, $a_{-j} = b_{n-j}$ for $0 \le j \le n$. Actually, the a's are known even further back because one knows that $u_{-n} = 0$ so that $a_{-n-i} = a_{-n+i}$ for $0 \le i \le n$. Therefore, the curves in the (X, U) plane which were discussed in the preceding section and are sketched in Fig. 4, correspond to trajectories for which β is defined within the rather narrow bounds $\pm(\frac{1}{2})^{2n}$. Viewed from this angle one notices a very striking feature which is described in the following:

Proposition 3: The domain D in the (X, U) plane is covered with a set of nonintersecting curves, each going from the lower boundary $U = -\sqrt{\mu(\pi/2)}$ to the upper boundary $U = +\sqrt{\mu(\pi/2)}$, so that they can be ordered with respect to their intersection with $U = 0$. Each of these curves also has a binary characteristic $(a_0\, a_{-1} \cdots a_{-n})$, and each point on these curves has a real number β associated with it where β is known only to the precision $\pm(\frac{1}{2})^{2n}$. The ordering with respect to β is the same as the ordering with respect to the intersection with $U = 0$. Again, no exception to this proposition has been found in extensive computations.

The initial conditions are, therefore, mapped into the real numbers β. The lines of constant β are given in the preceding section as the image of the X axis under the $\frac{1}{2}n$th iterate of the Poincaré map. The map from the initial conditions into β is essentially continuous (forgetting complications at the boundaries of D).

Because of the symmetry with respect to the x axis we can associate with a trajectory of binary characteristic $(b_0\, b_1 \cdots b_n)$ in the preceding section, a trajectory whose initial conditions are $x = x_n, u = -u_n, y = 0, v > 0$. Such a trajectory is characterized by $a_j = b_{n-j}$ for $0 \le j \le n$ and $a_{n+i} = b_i$ for $0 \le i \le n$. Its associated real number α is again known with a precision $\pm(\frac{1}{2})^{2n}$. Curves of constant α are the same as those of constant β after reflection on the X axis. Therefore, the above proposition holds just as well for α as for β.

In conclusion, the real numbers α and β provide a coordinate system for the domain D in the manner shown in

Fig. 6 for $m_1 = 5m_2$. The mapping from D into the two squares of Fig. 5 can be understood as follows. The X axis of D goes into the diagonal $\alpha = \beta$ because, if $u = 0$ initially, the sequence of intersections of the trajectory with the x axis is the same going forward and going backward in time. The upper and lower boundaries of D are mapped into various sides of the two squares, e.g., $(0 < X < 1, U = \sqrt{\mu}(\pi/2))$ into $(\alpha = 0, 0 < \beta < 2)$, $(1 < X < 2, U = \sqrt{\mu}(\pi/2))$ into $(0 < \alpha < 2, \beta = 2)$, $(2 > X > 1, U = -\sqrt{\mu}(\pi/2))$ into $(\alpha = 2, 2 > \beta > 0)$, etc. These boundary points correspond to trajectories which just had a collision or are just getting out of it, i.e., where either in the forward or in the backward direction all intersections with the x axis have the same sign, while in the opposite direction the trajectory behaves quite smoothly and regularly. The vertical boundaries of D, i.e., $X = -2, X = 0$, and $X = 2$, are mapped into the points $\alpha = \beta = -2, \alpha = \beta = 0$, and $\alpha = \beta = 2$.

The map from D to $\{\alpha, \beta\}$ has been constructed numerically, and the procedure demands at this time the explicit integration of the equations of motion in order to obtain a grid of approximate curves $\alpha = \text{const}$, or $\beta = \text{const}$. It is interesting to note that this works better for large anisotropy which leads to a grid of relatively even mesh size for short trajectories, i.e., small n, whereas for small anisotropy the curves of the same constant value of α or β are driven toward the boundaries of D, leaving the interior only poorly covered.

BERNOULLI SCHEMES AND PERIODIC ORBITS

Each trajectory in the anisotropic Kepler problem yields a binary sequence $(\cdots a_{-1} a_0 a_1 \cdots)$, and conversely each binary sequence gives rise to two trajectories, as has been shown in the previous section. The ambivalence comes from the symmetry with respect to the x axis, unless it is specified that the zero-crossing, a_0, has a positive momentum in the y direction, $v > 0$.

Binary sequences form a particularly simple example of a dynamical system whose basic ingredient is the tossing of a coin. The binary characteristic a_j indicates whether the result of the jth toss was "head" or "tail." Each sequence can be considered as an event

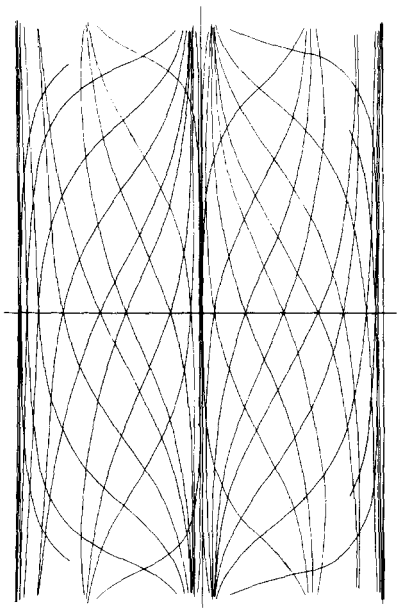

FIG. 6. Approximate coordinate grid in the (X, U) plane. Each curve corresponds to $\alpha = \text{const}$ (curves from upper left to lower right) or $\beta = \text{const}$ (curves from lower left to upper right). The curves are directly taken from Fig. 4.

or a point in a space, and one can define an algebra of measurable sets in this space. E.g., the union of the two squares in formula (48) and in Fig. 5 gives a picture of this space. There is a natural automorphism which maps every sequence $(\cdots a_{-1} a_0 a_1 \cdots)$ into another one $(\cdots a'_{-1} a'_0 a'_1 \cdots)$ through the formula

$$a'_j = a_{j+1}, \tag{49}$$

called a shift. The whole thing is called a Bernoulli scheme.[6]

The main result of this paper can now be stated as a theorem (if one is willing to accept the evidence of the two preceding sections) or, otherwise, as a

Conjecture: There is a one-to-one, continuous mapping between the anisotropic Kepler problem and the binary Bernoulli scheme, such that the Poincaré map for the (x, u) plane is equivalent to a double shift of the binary sequences.

The measure which is ordinarily used in the discussion of Bernoulli schemes, and which corresponds to the probability of success in tossing a coin, is essentially the area in Fig. 5. A comparison of Figs. 5 and 6 makes it doubtful whether corresponding meshes have equal areas. Therefore, it looks as if the Poincaré map for the anisotropic Kepler problem yields a measure for the Bernoulli scheme different from the usual one. As a matter of fact, it appears that each value of the ratio m_1/m_2 gives a different measure, which is, however, conserved in the shift.

As an application of the isomorphism with the Bernoulli schemes, the set of periodic orbits in the anisotropic Kepler problem will be discussed in this section. So far, only those periodic orbits were found which intersect the x axis perpendicularly in two places. Their binary sequences have an even period $2n$ with the additional symmetry $a_{n+j} = a_{n-j}$ for $0 \le j \le n$. The sequence is then symmetric with respect to 0, i.e., $a_{-i} = a_i$ for $i \ge 0$, so that $\alpha = \beta$ correspond to $u_0 = 0$. Similarly, $a_{n+j} = a_{n-j}$ for $j \ge 0$ so that $u_n = 0$. With the help of the binary sequences it is now possible to find the answer to the following kind of question: Are there any periodic orbits which intersect the x axis perpendicularly in only one place? Are there any periodic orbits which never cut the x axis perpendicularly?

Clearly, a periodic binary sequence gives rise to a periodic orbit, and vice versa. It is of some interest to start enumerating the periodic orbits in the order of the length n of their period in the associated binary sequence. If $n = 1$, one has either $\alpha = \beta = 2$ or $\alpha = \beta = -2$, and in both cases the corresponding point is not in the interior of the domain D in the (X, U) plane. If $n = 2$, one gets only the alternating sequence and, therefore, the pseudo-circular orbit.

The first novel case comes with $n = 3$ where the only possibility is $(\cdots + - - + - - + - - \cdots)$ apart from shifts and overall change in sign. With $a_{3l} = +1$ and $a_{3l+1} = a_{3l+2} = -1$, where l is any integer, one finds $\alpha_{3l} = \beta_{3l} = \frac{2}{7}$; $\alpha_{3l+1} = -\frac{10}{7}$ and $\beta_{3l+1} = -\frac{6}{7}$; and $\alpha_{3l+2} = -\frac{6}{7}$ and $\beta_{3l+2} = -\frac{10}{7}$. One is forced to conclude the existence of a periodic orbit which starts with $u_0 = 0$ and $v_0 > 0$ at some well-determined initial position $x_0, y_0 = 0$; and which, after two intersections with the x axis, returns to that same position again with $u = 0$, but this time with $v < 0$. Also, since the α and β

values for the two intermediate intersections with the x axis are symmetric with respect to each other, the corresponding values of the x coordinate are identical, while the corresponding values of the u momentum are opposite. All this is possible only if the particle retraces its initial trajectory to the opposite direction. And that, in turn, requires the trajectory to go to a point where its kinetic energy vanishes, i.e., a point on the boundary $x^2 + y^2 = 4$.

With all these indications, it is not hard to find the exact value of x_0 using the coordinate grid of Fig. 6; then do the integration numerically to check whether there is, indeed, a periodic orbit which retraces itself after three intersections with the x axis. The result is shown

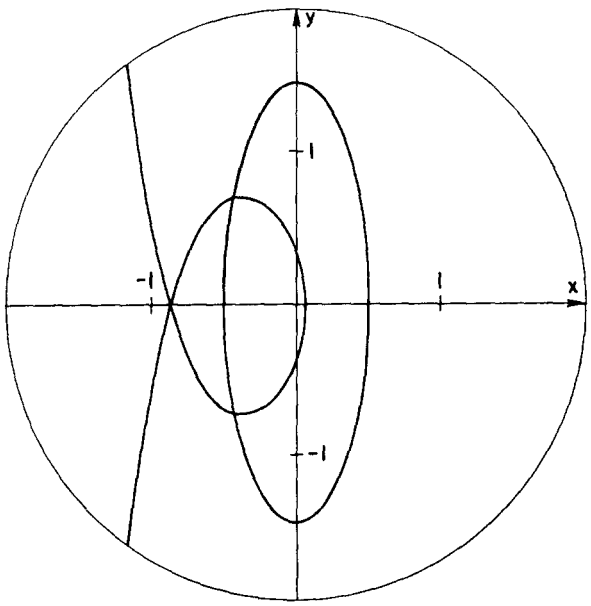

FIG. 7. The periodic orbits in the (x,y) plane corresponding to the periodic sequences $(\cdots + - + - + - \cdots)$ and $(\cdots + - - + - - \cdots)$. The former is the pseudo-circular orbit, and the latter is the first self-retracing orbit which goes up to the limiting circle $x^2 + y^2 = 4$. $(\mu^2 = 5)$.

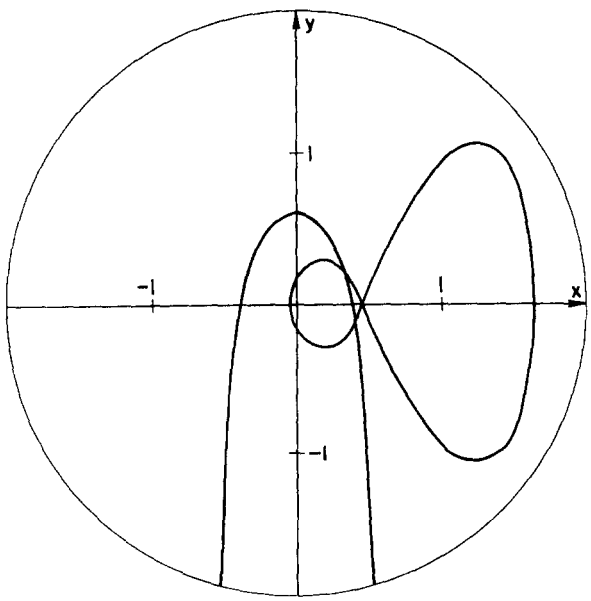

FIG. 8. Periodic orbits in the (x,y) plane which correspond to the periodic sequences $(\cdots + + + - + + + - \cdots)$ and $(\cdots + + - - + + - - \cdots)$. The former intersects the x axis perpendicularly in two different places, and the latter intersects only the y axis perpendicularly.

in Fig. 7. Obviously, this is the prototype of a large class of periodic orbits with an odd period n, such that $a_j = a_{n-j}$ for $0 \leq j \leq n$.

The existence of these orbits could hardly be understood if there was not the isomorphism with the Bernoulli schemes. Without this isomorphism one would rather believe that an orbit can be found which starts at $y = 0$ with $u = 0$ and $v > 0$, goes into the negative half-plane $x < 0$, cuts the negative x axis at two different places, comes back to the positive half-plane $x > 0$, and finally intersects the positive x axis perpendicularly, but at a place different from the starting point. There seems to be more freedom to adjust to the initial and final condition $u = 0$ when intersecting the positive x axis. But the only orbit of this kind turns out to have the same initial and final point on the x axis.

When $n = 4$, there are two essentially different cases, the sequence $(\cdots + + + - + + + - + + + - \cdots)$ and the sequence $(\cdots + + - - + + - - + + - - \cdots)$. The first gives a periodic orbit which intersects the x axis perpendicularly at two different places, once at $x > 0$ and once at $x < 0$. The second sequence is of interest because its associated orbit intersects the y axis perpendicularly, but not the x axis. Moreover, the two intersections with the positive x axis have symmetric values for α and β, as do the two intersections with the negative x axis. Therefore, one has again a self-retracing orbit which goes all the way to the boundary $x^2 + y^2 = 4$. Its picture is given in Fig. 8.

For a period $n = 5$, one finds three different self-retracing periodic orbits all of which intersect the x axis perpendicularly. For a period $n = 6$, there appears a new orbit of special interest because it intersects neither the x axis nor the y axis perpendicularly. Its sequence is $(\cdots + + + + - - + + + + - - \cdots)$. It is self-tracing and seems to float rather freely in the (x,y) plane although it hits the boundary $x^2 + y^2 = 4$ at two different points, as shown in Fig. 9. Again, one is struck by the special character of this orbit, and the difficulty of finding it without the Bernoulli scheme.

The binary sequences provide a natural method of enumerating all periodic orbits of the anisotropic Kepler problem and for ordering them according to their complication. It is satisfying to find the pseudo-circular orbit to be the first one in this scheme. This is not the place to indulge in an exhaustive study of all the various types which arise, as well as the number of their conjugate points and their stability. But it should be kept in mind that if the periodic orbits are important for the quasiclassical response function $\tilde{G}(E)$, a more complete investigation of their behavior is necessary.

THE BAKER TRANSFORMATION

It was explained in the first section how the classical approximation $\tilde{G}(q''q'\ E)$ for the quantum mechanical Green's function $G(q''q'\ E)$ is obtained. Among other things, one has to find all the classical trajectories which start at the position q' and end at the position q'' while moving with the fixed energy E. The problem of enumerating all these trajectories is, of course, even more difficult than finding all the periodic orbits. However, it will be shown how the binary sequences give at least a qualitative idea of the solution.

Consider a point (x,y) in the first quadrant, i.e., $x > 0$ and $y > 0$, but inside the limiting circle, i.e., $x^2 + y^2 < 4$.

All the other cases are either limits of this one, or can be obtained by reflection on the x axis and/or y axis. Any trajectory through the point (x,y) has a binary sequence associated with it, and there are two real numbers to describe this binary sequence. For the purpose of this section, these two real numbers will be defined slightly differently.

Let the binary sequence be $(\cdots a_{-1}a_0a_1\cdots)$ as before, where a_0 gives the sign of the first intersection with the x axis in the forward direction, after the point (x,y) has been traversed. Since $y>0$, the v momentum at this intersection is negative. Similarly, a_{-1} gives the sign of the last intersection with the x axis before the point (x,y) is reached. The corresponding v momentum is positive. The binary sequence is completely fixed (except where the particle collides with the origin) by the real numbers

$$\xi=\sum_0^\infty a_n(\tfrac{1}{2})^{n+1},\qquad \eta=\sum_1^\infty a_{-n}(\tfrac{1}{2})^n \tag{50}$$

whose values are in the square $-1<\xi<1, -1<\eta<1$. In terms of the previous notation one has $\alpha=2\xi$ and $\beta=a_0+\eta$. The advantage of the present notation is that it clearly distinguishes between the forward and the backward half of the trajectory in a symmetric fashion. As long as $x=0$, such a distinction cannot be made in a natural way.

As the trajectory through the point (x,y) varies its initial direction, the numbers ξ and η vary continuously. The point (ξ,η) traces a continuous curve in the (ξ,η) plane. Several examples are given in Fig. 10. These curves are symmetric with respect to the diagonal $\xi=\eta$, and intersect themselves on the diagonal, exactly once. This last fact is of interest, because it shows that there is exactly one trajectory through (x,y) which looks the same in the forward and in the backward direction. Again this trajectory runs directly from the point (x,y) to the limiting circle from whence it retraces itself. Notice that the point of return lies in the same half-plane $y>0$ as (x,y) itself. Thus, one gets the somewhat unexpected

Proposition 4: Through each point (x,y) of one quadrant there is exactly one trajectory which runs directly to the limiting circle in the same half-plane with respect to x. These special trajectories do not intersect one another in that half-plane.

Consider now two different points (x_1,y_1) and (x_2,y_2), and let them belong to the same first quadrant for simplicity's sake. Each has its curve in the (ξ,η) plane. Wherever these two curves intersect, one has a trajectory which connects them. For instance, if the two points lie on the same trajectory to the limiting circle, their curves in the (ξ,η) plane will have the same point on the diagonal $\xi=\eta$. The trajectories between (x_1,y_1) and (x_2,y_2) which are obtained in this manner, do not intersect the x axis in the interval between (x_1,y_1) and (x_2,y_2). They are, obviously, the simplest that exist. How does one get the others?

Let now (ξ,η) be a trajectory which starts at (x_1,y_1) and goes through (x_2,y_2) after having intersected the x axis twice. (Clearly, since both $y_1>0$ and $y_2>0$, one always needs an even number of intersections with the x axis to get back to the same half-plane.) Viewed from (x_1,y_1) its binary sequence would be $(\cdots a_{-1}a_0a_1\cdots)$, and viewed from (x_2,y_2) its binary sequence would be $(\cdots a''_{-1}a''_0a''_1\cdots)$ where $a''_n=a_{n+2}$ for all integer n. It is not the curve in the (ξ,η) plane which characterizes the point (x_1,y_1), but rather what becomes of it after two shifts (49), that has to intersect with the curve which characterizes (x_2,y_2). Thus, one is led to ask: What becomes of any curve in the (ξ,η) plane if the corresponding binary sequences undergo a shift?

If (ξ',η') are the two real numbers which are associated with the binary sequence after one shift (49), one finds immediately from (50) the formulas

$$\xi'=2\xi-\operatorname{sgn}\xi,\qquad \eta'=\tfrac{1}{2}(\eta+\operatorname{sgn}\xi). \tag{51}$$

That is the transformation of the square $-1<\xi<+1$, $-1<\eta<+1$ into itself which a baker uses when he

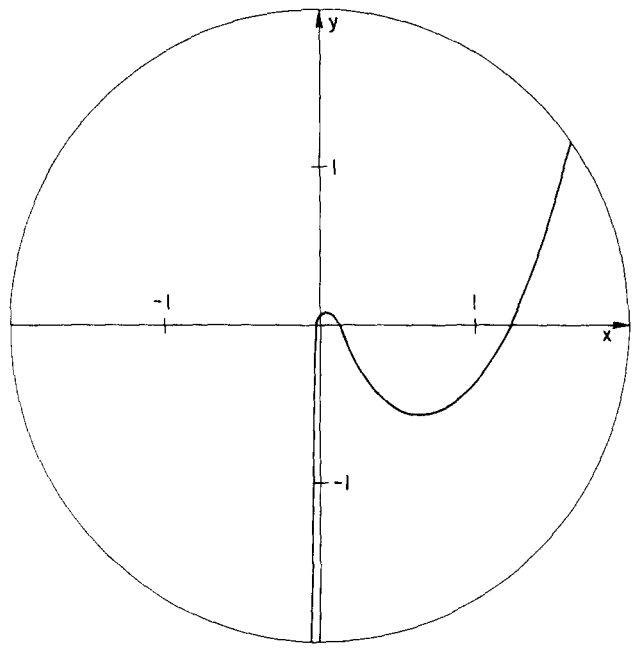

FIG. 9. Periodic orbit in (x,y) plane corresponding to the periodic sequence $(\cdots ++++--++++--\cdots)$. It intersects neither the x axis nor the y axis perpendicularly, but it is self-retracing.

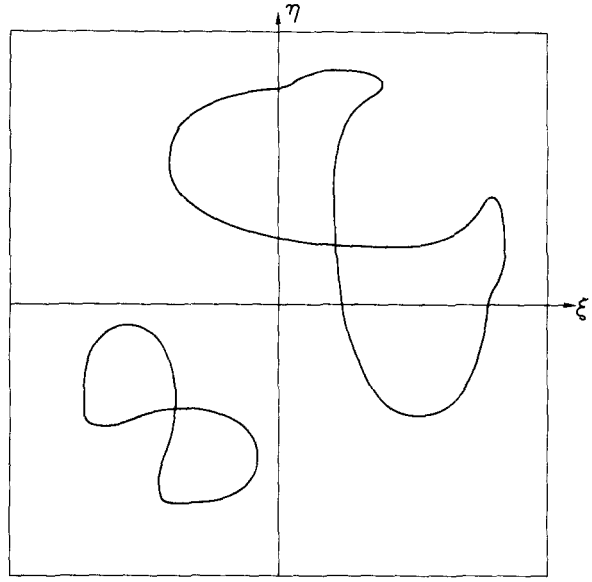

FIG. 10. Curves in the square $-1<\xi<1$, $-1<\eta<1$ which correspond to the positions $x=1.0, y=1.0$ and $x=0.5, y=0.5$. Only from the latter point are there any trajectories which lead directly to a collision ($\xi=0$ or $\eta=0$). The first curve is inverted at the origin.

rolls out the dough. He pulls the square along the ξ-axis by a factor 2, while reducing its thickness by $\frac{1}{2}$, then cuts the rectangle in two and places the right-hand side on top of the left-hand side.

The baker transformation is an area-preserving map which introduces discontinuities by tearing the original square along the η axis, i.e., for trajectories in the neighborhood of $\xi = 0$. Such a trajectory has either $a_0 > 0$ and $a_j < 0$ for $j > 0$ or $a_0 < 0$ and $a_j > 0$ for $j > 0$. The first x crossing in the forward direction happens to be a collision with the origin in this case. If the baker transformation is iterated once, new discontinuities will appear which arise for $\xi = \pm \frac{1}{2}$. These trajectories have a collision at their second x crossing in the forward direction, and so on.[7]

Also, the baker transformation allows to recognize immediately the stable and the unstable submanifold which belong to a given periodic orbit. For instance, the pseudo-circular orbit is represented by the points $(\frac{1}{3}, -\frac{1}{3})$ and $(-\frac{1}{3}, \frac{1}{3})$ in the (ξ, η) square. Obviously, any trajectory with $\xi = \pm \frac{1}{3}$ and arbitrary η moves closer to the pseudo-circular orbit with each baker transformation, while any trajectory with $\eta = \pm \frac{1}{3}$ and arbitrary ξ moves away. Both of these submanifolds are actually larger because, e.g., the trajectories with $\xi = \pm \frac{2}{3}$ and arbitrary η move into the stable submanifold $\xi = \pm \frac{1}{3}$ after one transformation. The stable and the unstable manifold of the pseudo-circular orbit intersect at $(\frac{1}{3}, \frac{1}{3})$ and $(-\frac{1}{3}, -\frac{1}{3})$ for the first time, giving rise to a homoclinic point, i.e., a trajectory which closes in on the pseudo-circular orbit both forward and backward in time, without becoming identical to the periodic orbit. In this particular case, the homoclinic point arises from a self-retracing trajectory because the sequence of intersections with the x axis is the same, forward and backward in time.

---

[1] Cf. V. Szebehely, *Theory of orbits, the restricted problem of three bodies* (Academic, New York, 1967), Chap. 9.

[2] J. Hadamard, "Sur le billiard non euclidien", Soc. Sci. Bordeaux, Procès Verbaux, **1898**, 147, (1898); J. Hadamard, J. Math. Pure Appl. **4**, 27 (1898); M. Morse, Am. J. Math. **43**, 33 (1921); E. Artin, Abhandl. Math. Sem. Hamburg 3, 170 (1924).

[3] M. C. Gutzwiller, J. Math. Phys. **8**, 1979 (1967); J. Math. Phys. **10**, 1004 (1969); J. Math. Phys. **11**, 1791 (1970); J. Math. Phys. **12**, 343 (1971).

[4] C. Siegel, *Vorlesungen uber Himmelsmechanik* (Springer-Verlag, Berlin, 1956), p. 133.

[5] J. Moser, Commun. Pure Appl. Math. **23**, 609 (1970).

[6] V. I. Arnold and A. Avez, *Ergodic problems of classical mechanics* (Benjamin, New York, 1968), p. 7.

[7] Cf. Ref. 6, p. 8.

Erratum: Analytic treatment of the Coulomb potential in the path integral formalism by exact summation of a perturbation expansion [J. Math. Phys. 13, 1070 (1972)]

M. J. Goovaerts and J. T. Devreese
Institute for Applied Mathematics, Faculty of Science, University of Antwerp, Middelheimlaan 1, Belgium
(Received 8 August 1972)

A number of paragraphs in the Introduction of this paper have to be interchanged:

The paragraphs 5: "Given this general ...,"
6: "For examples like ...,"
7: "The present paper is divided ...,"
8: "This leads to an expression ...,"
9: "Using Grosjean's theorem ...,"

should, *in this order*, come after

$$W = \int \exp(-\beta E) d\psi$$

(right before the last paragraph of the Introduction):

"In Appendices A–C ...."