Formal solutions of inverse scattering problems. II*

Reese T. Prosser

Department of Mathematics, Dartmouth College, Hanover, New Hampshire 03755
(Received 28 October 1975)

The formal solutions of inverse scattering problems presented in Paper I [J. Math. Phys. 10, 1819 (1969)] are shown here to converge in certain cases of potential scattering for sufficiently weak potentials, and in certain cases of refractive scattering for sufficiently weak variations in the index of refraction. The solutions for the cases of boundary scattering, on the other hand, are not likely to converge, because there is no way to make the effect of the boundary sufficiently weak.

INTRODUCTION

In Part I of this series,[1] formal solutions of certain inverse scattering problems were developed from a procedure suggested by Jost and Kohn[2] and developed by Moses.[3] In this paper it is shown that these formal solutions do in fact converge to give true solutions for some of these problems, and that in these cases the true solution is actually given by a constructive iteration procedure suitable for numerical computation. In each case the convergence requires that the disturbance causing the scattering be sufficiently weak that the direct scattering problem admit an iterative solution. This direct solution is then inverted to give the inverse solution.

As in Part I, three classes of problems are considered separately: problems of potential scattering, refractive scattering, and boundary scattering.

POTENTIAL SCATTERING

The scattering of a quantum mechanical wave function $\varphi(\mathbf{x},\mathbf{k})$ from a fixed potential $V(\mathbf{x})$ is governed by the time-independent Schrödinger equation

$$(\nabla^2+k^2)\varphi(\mathbf{x},\mathbf{k})=V(\mathbf{x})\varphi(\mathbf{x},\mathbf{k}). \tag{1}$$

The solution, which is to consist of an ingoing plane wave plus an outgoing scattered wave, may be expressed as

$$\varphi(\mathbf{x},\mathbf{k})=e^{i\mathbf{k}\cdot\mathbf{x}}+\int\frac{e^{i|\mathbf{k}||\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x}-\mathbf{y}|}V(\mathbf{y})\varphi(\mathbf{y},\mathbf{k})\,d\mathbf{y}. \tag{2}$$

As $|\mathbf{x}|\to\infty$ the behavior of $\varphi(\mathbf{x},\mathbf{k})$ is given by

$$\varphi(\mathbf{x},\mathbf{k})\to e^{i\mathbf{k}\cdot\mathbf{x}}+\frac{e^{i|\mathbf{k}||\mathbf{x}|}}{4\pi|\mathbf{x}|}T(\mathbf{k}',\mathbf{k})+O(1/|\mathbf{x}|^2). \tag{3}$$

Here $\mathbf{k}'=(|\mathbf{k}|/|\mathbf{x}|)\mathbf{x}$, and $T(\mathbf{k}',\mathbf{k})$ is given by

$$T(\mathbf{k}',\mathbf{k})=\int e^{-i\mathbf{k}'\cdot\mathbf{y}}V(\mathbf{y})\varphi(\mathbf{y},\mathbf{k})\,d\mathbf{y}. \tag{4}$$

Thus $T(\mathbf{k}',\mathbf{k})$ contains the scattering data. An iterative solution for $T(\mathbf{k}',\mathbf{k})$ is obtained by first solving (2) for $\varphi(\mathbf{x},\mathbf{k})$ and then substituting the result in (4):

$$\begin{aligned}T(\mathbf{k}',\mathbf{k})=&\int e^{-i\mathbf{k}'\cdot\mathbf{y}}V(\mathbf{y})\,e^{i\mathbf{k}\cdot\mathbf{y}}\,d\mathbf{y}\\&+\int\int e^{-i\mathbf{k}'\cdot\mathbf{y}_1}V(\mathbf{y}_1)\frac{e^{i|\mathbf{k}||\mathbf{y}_1-\mathbf{y}_2|}}{4\pi|\mathbf{y}_1-\mathbf{y}_2|}V(\mathbf{y}_2)\\&\times e^{i\mathbf{k}\cdot\mathbf{y}_2}d\mathbf{y}_2\,d\mathbf{y}_1+\cdots\\&+\int\cdots\int e^{-i\mathbf{k}'\cdot\mathbf{y}_1}V(\mathbf{y}_1)\frac{e^{i|\mathbf{k}||\mathbf{y}_1-\mathbf{y}_2|}}{4\pi|\mathbf{y}_1-\mathbf{y}_2|}\\&\times V(\mathbf{y}_2)\cdots V(\mathbf{y}_{n-1})\frac{e^{i|\mathbf{k}||\mathbf{y}_{n-1}-\mathbf{y}_n|}}{4\pi|\mathbf{y}_{n-1}-\mathbf{y}_n|}\\&\times V(\mathbf{y}_n)\,d\mathbf{y}_n\cdots d\mathbf{y}_1+\cdots.\end{aligned} \tag{5}$$

If Fourier transforms are taken throughout, then

$$\begin{aligned}T(\mathbf{k}',\mathbf{k})=&V(\mathbf{k}'-\mathbf{k})+\int V(\mathbf{k}'-\mathbf{k}'')\frac{1}{k''^2-k^2+i0}V(\mathbf{k}''-\mathbf{k})\,d\mathbf{k}''\\&+\cdots+\int\cdots\int V(\mathbf{k}'-\mathbf{k}'')\frac{1}{k''^2-k^2+i0}\\&\times V(k''-k''')\cdots V(\mathbf{k}^{(n)}-\mathbf{k})\,d\mathbf{k}^{(n)}\cdots d\mathbf{k}''+\cdots\end{aligned} \tag{6}$$

or, more formally,

$$T=V+V(\Gamma V)+V(\Gamma V(\Gamma V))+\cdots, \tag{7}$$

where ΓV is the kernel

$$(\Gamma V)(\mathbf{k}',\mathbf{k})=(\mathbf{k}'^2-\mathbf{k}^2+i0)^{-1}V(\mathbf{k}'-\mathbf{k}). \tag{8}$$

Now it is known that this formal solution (6) of the direct problem converges provided that the potential V is sufficiently weak.[4] To see this, we define a (well-behaved) class of integral kernels $K(\mathbf{k}',\mathbf{k})$, together with a norm $\|\ \|$ for this class, such that the class is complete in this norm, and if K and M are in the class, then

$$\|K\|\ \ \|M\|<\infty \tag{9}$$

and

$$\|K(\Gamma M)\|\leq\|K\|\ \ \|M\|, \tag{10}$$

where

$$(K(\Gamma M)(\mathbf{k}',\mathbf{k}))=\int K(\mathbf{k}',\mathbf{k}'')(\mathbf{k}''^2-\mathbf{k}^2+i0)^{-1}M(\mathbf{k}'',\mathbf{k})\,d\mathbf{k}''. \tag{11}$$

It is plain that if the kernel $V(\mathbf{k}'-\mathbf{k})$ is in this class, and if

$$\|V\|<1, \tag{12}$$

then the kernel $T(\mathbf{k}',\mathbf{k})$ is also in this class and the series (6) converges to T in norm.

There are several ways of defining such a norm. One way was originally given by K.O. Friedrichs in his study of the direct problem,[4] and for this reason these classes are sometimes called Friedrichs classes with Friedrichs norms; we summarize his results in the Appendix.

For the solution of the inverse problem we have only to invert Eq. (6). Suppose first that we know the backscattering data $T(-\mathbf{k},\mathbf{k})$ for all values of $\mathbf{k}$. Then we proceed as follows: We put $\mathbf{k}'$ equal to $-\mathbf{k}$, replace $T(-\mathbf{k},\mathbf{k})$ by $\epsilon T(-\mathbf{k},\mathbf{k})$ and $V(\mathbf{k})$ by $\sum_{m=1}^{\infty}\epsilon^m V_m(\mathbf{k})$, and substitute into (6). Then we equate the coefficients of ϵ^m. The result is

$$\begin{aligned}&T(-\mathbf{k},\mathbf{k})=V_1(-\mathbf{k}-\mathbf{k})=V_1(-2\mathbf{k}),\quad m=1,\\&0=V_m(-2\mathbf{k})+\sum_{i=2}^{m}\sum_{r_1+\cdots+r_i=m}\\&\int\cdots\int V_{r_1}(-\mathbf{k}-\mathbf{k}'')(\mathbf{k}''^2+\mathbf{k}^2+i0)^{-1}V_{r_2}(\mathbf{k}''-\mathbf{k}'')\cdots\\&\times V_{r_i}(\mathbf{k}^{(i)}-\mathbf{k})\,d\mathbf{k}^{(i)}\cdots d\mathbf{k}'',\quad m>1.\end{aligned}\tag{13}$$

Hence if we put

$$\begin{aligned}&T_1(\mathbf{k}',\mathbf{k})=T(\mathbf{k}',\mathbf{k}),\\&V_1(-2\mathbf{k})=T_1(-\mathbf{k},\mathbf{k}),\quad m=1,\\&T_m(\mathbf{k}',\mathbf{k})=-\sum_{i=2}\sum_{r_i}V_{r_i}(\Gamma V_{r_2}\cdots(\Gamma V_{r_i})\cdots)(\mathbf{k}',\mathbf{k}),\\&V_m(-2\mathbf{k})=T_m(-\mathbf{k},\mathbf{k}),\quad m>1,\end{aligned}\tag{14}$$

then

$$V'(-2\mathbf{k})=\sum_{m=1}^{\infty}V_m(-2\mathbf{k}).\tag{15}$$

Equation (14) gives a potential $V'(-2\mathbf{k})$ in terms of the backscattering data $T(-\mathbf{k},\mathbf{k})$ and thus provides a formal solution of the inverse potential problem. It remains to show that if V is sufficiently weak, i.e., if the Friedrichs norm of V is sufficiently small, then the series (15) actually converges to V' and that V' actually reproduces the backscattering data.

In order to do so, we must first verify that the Friedrichs norms have the following property. If K is any kernel of the Friedrichs class and M is derived from K by the following formula,

$$M(\mathbf{k}',\mathbf{k})=K((\mathbf{k}-\mathbf{k}')/2,\ (\mathbf{k}'-\mathbf{k})/2),\tag{16}$$

then the Friedrichs norms of M and K are related by

$$\|M\|\leq\|K\|.\tag{17}$$

The verification of (16) is included in the Appendix for the norms defined there.

Now if $V(\mathbf{k}'-\mathbf{k})$ belongs to the Friedrichs class, with $\|V\|<1$, we know from (7) that $T(\mathbf{k}',\mathbf{k})$ also belongs to the Friedrichs class, and

$$\|T\|\leq\sum_{k=1}^{\infty}\|V\|^k=\frac{\|V\|}{1-\|V\|}\cdot\tag{18}$$

Now from (14), (16), and (17) it follows that V_1 also belongs to the Friedrichs class and $\|V_1\|\leq\|T\|$. Hence so does V_m, and

$$\|V_m\|\leq\|T_m\|=\sum_{i=2}^{m}\|V_{r_1}\|\cdots\|V_{r_i}\|.\tag{19}$$

Now, following Jost and Kohn,[2] we put

$$\begin{aligned}&J_1=\|V_1\|,\\&J_m=\sum_{i=2}^{m}\sum_{r_1+\cdots+r_i=m}J_{r_1}\cdots J_{r_i};\end{aligned}\tag{20}$$

then we see from (19) that

$$\|V_m\|\leq J_m.\tag{21}$$

Now we put

$$w=w(z)=\sum_{m=1}^{\infty}J_m z^m\tag{22}$$

and observe that, because of (20),

$$\begin{aligned}w(z)-J_1z&=1/[1-w(z)]-1-w(z)\\&=w^2(z)/[1-w(z)],\end{aligned}\tag{23}$$

or

$$J_1z=w-w^2/(1-w).\tag{24}$$

Hence, if we solve (24) for w,

$$w=\{(1+J_1z)\pm[(1+J_1z)^2-8J_1z]^{1/2}\}/4,\tag{25}$$

then we see that, as a function of z, w is analytic at $z=0$ and out to the nearest singularity, where

$$(1+J_1z)^2-8J_1z=0\tag{26}$$

or where

$$J_1z=3-2\sqrt{2}=0.172.\tag{27}$$

Hence, if $J_1<0.172$, then the series (22) for $w(z)$ converges for $|z|\leq1$, and it follows from (21) that the series (15) for V' converges in the Friedrichs norm.

It remains to show that the sum of this series reproduces the backscattering data, i.e., that the sum satisfies Eq. (6). But, if we put

$$V^N=\sum_{m=1}^{N}V_m\tag{28}$$

and

$$T^N=V^N+V^N(\Gamma V^N)+V^N(\Gamma V^N(\Gamma V^N))+\cdots,\tag{29}$$

then we find

$$\begin{aligned}&\sup_{\mathbf{k}}|T(-\mathbf{k},\mathbf{k})-T^N(-\mathbf{k},\mathbf{k})|\\&\leq\sup_{\mathbf{k}}\Big|\sum_{m=N+1}^{\infty}\sum_{i=2}^{m}\sum_{\substack{r_1+\cdots+r_i=m\\ r_i\leq N}}V_{r_i}(\Gamma V_{r_2}\cdots(\Gamma V_{r_i})-)(-\mathbf{k},\mathbf{k})\Big|\\&\leq\sum_{m=N+1}^{\infty}\sum_{i=2}^{m}\sum_{\substack{r_1+\cdots+r_i=m\\ r_1\leq N}}\|V_{r_1}\|\cdots\|V_{r_i}\|\\&\leq\sum_{m=N+1}^{\infty}J_m\to0\ \text{ as } N\to\infty.\end{aligned}\tag{30}$$

It follows that if one starts with the backscattering data $T(-\mathbf{k},\mathbf{k})$ and constructs a potential V' according to (14) and (15), and then reconstructs the backscattering data from (6), one recovers $T(-\mathbf{k},\mathbf{k})$.

There is no guarantee, in general, that the potential V' so constructed will reproduce the *rest* of the T matrix, i.e., there is no guarantee that $T^N(\mathbf{k}',\mathbf{k})\to T(\mathbf{k}',\mathbf{k})$ for $\mathbf{k}'\neq-\mathbf{k}$, or that $\|T-T^N\|\to0$ as $N\to\infty$.

Consequently, there is no gaurantee, in general, that the potential V' so constructed will coincide with the original potential V of the problem. In fact, it follows from (7) that

$$T=V+V\Gamma T\tag{31}$$

so that, as kernels,

$$T(1+\Gamma T)^{-1}=V. \tag{32}$$

Now clearly, if $V'=V$, then V' will reproduce the entire T matrix through (7); and conversely, if V' does reproduce the entire T matrix, then $V'=V$ through (32). We know that $V'=V$ in the case of radial potentials through the Gel'fand—Levitan theory, but we can say nothing more in the general case.

If instead of the backscattering data $T(-\mathbf{k},\mathbf{k})$ for all $\mathbf{k}$ we are given the fixed aspect data $T(\mathbf{k}',\mathbf{k})$ for fixed aspect angle $\mathbf{k}/|\mathbf{k}|$ and all energies $|\mathbf{k}'|=|\mathbf{k}|$ and scattering angles $\mathbf{k}'/|\mathbf{k}'|$, then we modify the inversion procedure as follows.[3] We put

$$V_m(2\mathbf{h})=T_m(\mathbf{k}',\mathbf{k}),\quad 2\mathbf{h}=\mathbf{k}'-\mathbf{k}. \tag{33}$$

This determines V_m only in the half-space $\mathbf{h}\cdot\mathbf{k}<0$. But, if $V(\mathbf{x})$ is to be real, then we must have $V(-\mathbf{k})=\overline{V(\mathbf{k})}$. Hence we must supplement (33) with

$$V_m(-2\mathbf{h})=\overline{V_m(2\mathbf{h})}.$$

The rest of the argument goes through as before.

A similar modification works if we are given fixed scattering angle data $T(\mathbf{k}',\mathbf{k})$ for fixed $\mathbf{k}'/|\mathbf{k}|$, all $|\mathbf{k}|=|\mathbf{k}'|$ and all aspects $\mathbf{k}/|\mathbf{k}|$. The roles of $\mathbf{k}$ and $\mathbf{k}'$ are now interchanged, and (33) determines $V_m(2\mathbf{h})$ only in the half-space $\mathbf{h}\cdot\mathbf{k}'>0$, while (34) determines $V_m(2\mathbf{h})$ in the other half-space. Otherwise the result is the same as before.

We are not able to prove, however, that the potentials constructed from the three kinds of data described above agree with the original potential or with each other!

REFRACTIVE SCATTERING

The scattering of an acoustic wavefunction $\varphi(\mathbf{x},\mathbf{k})$ from a variable index of refraction $n(\mathbf{x})$ is governed by the wave equation

$$(\nabla^2+\mathbf{k}^2 n(\mathbf{x}))\varphi(\mathbf{x},\mathbf{k})=0. \tag{34}$$

If we put $W(\mathbf{x})=1-n(\mathbf{x})$, then (34) becomes

$$(\nabla^2+k^2)\varphi(\mathbf{x},\mathbf{k})=k^2W(\mathbf{x})\varphi(\mathbf{x},\mathbf{k}). \tag{35}$$

This equation resembles (1) with $V(\mathbf{x})$ replaced by $k^2W(\mathbf{x})$. Again we seek a solution of the form

$$\varphi(\mathbf{x},\mathbf{k})=e^{i\mathbf{k}\cdot\mathbf{x}}+\int\frac{e^{i|\mathbf{k}||\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x}-\mathbf{y}|}k^2W(\mathbf{y})\varphi(\mathbf{y},\mathbf{k})\,dy. \tag{36}$$

Again as $|\mathbf{x}|\to\infty$, we have

$$\varphi(\mathbf{x},\mathbf{k})\to e^{i\mathbf{k}\cdot\mathbf{x}}+\frac{e^{i|\mathbf{k}||\mathbf{x}|}}{4\pi|\mathbf{x}|}T(\mathbf{k}',\mathbf{k})+O\left(\frac{1}{|\mathbf{x}|^2}\right), \tag{37}$$

where now

$$T(\mathbf{k}',\mathbf{k})=\int e^{-i\mathbf{k}'\cdot\mathbf{y}}k^2W(\mathbf{y})\,\varphi(\mathbf{y},\mathbf{k})\,d\mathbf{y}. \tag{38}$$

Here we may substitute the solution of (36) for $\varphi(\mathbf{y},\mathbf{k})$ and obtain the analog of (6),

$$\begin{aligned}T(\mathbf{k}',\mathbf{k})=&\,k^2W(\mathbf{k}'-\mathbf{k})\\&+\int k^2W(\mathbf{k}'-\mathbf{k}'')(\mathbf{k}''^2-\mathbf{k}^2+i0)^{-1}k^2W(\mathbf{k}''-\mathbf{k})\,d\mathbf{k}''\\&+\cdots,\end{aligned} \tag{39}$$

or the analog of (7),

$$T=k^2[W+W(\Delta W)+W(\Delta W(\Delta W))+\cdots], \tag{40}$$

where now we have put

$$\begin{aligned}(\Delta K)(\mathbf{k}',\mathbf{k})&=(\mathbf{k}'^2-\mathbf{k}^2+i0)^{-1}\mathbf{k}^2K(\mathbf{k}',\mathbf{k})\\&=(\mathbf{k}'^2-\mathbf{k}^2+i0)^{-1}\mathbf{k}'^2K(\mathbf{k}',\mathbf{k})-K(\mathbf{k}',\mathbf{k}).\end{aligned} \tag{41}$$

It follows that we can reproduce our results for potential scattering word for word as soon as we define a Friedrichs class and Friedrichs norm appropriate for Δ instead of Γ; i.e., instead of (10) we must now have

$$\|K\Delta M\|\leqslant\|K\|\,\|M\|. \tag{42}$$

The proof that this can also be done is sketched in the Appendix.

In terms of the new Friedrichs norm we see that the series (40) for T converges to T provided that W lies in the Friedrichs class and $\|W\|<1$. In this case we may define the inversion procedure as follows.

If we are given the backscattering data $T(-\mathbf{k},\mathbf{k})$ for all $\mathbf{k}$, then we set

$$\begin{aligned}&T_1(\mathbf{k}',\mathbf{k})=T(\mathbf{k}',\mathbf{k}),\\&\mathbf{k}^2W_1(-2\mathbf{k})=T_1(-\mathbf{k},\mathbf{k}),\\&T_m(\mathbf{k}',\mathbf{k})=-\sum_{i=3}^{m}\sum_{r_i}W_{r_1}(\Delta W_{r_2}\cdots(\Delta W_{r_1})\cdots)(\mathbf{k}',\mathbf{k}),\\&\mathbf{k}^2W_m(-2\mathbf{k})=T_m(-\mathbf{k},\mathbf{k}),\end{aligned} \tag{43}$$

and

$$\mathbf{k}^2W'(-2\mathbf{k})=\sum_{m=1}^{\infty}k^2W_m(-2\mathbf{k}). \tag{44}$$

Then the proof of the convergence of (44) to an index of refraction which reproduces the backscattering data is exactly the same as for the case of potential scattering given above and holds under exactly the same conditions. Similar results hold in the cases of fixed aspect angle data and fixed scattering angle data.

BOUNDARY SCATTERING

The scattering of an acoustic wavefunction $\varphi(\mathbf{k},\mathbf{x})$ from an acoustically soft boundary is governed by the wave equation with Dirichlet boundary condition

$$\begin{aligned}(\nabla^2+k^2)\varphi(\mathbf{x},\mathbf{k})&=0,\quad x\in R',\\\varphi(\mathbf{x},\mathbf{k})&=0,\quad x\in\partial R',\end{aligned} \tag{45}$$

Here R' is the exterior of a compact region R in E^3 with smooth boundary ∂R. Again the solution may be expressed in the form

$$\varphi(\mathbf{x},\mathbf{k})=e^{i\mathbf{k}\cdot\mathbf{x}}+2\int_{\partial R}\frac{e^{i|\mathbf{k}||\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x}-\mathbf{y}|}\frac{\partial}{\partial\mathbf{n}(\mathbf{y})}\varphi(\mathbf{y},\mathbf{k})\,dy. \tag{46}$$

Here $\partial/\partial\mathbf{n}(\mathbf{y})$ denotes the exterior normal derivative on the boundary, and the integration is taken over the boundary. As $|\mathbf{x}|\to\infty$

$$\varphi(\mathbf{x},\mathbf{k})\to e^{i\mathbf{k}\cdot\mathbf{x}}+\frac{e^{i|\mathbf{k}||\mathbf{x}|}}{4\pi|\mathbf{x}|}T(\mathbf{k}',\mathbf{k})+O\left(\frac{1}{|\mathbf{x}|^2}\right), \tag{47}$$

where now

$$T(\mathbf{k}',\mathbf{k})=2\int_{\partial R}e^{-i\mathbf{k}'\cdot\mathbf{y}}\frac{\partial}{\partial\mathbf{n}(\mathbf{y})}\varphi(\mathbf{y}\cdot\mathbf{k})\,d\mathbf{y}. \tag{48}$$

Solving (46) formally for $\varphi(\mathbf{y},\mathbf{k})$ and substituting in (48), we find

$$T(\mathbf{k}',\mathbf{k})=2\int_{\partial R} e^{-i\mathbf{k}'\cdot\mathbf{y}}\frac{\partial}{\partial \mathbf{n}(\mathbf{y})}e^{i\mathbf{k}\cdot\mathbf{y}}\,d\mathbf{y}$$
$$+4\iint e^{-i\mathbf{k}'\cdot\mathbf{y}_1}\frac{\partial}{\partial \mathbf{n}(\mathbf{y}_1)}\frac{e^{i|\mathbf{k}||\mathbf{y}_1-\mathbf{y}_2|}}{4\pi|y_1-y_2|}$$
$$\times\frac{\partial}{\partial \mathbf{n}(\mathbf{y}_2)}e^{i\mathbf{k}\cdot\mathbf{y}_2}d\mathbf{y}_2d\mathbf{y}_1+\cdots. \tag{49}$$

This may be written in terms of the characteristic function $X_R(\mathbf{x})$ of R

$$X_R(\mathbf{x})\begin{cases}1, & \mathbf{x}\in R\cup\partial R,\\ 0, & \mathbf{x}\in R'\end{cases} \tag{50}$$

and volume integrals

$$T(\mathbf{k}',\mathbf{k})=2\int_{E^3} e^{-i\mathbf{k}'\cdot\mathbf{y}}\nabla X_R(\mathbf{y})\cdot\nabla e^{i\mathbf{k}\cdot\mathbf{y}}d\mathbf{y}$$
$$+4\iint_{E^3} e^{-i\mathbf{k}'\cdot\mathbf{y}}\nabla X_R(\mathbf{y})\cdot\nabla\frac{e^{i|\mathbf{k}||\mathbf{y}_1-\mathbf{y}_2|}}{4\pi|\mathbf{y}_1-\mathbf{y}_2|}$$
$$\times\nabla X_R(\mathbf{y}_2)\cdot\nabla e^{i\mathbf{k}\cdot\mathbf{y}_2}d\mathbf{y}_2d\mathbf{y}_1+\cdots. \tag{51}$$

If Fourier transforms are taken throughout, then

$$T(\mathbf{k}',\mathbf{k})=2X_R(\mathbf{k}'-\mathbf{k})(\mathbf{k}'-\mathbf{k})\cdot\mathbf{k}$$
$$+4\int X_R(\mathbf{k}'-\mathbf{k}'')(\mathbf{k}'-\mathbf{k}'')\cdot\mathbf{k}''(\mathbf{k}''^2-\mathbf{k}^2+i0)^{-1}$$
$$\cdot X_R(\mathbf{k}''-\mathbf{k})(\mathbf{k}''-\mathbf{k})\cdot\mathbf{k}\,d\mathbf{k}''\cdots, \tag{52}$$

or formally

$$T=Z(\Gamma Z)+Z(\Gamma Z(\Gamma Z))+\cdots, \tag{53}$$

where

$$Z(\mathbf{k}',\mathbf{k})=2X_R(\mathbf{k}'-\mathbf{k})(\mathbf{k}'-\mathbf{k})\cdot\mathbf{k} \tag{54}$$

and Γ is given by (8) as before.

Thus the series (53) will converge to T as soon as the kernel Z belongs to a Friedrichs class with Friedrichs norm $\|Z\|<1$. If this were the case, then (53) could be inverted by the same procedure used to invert (7) above. In fact, it suffices to find a norm for which $\|Z\|<\infty$. To see this, recall that, under the dilation $\mathbf{x}\to a\mathbf{x}$, $\mathbf{k}\to a^{-1}\mathbf{k}$, the characteristic function $X_R(k)\to X_{aR}(k)=a^{-1}X_R(ak)$ and hence $Z(\mathbf{k}',\mathbf{k})=Z_R(\mathbf{k}',\mathbf{k})=aZ_R(a\mathbf{k}',a\mathbf{k})$.[5] Thus, if $\|Z_R\|<\infty$ and if a is sufficiently small, then $\|Z_{aR}\|<1$ and the series (53) converges. The inversion procedure would then give Z_{aR}, and from this aR, and hence R, could be recovered.

Unfortunately, we have been unable to find a Friedrichs norm for which $\|Z_R\|<\infty$ for any choice of region R. The difficulty lies in the fact that the discontinuity of $X_R(\mathbf{x})$ at the boundary ensures that $Z(\mathbf{k}',\mathbf{k})$ will not die out for large $\mathbf{k}$, even for the smoothest of boundaries, and it follows that Z will belong to none of the usual Friedrichs classes.

Moreover, we know that the solution of the direct boundary scattering problem considered here is a limit of a sequence of solutions of direct potential scattering problems whose potentials V_n have increasing Friedrichs norm: $\|V_n\|\uparrow\infty$.[5] Thus it seems unlikely that the Jost and Kohn inversion procedure presented here can be used in any but a formal way for boundary scattering problems.

APPENDIX: FRIEDRICHS CLASSES

We briefly summarize here the essential features of those classes of integral kernels first introduced by Friedrichs in his studies of the direct potential scattering problems.[4] We shall deal here only with kernels in three dimensions.[6,7]

Let $K(\mathbf{k}',\mathbf{k})$ be an integral kernel defined for $\mathbf{k}',\mathbf{k}\in\mathbb{R}^3$, and let $0<\theta$ and $0\leq\mu\leq 1$ be positive numbers. We say that K has *decay of order θ at ∞* if

$$|K(\mathbf{k}',\mathbf{k})|<C(1+|\mathbf{k}'-\mathbf{k}|)^{-\theta} \tag{A1}$$

and *smoothness of order μ locally* if

$$|K(\mathbf{k}'+\mathbf{h}',\mathbf{k}+\mathbf{h})-K(\mathbf{k},\mathbf{h})|$$
$$<C(1+|\mathbf{k}'-\mathbf{k}|)^{-\theta}(|\mathbf{h}'|^{\mu}+|\mathbf{h}|^{\mu}). \tag{A2}$$

These bounds are to hold for all $\mathbf{h}',\mathbf{k}',\mathbf{h},\mathbf{k}\in\mathbb{R}^3$ with $|\mathbf{h}'|\leq 1$, $|\mathbf{h}|\leq 1$.

Let $\mathcal{K}(\theta,u)$ be the Friedricks class of all integral kernels with decay of order θ and smoothness or order μ. For each member K of this class we define the Friedrichs norm

$$\|K\|_{\theta,u}=a\sup_{\substack{\mathbf{k}',\mathbf{k}\\|\mathbf{h}'|<1,\,|\mathbf{h}|<1}}\left\{[1+(\mathbf{k}'-\mathbf{k})]^{-\theta}\left(|K(\mathbf{k}',\mathbf{k})|+\frac{(K(\mathbf{k}'+\mathbf{h}',\mathbf{k}+\mathbf{h})-K(\mathbf{k}',\mathbf{k}))}{|\mathbf{h}'|^{\mu}+|\mathbf{h}|^{\mu}}\right)\right\}, \tag{A3}$$

where a is a constant to be chosen below.

With these definitions it is then an arduous but straightforward task to show that $\mathcal{K}$ is a linear space which is complete in this norm.

Moreover, if K and M are any two members of $\mathcal{K}$ and if we define ΓM by

$$(\Gamma M)(\mathbf{k}',\mathbf{k})=(\mathbf{k}'^2-\mathbf{k}^2+i0)^{-1}M(\mathbf{k}',\mathbf{k}), \tag{A4}$$

then ΓM is a singular kernel not in $\mathcal{K}$; but, if $0<\mu<1$ and $1<\theta$, then the composition $K(\Gamma M)$ is again in $\mathcal{K}$ and

$$\|K\Gamma M\|\leq b\,\|K\|\,\|M\| \tag{A5}$$

for some choice of constant b depending only on μ and θ. This result is essentially a consequence of the Privalov lemmas.[6]

If we now choose a in (A3) so that $a=b^{-1}$, then we arrive at (10), provided $1<\theta$, $0<\mu<1$.

Finally, we note that if M is derived from K by the relation

$$M(\mathbf{k}',\mathbf{k})=K((\mathbf{k}'-\mathbf{k})/2,(\mathbf{k}-\mathbf{k}')/2) \tag{A6}$$

then it is clear from (A3) that $\|M\|\leq\|K\|$. and so (17) holds.

For refractive scattering we must replace ΓM by ΔM, defined by

$$\Delta M(\mathbf{k}',\mathbf{k}) = (\mathbf{k}'^2 - \mathbf{k}^2 + i0)^{-1}\mathbf{k}^2 M(\mathbf{k}',\mathbf{k})$$
$$= (\mathbf{k}'^2 - \mathbf{k}^2 + i0)^{-1}\mathbf{k}'^2 M(\mathbf{k}',\mathbf{k}) + M(\mathbf{k}',\mathbf{k}). \quad \text{(A7)}$$

Again ΔM is a singular kernel not in K, but now if $0 < \mu < 1$ and $\frac{3}{2} < \theta$, then the composition $K(\Delta M)$ is again in K, and

$$\|K\Delta M\| \leq c\|K\|\,\|M\| \quad \text{(A8)}$$

for some constant c depending only on μ and θ. The proof is an obvious modification of that of (A5). If we now choose a in (A3) so that $a = c^{-1}$ then we arrive at (42), as required, provided $0 < \mu < 1$, $\frac{3}{2} < \theta$.

*This work was supported in part by AFOSR Grant No. 71-2085.

[1]R. T. Prosser, "Formal Solutions of Inverse Scattering Problems," J. Math. Phys. 10, 1819 (1969).
[2]R. Jost and W. Kohn, Phys. Rev. 37, 977 (1952).
[3]Harry Moses, Phys. Rev. 102, 559 (1956).
[4]K. O. Friedrichs, *Perturbations of Spectra in Hilbert Spaces* (Amer. Math. Soc., Providence, R.I., 1965).
[5]P. D. Lax and R. L. Phillips, Comm. Pure Appl. Math. 22, 737 (1969).
[6]L. D. Faddeev, *Mathematical Aspects of the Three Body Problem* (Steklov Math. Inst., 1963), Vol. 69 [Eng. transl.: Sivan, Jerusalem, 1965)].
[7]J. Schwartz, Comm. Pure Appl. Math. 13, 609 (1960).

On induced representations for finite groups

Nigel B. Backhouse

Department of Applied Mathematics & Theoretical Physics, University of Liverpool, Liverpool, United Kingdom

Patricia Gard

The Cavendish Laboratory, Madingley Road, Cambridge, United Kingdom

(Received 28 October 1975)

Two new notions are introduced as tools for the study of representations of finite groups. First, in the spirit of duality, a basis set of class orientated characters is shown to possess nice properties with respect to induction and subduction, which lead to simple proofs of some well-known theorems. Secondly, as a useful device in constructive representation theory, a study is made of the subrepresentations which often naturally occur when a representation is induced from a subgroup to a supergroup.

1. INTRODUCTION

In this paper we explore two areas of representation theory which have been of interest to applied group theorists, namely duality for finite groups and the construction of group representations by induction from subgroups.

Duality theory is seen to best advantage in the case of the symmetric groups, these being remarkable among finite groups for many reasons, but in particular for the natural correspondence which exists between their conjugacy classes and their unitary irreducible representations (UIR's). The correspondence is achieved by labelling the elements of both sets by Young diagrams. It has proved to be of considerable calculational value. Unfortunately, no such explicit relationship has yet been found for arbitrary finite groups, but, nevertheless, there is a certain duality between the algebra of classes and the algebra of representations which has yielded many new and interesting results (see Refs. 1—7). Typically in duality theory, by the replacement of simple characters by suitably normalized class sums, a representation-theoretic formula is transformed into a class-theoretic formula whose validity, though not guaranteed, is often the case.

In the first three sections of this paper we develop a new relationship between classes and characters by introducing the notion of a lonely character. This is defined as follows: If C_i denotes the ith conjugacy class of the finite group G, then the ith lonely character is the class function χ^i which takes the integer value $|G|/|C_i|$ on the elements of the class C_i and the value zero elsewhere. This definition and terminology is due to C.J. Bradley. We show that these class functions possess very simple properties with respect to the operations of inducing and restricting, which give rise to trivial proofs of the Frobenius reciprocity theorem, the permanence theorem and its inverse, the latter involving Robinson's inverse restriction operation.

In the last two sections we introduce the concept of partial induction, for which the basic objects are a finite group G and a representation D of a subgroup H. The operation of inducing D to obtain a representation of G is well defined. However, it frequently occurs in concrete examples that the carrier space V of D is contained in a larger space W which is invariant under G. Then it is possible to lift D to the largest subgroup $K \geq H$ of G which maps V into itself, before inducing up to G. It turns outs that the resulting partially induced representation is a subrepresentation of D induced to G. Hence partial induction lies somewhere between lifting and full induction. We find that this concept allows us to rederive rather simply an important formula on outer products for symmetric groups and also a key result in little group theory for finite groups, hinting at its power in constructive representation theory.

In this paper we use the following notation: χ_i^μ is the value of the μth simple character χ^μ on the ith class C_i, and χ_j^i is the value of the ith lonely character on C_j. We distinguish between χ_i and $\chi(C_i)$, the latter having value $|C_i|\chi_i$, regarding the character χ as a linear functional on the module structure of the group algebra of G. The symbols R and I placed before a character indicate the operations of restricting and inducing, respectively. It is usually obvious from context which groups are involved. Also, where it is well defined, R^{-1} denotes the inverse of restricting as defined by Robinson.

2. LONELY CHARACTERS

The ith lonely character χ^i can be more compactly defined to be that class function whose value on the jth conjugacy class is

$$\chi_j^i = (|G|/|C_i|)\delta_{ij}, \qquad (2.1)$$

where δ_{ij} is the Kronecker delta. It easily follows that

$$\chi^\mu = (1/|G|)\sum_i |C_i|\chi_i^\mu \chi^i. \qquad (2.2)$$

We recall from Ref. 8 that the set of complex-valued class functions, of which χ^i and χ^μ are elements, form a ring $C(G)$ under pointwise addition and multiplication. It is clear from the definition that $\chi^i\chi^j = (|G|/|C_i|\chi^i\delta_{ij}$; hence, the lonely characters form a set of orthogonal essential idempotents within $C(G)$. We also know that $C(G)$ is a vector space over the complex numbers and possesses an inner product $\langle\ ,\ \rangle$ defined by $\langle\phi, \psi\rangle = (1/|G|)\sum_{g\in G}\overline{\phi}(g)\psi(g)$. Equation (2.2) is a particular case of the statement that the lonely characters form a basis for $C(G)$, and in particular we find

TABLE I. In row 1 an entry gives a class label and the size of the corresponding class.

C_i	$(1^3)1$	$(2,1)3$	$(3)2$
χ^μ			
[3]	1	1	1
[2,1]	2	0	−1
$[1^3]$	1	−1	1

$$\chi^i = \sum_\mu \bar{\chi}_i^\mu \chi^\mu, \tag{2.3}$$

using the orthogonality properties of the χ^μ with respect to the inner product. The latter property is also possessed by the lonely characters, for we find

$$\langle \chi^i, \chi^j \rangle = (1/|G|)\sum_{g\in G} \bar{\chi}^i(g)\chi^j(g) = (|G|/|C_i|)\delta_{ij}. \tag{2.4}$$

From an algebraic point of view it would seem that lonely characters are much easier to handle than simple characters, yet contain as much information. Of course, they have the disadvantage of not being associated directly with representations.

We now consider an example. The ordinary character table of the permutation group S_3 is displayed in Table I, where the rows and columns are labelled by simple characters and conjugacy classes, respectively. According to definition the lonely character table is the sparse array of numbers in Table II. We quickly verify Eq. (2.3) by directly computing $\chi^{(1^3)} = [3] + 2[2,1] + [1^3]$, $\chi^{(2,1)} = [3] - [1^3]$, $\chi^{(3)} = [3] - [2,1] + [1^3]$. In this case, since S_3 has a real ordinary character table, the linear combinations of the simple characters required to form the lonely characters are obtained by reading the columns of Table I. The orthogonality of the rows of Table II is obvious. We note that, in the case of S_3 (generally S_n), the lonely characters are examples of generalized characters, that is, integral linear combinations of simple characters. Such characters are of great importance in group theory (see, for example, Ref. 8)

We have formed the lonely character table of G by taking linear combinations of the rows of its table of simple characters leaving unchanged the column labels, the classes. In the spirit of duality the same array of numbers can be formed by evaluating the simple characters on certain linear combinations of the classes. Define the μth lonely class

$$C_\mu = \sum_i \bar{\chi}_i^\mu C_i, \tag{2.5}$$

this being a member of the group algebra $A(G)$, where $C_i = \sum_{c_i \in C_i} c_i$. Then $\chi^\nu(C_\mu) = |G|\delta_{\mu\nu}$, the analog of (2.1), since the latter can be rewritten as $\chi^i(C_j) = |G|\delta_{ij}$. In our example we see that $C_{[3]} = (1^3) + (2,1) + (3)$, $C_{[2,1]} = 2(1^3) - (3)$, $C_{[1^3]} = (1^3) - (2,1) + (3)$. We also note the relations

$$\chi^i(C_\mu) = |G|\bar{\chi}_i^\mu = (|G|/|C_i|)\bar{\chi}^\mu(C_i). \tag{2.6}$$

We have seen that the two processes of replacing simple characters by lonely characters and replacing ordinary classes by lonely classes independently lead to the lonely character table. Equation (2.6) shows that if these two processes are carried out simultaneously, apart from the factor $|G|$, we produce the complex conjugate of the ordinary character table.

The quantity C_μ which we have defined occurs naturally in another context, namely in the analysis of the regular representation of a group. Indeed, in Ref. 8, it is shown that the rational multiple $(d_\mu/|G|)C_\mu$, $d_\mu = \dim\chi^\mu$, is the uniquely determined central idempotent in $A(G)$ which generates the two-sided ideal affording the d_μ copies of the UIR associated with χ^μ in the regular representation of G.

All of these results have projective analogs, once we recall (Ref. 9) that any projective representation can be so adjusted that its trace function becomes a class function. Thus the lonely characters can also be used as a basis set for expanding projective characters.

3. INDUCING AND RESTRICTING

In this section we prove that an induced lonely character is a lonely character and that the restriction of a lonely character of G to a subgroup H is an integral sum of lonely characters of H. First we need some elementary definitions and results.

If C_i is the ith class of G, then $C_i \cap H$ is a disjoint union of whole classes of H, written $\cup_\alpha C_{i,\alpha}$. Let $c_{i,\alpha} \in C_{i,\alpha}$ be fixed once and for all. Then $H_{i,\alpha} = \{g \in G: g^{-1}c_{i,\alpha}g \in C_{i,\alpha}\}$ consists of whole cosets of H in G and is a subgroup if H is normal in G. Define $N^G{}_{i,\alpha} = \{g\in G: g^{-1}c_{i,\alpha}g = c_{i,\alpha}\}$, then $N^G{}_{i,\alpha}$, a subgroup of the set $H_{i,\alpha}$, contains $N^H_{i,\alpha} = N^G_{i,\alpha}\cap H$ and, moreover, $|H_{i,\alpha}| \times |N^G_{i,\alpha}| = |C_{i,\alpha}|\,|N^G_{i,\alpha}|$. Then we may prove

Theorem 1: Let H be a subgroup of G and let χ^i, $\chi^{i,\alpha}$ be lonely characters corresponding to the conjugacy classes C_i, $C_{i,\alpha}$ of G, H, respectively, where $C_i \cap H = \cup_\alpha C_{i,\alpha}$. Then $I(\chi^{i,\alpha}) = \chi^i$, for all α.

Proof: If $g \in G$, then by the definition of an induced character (Ref. 10)

$$I(\chi^{i,\alpha})(g) = \chi^{i,\alpha}_{i,\alpha}\,\eta_{i,\alpha}(g),$$

where $\chi^{i,\alpha}_{i,\alpha} = |H|/|C_{i,\alpha}|$ and $\eta_{i,\alpha}(g)$ is the number of coset representatives of H in G which conjugate g into $C_{i,\alpha}$. Clearly $\eta_{i,\alpha}(g)$ is zero if $g \notin C_i$ and is equal to $\eta_{i,\alpha} = |H_{i,\alpha}|/|H|$ if $g\in C_i$. Hence

$$I(\chi^{i,\alpha}) = \frac{|H|}{|C_{i,\alpha}|}\times\frac{|H_{i,\alpha}|}{|H|}\times\frac{|C_i|}{|G|}\chi^i.$$

But $|H_{i,\alpha}| = |C_{i,\alpha}|\,|N^G_{i,\alpha}|$ and $|N^G_{i,\alpha}| = |G|/|C_i|$, giving $I(\chi^{i,\alpha}) = \chi^i$, as required.

TABLE II.

C_i	(1^3)	$(2,1)$	(3)
χ^i			
$\chi^{(1^3)}$	6	0	0
$\chi^{(2,1)}$	0	2	0
$\chi^{(3)}$	0	0	3

Theorem 2: Using the notation of Theorem 1 and its proof, we have

$$R(\chi^i)=\delta_i\sum_\alpha \eta_{i,\alpha}\chi^{i,\alpha}, \tag{3.1}$$

where $\delta_i=0$ if $C_i\cap H$ is empty but is unity otherwise.

Proof: Clearly $R(\chi^i)$ is zero if $C_i\cap H$ is empty. So assume that $C_i\cap H$ is not empty; then χ^i takes the value $|G|/|C_i|$ on each element of $C_{i,\alpha}$, for every α, but is zero outside C_i. Thus we have

$$R(\chi^i)=\sum_\alpha \frac{|G|}{|C_i|}\times\frac{|C_{i,\alpha}|}{|H|}\chi^{i,\alpha}$$

$$=\sum_\alpha \eta_{i,\alpha}\chi^{i,\alpha},$$

as required

Corollary: If H is a normal subgroup of G, $\eta_{i,\alpha}$ is independent of α and therefore can be denoted η_i. Then we have

$$R(\chi^i)=\delta_i\eta_i\sum_\alpha \chi^{i,\alpha}.$$

We now give two applications of lonely characters.

Frobenius reciprocity theorem: Let H be a subgroup of G and let θ, ψ be characters of H, G, respectively. Then

$$\langle I(\theta),\psi\rangle_G=\langle\theta,R(\psi)\rangle_H, \tag{3.2}$$

where $\langle\ ,\ \rangle_G$, $\langle\ ,\ \rangle_H$ denote the usual inner products on $C(G)$, $C(H)$, respectively.

Proof: By using the bilinearity of the inner products, it is clear that the validity of (3.2) can be tested by letting θ, ψ run over basis sets in $C(H)$, $C(G)$, respectively. For example the lonely characters will do. So take $\theta=\chi^{i,\alpha}$ and $\psi=\chi^j$. Then

$$\langle I(\chi^{i,\alpha}),\ \chi^j\rangle_G=\langle\chi^i,\chi^j\rangle_G=(|G|/|C_i|)\delta_{ij},$$

using Theorem 1 and (2.4). Also

$$\langle\chi^{i,\alpha},R(\chi^j)\rangle_H$$

$$=\delta_j\sum_\beta \eta_{j,\beta}\langle\chi^{i,\alpha},\chi^{j,\beta}\rangle_H,\quad \text{by Theorem 2}$$

$$=\delta_j\sum_\beta \eta_{j,\beta}(|H|/|C_{i,\alpha}|)\delta_{(i,\alpha)(j,\beta)}\quad \text{by (2.4)},$$

$$=(|G|/|C_i|)\delta_{ij},$$

using the numerical relationships noted prior to and during the proof of Theorem 1.

Permanence theorem: Let H be a subgroup of G, and let θ, ψ be characters of H, G, respectively. Then

$$I(\theta)\psi=I(\theta R(\psi)). \tag{3.3}$$

Proof: It suffices to check the relationship (3.3) on the set of lonely characters, so take $\theta=\chi^{i,\alpha}$ and $\psi=\chi^j$. Then

$$I(\chi^{i,\alpha})\chi^j=\chi^i\chi^j=(|G|/|C_i|)\delta_{ij}\chi^i,$$

using Theorem 1 and the multipication formula for lonely characters. Now consider $\chi^{i,\alpha}R(\chi^j)$. Evidently, if $i\neq j$, this vanishes on H, but for $j=i$, $R(\chi^i)$ is nonzero on the set $C_i\cap H$ containing $C_{i,\alpha}$, and assumes the value $|G|/|C_i|$ there. So $\chi^{i,\alpha}R(\chi^j)=(|G|/|C_i|)\delta_{ij}\chi^{i,\alpha}$, and the theorem follows on application of Theorem 1.

In a sense we can think of the above two theorems as being self-dual, for our proofs are essentially class-theoretic to contrast with the usual proofs (see Ref. 10) which are representation-theoretic.

In the next section we consider the application of lonely characters to Robinson's work (Refs. 4, 11) on the inverse restriction operation.

4. INVERSE RESTRICTION

In connection with the representation theory of the symmetric group, Robinson (Ref. 11) considers the validity of the equation $I(\chi\lambda)=I(\chi)\lambda'$, where χ, λ are characters of a subgroup H of G and λ' is a character of G such that $R(\lambda')=\lambda$. First we investigate the conditions under which $R(\lambda')=\lambda$ has at least one solution λ' for every character λ or equivalently for every lonely character. Choose $\lambda=\chi^{i,\alpha}$ and $\lambda'=\sum_j r_j\chi^j$, then the condition $R(\lambda')=\lambda$ becomes

$$\sum_{j,\beta}\delta_j r_j\eta_{j,\beta}\chi^{j,\beta}=\chi^{i,\alpha}. \tag{4.1}$$

Clearly r_j is arbitrary if $\delta_j=0$ and is zero if $\delta_j=1$ and $i\neq j$. We are left with the equation

$$r_i\sum_\beta \eta_{i,\beta}\chi^{i,\beta}=\chi^{i,\alpha},$$

which only makes sense if $r_i\eta_{i,\beta}=\delta_{\alpha\beta}$. But every $\eta_{i,\beta}$ is nonzero, so that (4.1) only has a solution if $C_i\cap H$ is a single class of H.

Before stating a theorem it is convenient to relabel the classes of G so that for $i=1,2,\cdots,r$ we have $C_i\cap H\neq\{0\}$ and for $i>r$, $C_i\cap H=\{0\}$. Also, if $C_i\cap H$ is a single class of H, then the label α used previously is redundant and for example we can replace $\eta_{i,\alpha}$ by η_i. Then we have

Theorem 3: Let λ run over the characters of the subgroup H of G. Then the equation $R(\lambda')=\lambda$ has a solution λ' which is a character of G, for every λ, if and only if no class of G contains more than one class of H (nonsplitting property). When this holds, we may write the solution for lonely characters as

$$R^{-1}(\chi_H^i)=\frac{1}{\eta_i}\chi_G^i+\sum_{k>r}\gamma_{ik}\chi_G^k, \tag{4.2}$$

for $i=1,2,\ldots,r$, where the complex numbers γ_{ik} are arbitrary. Furthermore, $I(\chi\lambda)=I(\chi)R^{-1}(\lambda)$ for all characters χ, λ of H if and only if the nonsplitting property holds.

Proof: This is trivial using lonely characters.

In stating the theorem and, in particular, in writing (4.2) we have committed an abuse of notation, since the operator R^{-1}, even when it exists, is not unique. In fact, the solutions of $R(\lambda')=\chi_H^i$ form a hyperplane passing through the principal solution $R^{-1}(\chi_H^i)=(1/\eta_i)\chi_G^i$. However, since the induced character $I(\chi)$ only takes nonzero values on the first r classes we could replace

$R^{-1}(\lambda)$ by the principal solution in the result $I(\chi\lambda) = I(\chi)R^{-1}(\lambda)$.

Although the nonsplitting property is, in general, a strong property to impose on a subgroup H of G, pairs of symmetric groups do possess it. Indeed for such groups the principal solutions for lonely characters are

$$R^{-1}(\chi^{(\mu)}_{S_{n-m}}) = (1/\eta_\mu)\chi^{(1^m\mu)}_{S_n},$$

where (μ) denotes a class of S_{n-m}. Also it is possible to apply R^{-1} by stages through a chain of subgroups. For the pair of groups (S_3, S_4) we find (non-principal) solutions for the simple characters $\chi^{[\mu]}$ as $R^{-1}(\chi^{[3]}) = \chi^{[4]}$, $R^{-1}(\chi^{[2,1]}) = \chi^{[2^2]}$, $R^{-1}(\chi^{[1^3]}) = \chi^{[1^4]}$, by inspection of the character tables. In general, however, inverse restriction does not yield an irreducible character from an irreducible character. We find from branching rules that if $[\mu] = [\mu_1, \mu_2, \ldots, \mu_t]$ labels a UIR of S_{n-1} then the hyperplane $R^{-1}(\chi^{[\mu]})$ contains an irreducible character of S_n if and only if $\mu_1 = \mu_2 = \cdots = \mu_{r-1} = \mu_r + 1$, for some $r \leq n$ and all other μ's are zero. Then the solution set contains $\chi^{[\mu_1, \mu_2, \ldots, \mu_{r-1}, \mu_r+1]}$.

We now consider some rather different ideas connected with induced representations.

5. PARTIAL INDUCTION

We begin by recalling the inducing construction for finite groups. Let D be a representation of the subgroup H of G and let its carrier space V have basis $\{\phi_r: r = 1, 2, \ldots, d\}$. Let $G = \cup_\sigma p_\sigma H$ be a left coset decomposition of G relative to H. Now, for each σ, let V_σ denote the vector space with basis $\{p_\sigma\phi_r: r = 1, 2, \ldots, d\}$, where p_σ is regarded as a label. Then the spaces V_σ are considered distinct and the induced representation $I(D)$ acts on $\oplus_\sigma V_\sigma$ in the following way: Suppose $p_\lambda^{-1} g p_\tau = h \in H$, for $g \in G$, then

$$g(p_\tau \phi_r) = p_\gamma(h\phi_r) = \sum_{t=1}^{d} (p_\gamma \phi_t) D(h)_{tr}. \tag{5.1}$$

Now we said that p_σ is merely a label to distinguish the space V_σ from other such spaces, but it may well happen, for example, if V is a space of wavefunctions or physical tensors, that each p_σ has a definite action on V within some larger space W—so that V_σ already has a meaning within W. When this happens, we clearly obtain a representation of G by letting G generate from V an invariant subspace of W. This subspace is the linear span of the vectors $\{p_\sigma\phi_r\}$, which are now not necessarily independent. Indeed one might make precisely this construction if one misunderstood the induction procedure, for in the latter case one must consider as distinct the spaces V_σ and V even when they appear identical. We show now how these ideas lead to the notion of partial induction with a consequent partial reduction of $I(D)$.

Define $A = \{a \in G: V_a = V\}$, then A is clearly a subgroup of G containing H. Let $A = \cup_{i=1}^t a_i H$ be a left coset decomposition of A relative to H. For each i, $i = 1, 2, \ldots, t$, we have $V_{a_i} = V$ and so there exists a matrix A_i such that

$$a_i \phi_r = \sum_{s=1}^{d} \phi_s (A_i)_{sr}. \tag{5.2}$$

Let $\langle\phi|$ denote the row vector $(\phi_1, \phi_2, \ldots, \phi_d)$ then (5.2) becomes

$$a_i\langle\phi| = \langle\phi|A_i. \tag{5.3}$$

Hence if we only consider the group action, $a_i\langle\phi|A_i^{-1} = \langle\phi|$ is independent of the particular choice of i. This would not be true for the action of the induced representation.

Now consider the above from the point of view of the inducing construction where only H is allowed to act on the ϕ's and the spaces labelled by different coset representatives are distinct. Define $\langle\phi|^i = a_i\langle\phi|A_i^{-1}$, then if $a_i a_j = a_m h$, where $h \in H$, we have

$$a_i\langle\phi|^j = a_m h\langle\phi|A_j^{-1} = \langle\phi|^m A_m D(h) A_j^{-1}. \tag{5.4}$$

For the group action, since $\langle\phi|^j = \langle\phi|^m = \langle\phi|$, we could have written simply

$$a_i\langle\phi|^j = \langle\phi|^m A_i; \tag{5.5}$$

hence comparing (5.4), (5.5), we see that $a_i a_j = a_m h$ implies $A_i A_j = A_m D(h)$. This is not the quickest derivation, but it brings out the differences between the two actions.

Keeping the spaces $\langle\phi|^1, \langle\phi|^2, \ldots, \langle\phi|^t$ distinct, we now show that they form a basis for a representation of the symmetric group S_t. Let $b \in A$; then, if $ba_i \in a_j H$,

$$b\langle\phi|^i = \langle\phi|^j B(b) \tag{5.6}$$

and in particular if $b = a_m h$, where $h \in H$, we have $B(b)$ is equal to $A_m D(h)$ and is independent of i, j. The action of A on the cosets of H in A leads to a permutation representation which, as is well known, decomposes into the representation $[t] \oplus [t-1, 1]$ of S_t. Here we associate $b \in A$ with $\Pi_b \in S_t$ if $ba_i \in a_{\Pi(b)i}H$. Hence, if $I^A(D)$ denotes D induced up to A, then $I^A(D)$ decomposes naturally as

$$b \to B(b) \oplus [t-1, 1](\Pi_b) \otimes B(b) \tag{5.7}$$

for $b \in B$, where $b \to B(b)$ is the extension Δ of D to A. Thus $\Delta(b) = B(b) = A_m D(h)$ for $b = a_m h$. If D is irreducible, then so is Δ. Now consider $I^G(\Delta)$, which is Δ induced from A to G, and which we call the partially induced representation. From (5.7) and the well-known result on inducing by stages through intermediate subgroups we see that $I^G(\Delta)$ is a subrepresentation of $I^G(D)$.

It should be noted that $I^G(\Delta)$ is only identical to the representation obtained by allowing G to act on V within W if the vector spaces V_σ satisfy the condition: Either $V_\sigma = V_\tau$ or $V_\sigma \cap V_\tau = \{0\}$ for all σ, τ. In particular this will be true if H is a normal subgroup of G and D is a UIR of H, because then V and $p_\sigma V$ are both irreducible H-modules and hence their intersection is either V or $\{0\}$.

Another point is that in general the set $\{\Pi_b : b \in A\}$ will only give a subgroup of S_t and so the representation $[t-1, 1] \otimes \Delta$ will split further. For example, if $A = H \circledS P$, the coset representatives can be chosen to form the group P. Then $I^A(D)$ becomes $\text{reg}P \otimes \Delta$, where $\text{reg}P$ denotes the regular representation of P, which can be reduced into irreducibles.

This result can be used to obtain simply a result of Robinson (Refs. 12—14) for the symmetric groups. Take

$G=S_{mn}$ and $H=S_m\times S_m\times\cdots\times S_m$ (n factors). Let $[\alpha]$ be a representation of S_m; then we consider the reduction of $I([\alpha]\otimes\cdots\otimes[\alpha])$ arising from the interchangeability of the n factors. In this case $A=H\circledS S_n^*$, where S_n^* is isomorphic to S_n. If $\Pi\in S_n$ belongs to the class $(1^{a_1}2^{a_2}\cdots n^{a_n})$, then $\Pi^*\in S_n^*$ belongs to the class $(1^{ma_1}2^{ma_2}\cdots n^{ma_n})$. Let $\{\psi_i: i=1\cdots f\}$ be a basis for the UIR $[\alpha]$ of S_m; then a basis for $[\alpha]^n=[\alpha]\otimes\cdots\otimes[\alpha]$ (n factors) is the set of ordered n-tuples $\{(\psi_{i_1},\psi_{i_2},\ldots,\psi_{i_n})\ i_s=1,\ldots,f;\ s=1,\ldots,n\}$. The action of S_n^* on this basis is

$$\Pi^*(\psi_{i_1},\ldots,\psi_{i_n})=(\psi_{i_{\Pi^{-1}(1)}},\ldots,\psi_{i_{\Pi^{-1}(n)}})$$

$$=\sum_j(\psi_{j_1},\ldots,\psi_{j_n})A(\Pi^*)_{ji}. \qquad (5.8)$$

From the previous discussion

$$I^A([\alpha]^n)=\bigoplus_\beta f_\beta(\Delta\otimes[\beta]), \qquad (5.9)$$

where $[\beta]$ is a UIR of S_n of dimension f_β and Δ is defined by

$$\Delta(\Pi^*h)=A(\Pi^*)([\alpha]^n(h)). \qquad (5.10)$$

Robinson denoted $I^{S_{mn}}(\Delta\otimes[\beta])$ by $[\alpha]\odot[\beta]$. It is called the symmetrized outer product. Thus

$$I^{S_{mn}}([\alpha]^n)=\bigoplus_\beta f_\beta[\alpha]\odot[\beta]. \qquad (5.11)$$

6. LITTLE GROUP THEORY

The theory of the above section also allows us to rederive an important aspect of little group theory in a very straightforward manner when the little group can be expressed as a semidirect product, for example, for symmorphic space groups. Let G, H, D be as before. Define the little group $N=\{g\in G: D(g^{-1}hg)=D(h)\ \forall h\in H\}$. Suppose for simplicity that $N=H\circledS P$; then

$$D(p_i^{-1}hp_i)=P_i^{-1}D(h)P_i, \qquad (6.1)$$

for all $h\in H$, $p_i\in P$. Hence $p_i\in P$. Hence $p_i\to P_i$ is a projective representation of P will factor system w say.

Now

$$hp_i\langle\phi| = p_i\langle\phi|D(p_i^{-1}hp_i),$$
$$=p_i\langle\phi|P_i^{-1}D(h)P_i,$$

by (6.1), so that

$$h(p_i\langle\phi|P_i^{-1})=(p_i\langle\phi|P_i^{-1})D(h), \qquad (6.2)$$

$$p_j(p_i\langle\phi|P_i^{-1})=p_jp_i\langle\phi|P_i^{-1}P_j^{-1}p_j$$
$$=(p_k\langle\phi|P_k^{-1})w^*(j,i)Pj, \qquad (6.3)$$

where $p_jp_i=p_k$. Let $\Delta=I^N(D)$; then

$$\Delta(p_jh)=P_jD(h)\otimes \text{regular } w^*\text{-rep of } P.$$

On reducing the regular w^*-rep of P to irreducibles, we see that we have captured an important property of the little group, namely that it allows for a natural reduction of $I^G(D)$.

ACKNOWLEDGMENTS

The authors wish to thank Drs. C. J. Bradley, R. C. King, and B. L. Davies for posing certain questions whose answers led to this paper. One of us, P. G., wishes to thank New Hall, Cambridge for the award of a Research Fellowship.

[1]A. Gamba, J. Math. Phys. 9, 186 (1968).
[2]G. de B. Robinson, J. Algebra 20, 118 (1972).
[3]G. de B.Robinson, J. Math. Phys. 12, 2212 (1971).
[4]G. de B. Robinson, Can. J. Math. XXV, 1051 (1973).
[5]E. de Vries and A. J. van Zanten, Physica 49, 532 (1970).
[6]A. J. van Zanten and E. de Vries, J. Math. Phys. 14, 1423 (1973).
[7]N. B. Backhouse, J. Math. Phys. 16, 443 (1975).
[8]C. W. Curtis and I. Reiner, *Representation Theory of Finite Groups and Associative Algebras* (Interscience, New York, 1966).
[9]N. B. Backhouse, Physica 70, 505 (1973).
[10]C. J. Bradley, J. Math. Phys. 7, 1145 (1966).
[11]G. de B.Robinson, J. Lond. Math. Soc., to be published).
[12]G. de B. Robinson, Can. J. Math. 1, 166 (1949).
[13]G. de B. Robinson, Can. J. Math. 2, 334 (1950).
[14]A. Kerber, Proc. 2nd Int. Colloq. on Group Theoretical Methods in Physics (Nijmegen, 1973) B146.

Markovian subdynamics in quantum dynamical systems

Etang Chen*

University of Rochester, Rochester, New York 14627
and Harvard University, Cambridge, Massachusetts 02138
(Received 2 April 1976)

Subdynamical systems induced from a given quantum dynamical system are studied in the framework of operator algebras. Sufficient conditions are shown for induced subdynamics to be Markovian. It is also proved that the ergodicity of states in Markovian subsystems is preserved.

1. INTRODUCTION

Given a quantum system Σ described by some dynamics σ. Let Γ be a subsystem of Σ, we are concerned about the subdynamics γ of Γ induced from σ. There are, in general, two different types of subdynamics: reversible and irreversible. An interesting problem in the subdynamics is how to obtain an irreversible process γ from a given reversible process σ in the whole system. The conventional method employed in this problem is by means of the so-called "projection technique," i. e., by using an appropriate projection from Σ onto Γ so that γ is exactly the projected map of σ. For example, σ is described by Schrödinger equation in Σ, and, by projection technique, one can obtain in subsystem Γ an irreversible subdynamics γ, which is controlled by the so-called master equation.[1] Another aspect of interest in the subdynamics is the extraction of macroscopic subsystems from a quantum mechanical system such as developed in the theory of "independent subdynamics."[2]

In the present paper, we study the subdynamics in the framework of operator algebras, in particular, Markovian subdynamics induced from a reversible dynamical system. Some sufficient conditions are given so that induced subdynamics is Markovian. Then, the preservation of ergodicity is shown in Markovian subdynamical systems.

We begin with reversible subdynamics in the next section. The main purpose of this section is to show the connection between some well-known results in operator algebras and subdynamical systems. In Sec. 3, we study Markovian subdynamics. Sufficient conditions for a subsystem to be Markovian are shown (Theorem 3.1 and Proposition 3.5). In the final section, Sec. 4, the preservation of the ergodicity of states in a Markovian subdynamical system is proved (Proposition 4.1).

2. SUBDYNAMICAL SYSTEMS

A (quantum) *dynamical system* is a pair $(M, \alpha(\mathbb{R}))$ consisting of a quantum system M and dynamics $\alpha(\mathbb{R})$ of the system. Here, M is a von Nuemann algebra acting on a separable Hilbert space $\mathcal{H}$, and $\alpha(\mathbb{R})$ is an one-parameter group of automorphisms $\{\alpha_t; t\in\mathbb{R}\}$ of M. Let N be a von Neumann subalgebra of M, which represents a subsystem in the quantum system M. We are interested in the subdynamics of N induced from $\alpha(\mathbb{R})$ in the following way: Let ϵ_0 be a projection of norm one from M onto N, and $\gamma_t=\epsilon_0\alpha_t\epsilon_0$; $\{\gamma_t; t\in\mathbb{R}\}$ and $\{\gamma_t; t\geq 0\}$ are denoted by $\gamma(\mathbb{R})$ and $\gamma(\mathbb{R}^+)$ respectively; then we want to know whether $\gamma(\mathbb{R})$ and $\gamma(\mathbb{R}^+)$ would be dynamics of the subsystem N. If they are dynamics in N, then the subdynamics described by $\gamma(\mathbb{R})$ (resp. $\gamma(\mathbb{R}^+)$) is a reversible (resp. irreversible) process. More precisely, the pair $(N, \gamma(\mathbb{R}))$ is a (*reversible*) *subdynamical system*, if $\gamma(\mathbb{R})$ is an one-parameter group of automorphisms of N; and $(N, \gamma(\mathbb{R}^+))$ is a *Markovian subdynamical system*, if $\gamma(\mathbb{R}^+)$ is an one-parameter contraction semigroup of N.

We shall give some characterizations for $\gamma(\mathbb{R})$ to be reversible subdynamics in this section, and Markovian subdynamical systems will be studied extensively in the following sections.

Notice that one crucial point for $\gamma(\mathbb{R})$ to be dynamic in a subsystem N is $\gamma_t\circ\gamma_s=\gamma_{t+s}$ for $t,\ s\in\mathbb{R}$. This can be achieved by, for instance, $\epsilon_0\circ\alpha_t=\alpha_t\circ\epsilon_0$ for all $t\in\mathbb{R}$. In this section, we shall give necessary and sufficient conditions for the commutativity of ϵ_0 and α_t.

First, we recall a projection of norm one ϵ_0 from M onto N has the following properties[3]: (i) $\epsilon_0(x)=x$ for $x\in N$, (ii) $\|\epsilon_0(x)\|\leq\|x\|$ for $x\in M$, (iii) $\epsilon_0(x^*)=\epsilon_0(x)^*$ for $x\in M$, (iv) $\epsilon_0(x^*x)\geq 0$ for $x\in M$, (v) $\epsilon_0(axb)=a\,\epsilon_0(x)b$ for $a, b\in N$, $x\in M$, (vi) $\epsilon_0(x)^*\epsilon_0(x)\leq\epsilon_0(x^*x)$ for $x\in M$. ϵ_0 is a faithful normal projection of norm one if, in addition, (vii) $\epsilon_0(x^*x)=0$ implies $x=0$ for $x\in M$, and (viii) $\sup_\alpha\epsilon_0(x_\alpha)=\epsilon_0(\sup_\alpha x_\alpha)$ for each uniformly bounded directed set $\{x_\alpha\}$ of positive elements of M.

Remark 2.1: We note that a projection ϵ_0 of norm one from M onto N is necessarily positive and 2-side N-module mapping [i. e., (iv) and (v)]; and conversely, a projection ϵ_0 from M onto N is norm one if it is positive.[3]

The existence of such a projection of norm one is ensured from a theorem of Takesaki[4,5]: Let N be a subalgebra of a von Neumann algebra M, and φ a faithful normal state on M. Then, there is a faithful normal projection of norm one ϵ_0 from M onto N such that $\varphi\circ\epsilon_0=\varphi$ if and only if $\sigma_t^\varphi(N)=N$ for all $t\in\mathbb{R}$, where $\{\sigma_t^\varphi; t\in\mathbb{R}\}=\sigma^\varphi(\mathbb{R})$ is the modular automorphism of M characterized by the following conditions: for $x, y\in M$ there is an analytic function f on the strip $D=\{z\in\mathbb{C};\ \mathrm{Im}z\in(0,1)\}$, continuous on $\bar{D}$ such that $|f|$ is bounded and $f(t)=\varphi(\sigma_t^\varphi(x)y)$, $f(t+i)=\varphi(y\sigma_t^\varphi(x))$. ϵ_0 obtained in this way is a conditional expectation induced by φ.

In the sequal, a conditional expectation of M onto N will always mean a faithful normal projection of norm one induced by some faithful normal state on M.

If we consider dynamical system $(M,\sigma^\varphi(\mathbb{R}))$, and $\gamma_t = \epsilon_0\sigma_t^\varphi\epsilon_0$ with ϵ_0 a conditional expectation of M onto N, then $(N,\gamma(\mathbb{R}))$ is a subdynamical system. This is due to the fact that $\gamma(\mathbb{R})$ is a group of automorphisms of N. In fact, we have the following:

Proposition 2.2: For all $t\in\mathbb{R}$, $\sigma_t^\varphi\circ\epsilon_0=\epsilon_0\circ\sigma_t^\varphi \iff \sigma_t^\varphi(N)=N$.

Therefore, there is always a subdynamical system induced from the system $(M,\sigma^\varphi(\mathbb{R}))$ whenever there exists a conditional expectation of M onto N.

The proof "$\iff$" is essentially in Ref. 6. By assumption, $\sigma_t^\varphi(N)=N$, there is a conditional expectation ϵ_0 of M onto N. Let us consider $\sigma_{-t}^\varphi\circ\epsilon_0\circ\sigma_t^\varphi(x)=\sigma_{-t}^\varphi(\epsilon_0[\sigma_t^\varphi(x)] \in\sigma_{-t}^\varphi(N)$ for $x\in M$, then $\sigma_{-t}^\varphi\circ\epsilon_0\circ\sigma_t^\varphi$ is a faithful normal projection of norm one from M onto $\sigma_{-t}^\varphi(N)$. Moreover, $(\varphi\circ\sigma_t^\varphi)\circ(\sigma_{-t}^\varphi\circ\epsilon_0\circ\sigma_t^\varphi)=\varphi\circ\sigma_t^\varphi$ for all $t\in\mathbb{R}$. Hence, $\sigma_{-t}^\varphi\circ\epsilon_0\circ\sigma_t^\varphi$ is a conditional expectation of M onto $\sigma_{-t}^\varphi(N)$ induced by a faithful normal state $\varphi\circ\sigma_t^\varphi$. The modular automorphism is indeed $\sigma_t^{\varphi\circ\sigma_t^\varphi}=\sigma_t^\varphi$. Therefore, the corresponding conditional expectation is $\sigma_{-t}^\varphi\circ\epsilon_0\circ\sigma_t^\varphi=\epsilon_0$, which implies $\epsilon_0\circ\sigma_t^\varphi=\sigma_t^\varphi\circ\epsilon_0$. Conversely, we have $\sigma_t^\varphi(\epsilon_0(M)) =\epsilon_0(\sigma_t^\varphi(M))$ from the hypothesis, and note that $\sigma_t^\varphi(M)=M$. Hence, we have $\sigma_t^\varphi(N)=N$.

If we consider a subsystem N of macroscopic observables (e.g., in statistical mechanics), then the dynamics described by the modular automorphism $\sigma^\varphi(\mathbb{R})$ is not very interesting. In fact, the subdynamical system $(N,\gamma(\mathbb{R}))$ is trivial in the sense that $\gamma(\mathbb{R})$ is the identity automorphism of N. This can be seen easily from the following.

Proposition 2.3: N is Abelian, then $N\subseteq M_\varphi$, where $M_\varphi=\{x\in M;\varphi(xy)=\varphi(yx)$ for all $y\in M\}$.

Indeed, for $x\in N$, $y\in M$, $\varphi(xy)=\varphi(\epsilon_0(xy))=\varphi(x\epsilon_0(y)) =\varphi(\epsilon_0(y)x)=\varphi(\epsilon_0(yx))=\varphi(yx)$. It is well known that M_φ is a fixed point subalgebra of M,[5] i.e., $M_\varphi=\{x\in M; \sigma_t^\varphi(x)=x\}$, so that $\gamma_t=\epsilon_0\sigma_t^\varphi\epsilon_0$ is the identity automorphism. Therefore, a nontrivial subdynamical system of macroscopic observables should be induced from another automorphism $\alpha(\mathbb{R})$ different from the modular automorphism $\sigma^\varphi(\mathbb{R})$.

Given a dynamical system $(M,\alpha(\mathbb{R}))$ for arbitrary dynamics $\alpha(\mathbb{R})$ different from the modular automorphism, there is also a criterion similar to Proposition 2.2:

Proposition 2.4: Let ϵ_0 be a faithful normal projection of norm one from M onto N. (i) If $\alpha_t\epsilon_0=\epsilon_0\alpha_t$ for all $t\in\mathbb{R}$, then $\alpha_t(N)=N$. (ii) If $\alpha_t(N)=N$ for all $t\in\mathbb{R}$ and $N\supseteq N'\cap M$, then $\alpha_t\epsilon_0=\epsilon_0\alpha_t$.

(i) was proved in Proposition 2.2. The proof of (ii) follows again from arguments used by Connes.[6] Let ϵ': $M\to N$ be a map defined by $\epsilon'(x)=\alpha_{-t}\{\epsilon_0[\alpha_t(x)]\}$ for $x\in M$; then ϵ' is a linear, positive, normal surjective map.

Moreover, for $x,y\in M$,

$$\begin{aligned}\epsilon'(\epsilon'(x)y\epsilon'(x)) &=\alpha_{-t}\{\epsilon_0[\alpha_t(\epsilon'(x)y\epsilon'(x))]\}\\ &=\alpha_{-t}\{\epsilon_0[\alpha_t(\alpha_{-t}[\epsilon_0\alpha_t(x)]y\,\alpha_{-t}[\epsilon_0\alpha_t(x)])\}\\ &=\alpha_{-t}\{\epsilon_0[\epsilon_0\alpha_t(x)\cdot\alpha_t(y)\cdot\epsilon_0\alpha_t(x)]\}\\ &=\alpha_{-t}\{\epsilon_0\alpha_t(x)\cdot\epsilon_0\alpha_t(y)\cdot\epsilon_0\alpha_t(x)\}\\ &=\alpha_{-t}\{\epsilon_0\alpha_t(x)\}\alpha_{-t}\{\epsilon_0\alpha_t(y)\}\alpha_{-t}\{\epsilon_0\alpha_t(x)\}\\ &=\epsilon'(x)\epsilon'(y)\epsilon'(x).\end{aligned}$$

This means that ϵ' is a 2-side N-module mapping; hence ϵ' is a faithful normal projection of norm one from M onto N (see Remark 2.1). However, due to the uniqueness of faithful normal projection of norm one [Ref. 6, Theorem 1.5.5(a)], we have $\epsilon'=\epsilon_0$. Consequently, $\epsilon_0\alpha_t =\alpha_t\epsilon_0$ for all $t\in\mathbb{R}$.

Remark 2.5: The linearity, positivity and normality of ϵ' follows immediately from its definition; however, the surjectivity of ϵ' is due to the hypothesis of $\alpha_t(N)=N$ *for all* $t\in\mathbb{R}$. Actually, this is one of the main reasons which makes the subdynamics a reversible process.

Remark 2.6: We note that if N is a maximal Abelian subalgebra of M, then the condition $N\supseteq N'\cap M$ holds.

3. MARKOVIAN SUBDYNAMICS

Given a dynamical system $(M,\alpha(\mathbb{R}))$, and a conditional expectation ϵ_0 of M onto N induced by a faithful normal state φ on M, we want to see in this section when is the subsystem $(N,\gamma(\mathbb{R}^+))$ Markovian.

Theorem 3.1: Suppose φ is $\alpha(\mathbb{R})$-invariant, then $(N,\gamma(\mathbb{R}^+))$ is a Markovian subdynamical system if $\alpha_t(N)\supseteq N$ for all $t\geq 0$.

Proof of Theorem 3.1: Let π_φ, $\mathcal{H}_\varphi$, ξ_φ be the cyclic representation of M induced by φ. As φ is faithful, π_φ is also faithful; and φ is normal, hence $\pi_\varphi(M)''=\pi_\varphi(M)$. Thus, π_φ is an isomorphism of M onto $\pi_\varphi(M)$, and we may consider $M=\pi_\varphi(M)$ and $\mathcal{H}=\mathcal{H}_\varphi$. Let $\overline{N\xi_\varphi}=\mathcal{K}$ and let E_0 be a projection from $\mathcal{H}$ onto $\mathcal{K}$. Then, for $x\in M$ and $y\in N$, we have[5]

$$\langle E_0x\xi_\varphi, y\xi_\varphi\rangle=\varphi(y^*x)=\varphi(\epsilon_0(y^*x))=\varphi(y^*\epsilon_0(x)) =\langle\epsilon_0(x)\xi_\varphi,y\xi_\varphi\rangle;$$

hence, $E_0x\xi_\varphi=\epsilon_0(x)\xi_\varphi$ for all $x\in M$. Thus,

$$E_0(M)\xi_\varphi=\epsilon_0(M)\xi_\varphi=N\xi_\varphi. \quad (1)$$

By assumption, φ is $\alpha(\mathbb{R})$-invariant, $\alpha(\mathbb{R})$ is unitarily implementable in $\mathcal{H}_\varphi=\mathcal{H}$, i.e., $\alpha_t(x)=u_txu_t^*$ and $u_t\xi_\varphi=\xi_\varphi$ for $x\in M$. Define

$$E_t=u_tE_0u_t^* \quad \text{for } t\geq 0 \quad (2)$$

then E_t is a projection on $\mathcal{H}$. By hypothesis, $\alpha_t(N)\supseteq N$ for $t\geq 0$, we have $\alpha_t(N)\xi_\varphi\supseteq N\xi_\varphi$; hence $u_tNu_t^*\xi_\varphi=u_tN\xi_\varphi \supseteq N\xi_\varphi$ for $t\geq 0$. From (1) and (2),

$$E_t(M)\xi_\varphi=u_tE_0u_t^*(M)\xi_\varphi=u_tE_0(M)\xi_\varphi =u_tN\xi_\varphi\supseteq N\xi_\varphi=E_0(M)\xi_\varphi$$

for $t\geq 0$; this implies

$$E_t\geq E_0 \quad \text{for } t\geq 0. \quad (3)$$

By the definition of γ_t, we have for all $x\in N$ and $t\geq 0$,

$$\gamma_t(x)\xi_\varphi=\epsilon_0\alpha_t\epsilon_0(x)\xi_\varphi=\epsilon_0\alpha_t(E_0x)\xi_\varphi=\epsilon_0[u_tE_0(x)u_t^*]\xi_\varphi$$
$$=E_0u_tE_0(x)u_t^*\xi_\varphi=E_0u_tE_0(x)\xi_\varphi=E_0u_tx\xi_\varphi. \quad (4)$$

For $s,t\geq 0$ and $x\in N$, we have

$$\begin{aligned}\gamma_s\gamma_t(x)\xi_\varphi&=\gamma_s(\gamma_t(x))\xi_\varphi=E_0u_s(\gamma_t(x))\xi_\varphi\\&=E_0u_sE_0u_tx\xi_\varphi\\&=E_0u_sE_0u_s^*u_su_tx\xi_\varphi\\&=E_0E_su_{s+t}x\xi_\varphi &&[\text{by (2)}]\\&=E_0u_{s+t}x\xi_\varphi &&[\text{by (3)}]\\&=\gamma_{s+t}(x)\xi_\varphi &&[\text{by (4)}].\end{aligned}$$

As ξ_φ is separating for N, we have $\gamma_s\gamma_t(x)=\gamma_{s+t}(x)$ for $x\in N$; hence $\gamma_s\gamma_t=\gamma_{s+t}$ for $s,t\geq 0$. Clearly $\gamma_0=1$, and $\|\gamma_t\|=\|\epsilon_0\alpha_t\epsilon_0\|\leq\|\alpha_t\|\leq 1$; hence γ_t is contraction. Moreover, γ_t is a surjective map of N onto N: Indeed, $\gamma_t(N)=\epsilon_0\alpha_t(N)\supseteq N$ for $t\geq 0$ by assumption, and, on the other hand, $x\in\gamma_t(N)$ for $t\geq 0$, there is a $y\in N$ such that $x=\gamma_t(y)$ for some $t\geq 0$; it follows then $x=\gamma_t(y)=\epsilon_0\alpha_t(y)\in N$, so that $\gamma_t(N)\subseteq N$ for $t\geq 0$. This completes the proof.

Remark 3.2: Compare the condition in Theorem 3.1 with Remark 2.5, we see that this is an *a priori* condition to yield an irreversible process in a subsystem.

Remark 3.3: If $t\to\alpha_t(x)$ is continuous for $x\in M$, then $t\to\gamma_t(x)=\epsilon_0\alpha_t\epsilon_0(x)$ is also continuous for $x\in N$. Therefore, $\gamma(\mathbb{R}^+)$ is a strongly continuous contraction semigroup on N if $\alpha(\mathbb{R})$ is strongly continuous.

Remark 3.4: Notice that $\gamma(\mathbb{R}^+)$ is in fact a semigroup of norm one;

$$\|\gamma_t\|=\sup_{\substack{\|x\|=1\\x\in N}}\|\gamma_t(x)\|=\sup_{\substack{\|x\|=1\\x\in N}}\|\epsilon_0\alpha_t(x)\|$$
$$\geq\|\epsilon_0\alpha_t(x)\|=\|\alpha_t(x)\|=\|x\|=1.$$

In connection with Theorem 3.1, we give other sufficient conditions which yield Markovian subdynamics. Let $N_c=N'\cap M$, and $N_z=N\cap N'$. If ϵ_0 is a conditional expectation of M onto N induced by a faithful normal state φ of M, then there is a conditional expectation ϵ_c (resp. ϵ_z) of M onto N_c (resp. N_z) by Takesaki's theorem cited in Sec. 2. In fact, for all $t\in\mathbb{R}$, $\sigma_t^\varphi(N_c)=N_c$ and $\sigma_t^\varphi(N_z)=N_z$, respectively; this can be seen as follows: Let $x\in N_c$; then $\sigma_t^\varphi(x)\in M$. Furthermore, for each $y\in N$, there is a $w\in N$ such that $\sigma_t^\varphi(w)=y$, since $\sigma_t^\varphi(N)=N$ for all $t\in\mathbb{R}$. Hence, $\sigma_t^\varphi(x)y=\sigma_t^\varphi(x)\sigma_t^\varphi(w)=\sigma_t^\varphi(xw)=\sigma_t^\varphi(wx)=\sigma_t^\varphi(w)\sigma_t^\varphi(x)=y\sigma_t^\varphi(x)$, which means $\sigma_t^\varphi(x)\in N'$. Therefore $\sigma_t^\varphi(x)\in N'\cap M=N_c$. Similarly, one can show N_z is also σ_t^φ-invariant for all $t\in\mathbb{R}$.

Proposition 3.5: Let $\gamma_t=\epsilon_c\alpha_t\epsilon_0$ for $t\geq 0$. Suppose φ is $\alpha(\mathbb{R})$-invariant and $N\supseteq N_c$, then $(N_c,\gamma(\mathbb{R}^+))$ is a Markovian subdynamical system, if (i) $\epsilon_c\epsilon_0=\epsilon_z\epsilon_0$ and (ii) $\alpha_t(N_z)\supseteq N_z$ for $t\geq 0$.

Proof: We use the same notations in the proof of Theorem 3.1, we may define

$$E_0(M)\xi_\varphi=\epsilon_0(M)\xi_\varphi=N\xi_\varphi,$$
$$E_c(M)\xi_\varphi=\epsilon_c(M)\xi_\varphi=N_c\xi_\varphi,$$
$$E_z(M)\xi_\varphi=\epsilon_z(M)\xi_\varphi=N_z\xi_\varphi.$$

Let $E_{tc}=u_tE_cu_t^*$ for $t\geq 0$, which is a projection on $\mathcal{H}$. And notice that $\alpha_t(N_c)\supseteq\alpha_t(N_z)\supseteq N_z$ for $t\geq 0$ [by (ii)], $E_{tc}(M)\xi_\varphi\supseteq E_z(M)\xi_\varphi$, which means

$$E_{tc}\geq E_z \quad \text{for } t\geq 0. \quad (5)$$

From (i), we have for all $x\in M$

$$E_cE_0x\xi_\varphi=E_zE_0x\xi_\varphi. \quad (6)$$

Moreover, $N\supseteq N_c$ implies

$$E_0\geq E_c. \quad (7)$$

For $x\in M$ and $t\geq 0$, similar to Theorem 3.1, we may compute

$$\gamma_t(x)\xi_\varphi=\epsilon_c\alpha_t\epsilon_0(x)\xi_\varphi$$
$$=E_cu_tE_0x\xi_\varphi. \quad (8)$$

Therefore, for $s,t\geq 0$ and $x\in M$,

$$\begin{aligned}\gamma_s\gamma_t(x)\xi_\varphi&=E_cu_sE_0(\gamma_t(x))\xi_\varphi\\&=E_cu_sE_0E_cu_tE_0x\xi_\varphi\\&=E_cu_sE_cu_tE_0x\xi_\varphi &&[\text{by (7)}]\\&=E_cE_{sc}u_{s+t}E_0x\xi_\varphi\\&=E_cE_0(E_{sc}u_{s+t}E_0x)\xi_\varphi &&[\text{by (7)}]\\&=E_zE_0(E_{sc}u_{s+t}E_0x)\xi_\varphi &&[\text{by (6)}]\\&=E_0E_z(E_{sc}u_{s+t}E_0x)\xi_\varphi &&[E_0\geq E_z]\\&=E_0E_zu_{s+t}E_0x\xi_\varphi &&[\text{by (5)}]\\&=E_zE_0u_{s+t}E_0x\xi_\varphi\\&=E_cE_0u_{s+t}E_0x\xi_\varphi &&[\text{by (6)}]\\&=E_cu_{s+t}E_0x\xi_\varphi &&[\text{by (7)}]\\&=\gamma_{s+t}(x)\xi_\varphi &&[\text{by (8)}].\end{aligned}$$

By a similar argument used in Theorem 3.1, one can show that $\gamma(\mathbb{R}^+)$ is a semigroup of contraction. In fact, Remark 3.4 still holds; hence $\gamma(\mathbb{R}^+)$ is a semigroup of norm one.

It deserves some remarks on the assumptions given in this proposition.

Remark 3.6: First of all, from Theorem 3.1, (ii) implies that the subsystem $(N_z,\gamma(\mathbb{R}^+))$, where $\gamma_t=\epsilon_z\alpha_t\epsilon_z$ for $t\geq 0$, is Markovian. Therefore, the above proposition shows subsystem N_c can be Markovian, if its subsystem N_z is already Markovian.

Remark 3.7: (i) is a special version of Nelson's Markoff property in the theory of Markoff fields.[7] Indeed, given a topological space X, let $\{\mathcal{O}\}$ be a family of closed subsets of X containing X, and the boundary (resp. the complement) of $\mathcal{O}$ is denoted by $\partial\mathcal{O}$ (resp. $\mathcal{O}^c$).

For each $\mathcal{O}\in\{\mathcal{O}\}$, there is a von Neumann algebra $\mathcal{R}(\mathcal{O})$, and for $\mathcal{O}_1,\mathcal{O}_2\in\{\mathcal{O}\}$ with $\mathcal{O}_1\subset\mathcal{O}_2$, $\mathcal{R}(\mathcal{O}_1)\subset\mathcal{R}(\mathcal{O}_2)$. In particular, for two $\mathcal{O}_1,\mathcal{O}_2\in\{\mathcal{O}\}$ with $\mathcal{O}_1\subset\mathcal{O}_2$, we assume $\mathcal{R}(\mathcal{O}_1^c)=\mathcal{R}(\mathcal{O}_1)'\cap\mathcal{R}(\mathcal{O}_2)$ and $\mathcal{R}(\mathcal{O}_1)\cap\mathcal{R}(\mathcal{O}_2)=\mathcal{R}(\mathcal{O}_1\cap\mathcal{O}_2)$. Then $\mathcal{R}(\partial\mathcal{O})=\mathcal{R}(\mathcal{O}_1)\cap\mathcal{R}(\mathcal{O}_1^c)=\mathcal{R}(\mathcal{O}_1)'\cap\mathcal{R}(\mathcal{O}_1)$. If there exists a conditional expectation ϵ_0 of $\mathcal{R}(\mathcal{O}_2)$ onto $\mathcal{R}(\mathcal{O}_1)$, then as we have seen that there are also conditional expectations ϵ_z (resp. ϵ_c) of $\mathcal{R}(\mathcal{O}_2)$ onto $\mathcal{R}(\partial\mathcal{O})$ (resp. onto $\mathcal{R}(\mathcal{O}_1^c)$). Hence, (i) is a Nelson's Markoff property; $\epsilon_c\epsilon_0=\epsilon_z\epsilon_0$. For more details of this property in this setting, we refer to Accardi's paper.[8]

Remark 3.8: We note that if N is maximal Abelian, then $N=N_c$, and thus $N=N_c=N_z$. Hence, Proposition 3.5 reduces to Theorem 3.1.

4. PRESERVATION OF ERGODICITY

In addition to the time evolution, we consider in this section another group G of physical symmetry acting on this Markovian system $(N,\gamma(\mathbb{R}^+))$ induced from a dynamical system $(M,\alpha(\mathbb{R}))$. Let us assume that G is also represented by an automorphism group of N; $\tau: G \to \mathrm{Aut}(N)$. If ω is a G-invariant state on N, and π_ω, $\mathcal{H}_\omega$ is the cyclic representation of N induced by ω, then there is a unitary representation u_ω of G on $\mathcal{H}_\omega$ such that $\pi_\omega(\tau_g x)=u_\omega(g)\pi_\omega(x)u_\omega(g)^*$ for $x\in N$ and $g\in G$. A G-invariant state ω is *G-ergodic* if $\pi_\omega(N)'\cap\mathcal{U}_\omega(G)'=\{\lambda 1_\omega\}$.[9]

In this section, we assume that $\alpha(\mathbb{R})$ is unitarily implementable on subsystem N, i.e., $\alpha_t(x)=u_t x u_t^*$ for $x\in N$ (cf. the proof of Theorem 3.1). Furthermore, we assume $\gamma_t\tau_g=\tau_g\gamma_t$ for all $g\in G$ and $t\geq 0$. Hence, $\omega\gamma_t$ is G-invariant whenever ω is invariant under G.

Again, ϵ_0 is a conditional expectation of M onto N induced by a faithful normal state φ on M, and $\gamma_t=\epsilon_0\alpha_t\epsilon_0$ for $t\geq 0$.

Proposition 4.1: Let ω be a faithful, normal G-invariant state of N, if ω is G-ergodic, so is $\omega\gamma_t$ for $t\geq 0$.

We need some preliminary lemmas to prove this proposition.

Lemma 4.2: If ω is a faithful, normal state on N, then $\omega'=\omega\gamma_t$ for $t\geq 0$ is also faithful and normal.

Proof: Faithfulness can be seen as follows; for $x\in N$ and $t\geq 0$, if $\gamma_t(x^*x)=\epsilon_0[\alpha_t(x^*x)]=\epsilon_0[\alpha_t(x)^*\alpha_t(x)]=0$; then $\alpha_t(x)=0$ (due to the faithfulness of ϵ_0), and $x=0$ (because of the injectivity of α_t). Thus, $\omega'(x^*x)=\omega(\gamma_t(x^*x))=0$ implies $x=0$.

Let $\{x_\alpha\}$ be a uniformly bounded directed set of positive elements of N; then $\sup_\alpha\gamma_t(x_\alpha)=\sup_\alpha\epsilon_0[\alpha_t(x_\alpha)]=\epsilon_0[\sup_\alpha\alpha_t(x_\alpha)]$. Furthermore, α_t is unitarily implementable: $\alpha_t(x)=u_t x u_t^*$, for $x\in N$, we have $\sup_\alpha\alpha_t(x_\alpha)=\sup_\alpha u_t x u_t^*=u_t(\sup_\alpha x_\alpha)u_t^*=\alpha_t\sup_\alpha x_\alpha$. Then $\sup_\alpha\gamma_t(x_\alpha)=\epsilon_0[\alpha_t\sup_\alpha x_\alpha]=\gamma_t(\sup_\alpha x_\alpha)$. Therefore, $\sup_\alpha\omega'(x_\alpha)=\sup_\alpha\omega(\gamma_t x_\alpha)=\omega(\sup_\alpha\gamma_t x_\alpha)=\omega(\gamma_t\sup_\alpha x_\alpha)=\omega'(\sup_\alpha x_\alpha)$ which shows the normality of ω'.

Now, the key point in the proof of Proposition 4.1 is due to the fact that two faithful normal states on a von Neumann algebra is the same up to equivalence. Indeed, let ω and ω' be two faithful normal states on N, π_ω and $\pi_{\omega'}$ the corresponding cyclic representations with cyclic vectors ξ_ω and $\xi_{\omega'}$ respectively. Faithfulness of ω (resp. ω') implies π_ω (resp. $\pi_{\omega'}$) is faithful, and ξ_ω (resp. $\xi_{\omega'}$) is also a separating vector for $\pi_\omega(N)''$ (resp. $\pi_{\omega'}(N)''$). Moreover, due to the normalities of ω and ω', π_ω and $\pi_{\omega'}$ are W^*-representations of N. Therefore, two faithful W^*-representations π_ω and $\pi_{\omega'}$ of N are unitarily equivalent, since π_ω (resp. $\pi_{\omega'}$) has a separating and cyclic vector ξ_ω (resp. $\xi_{\omega'}$).[10] Hence, we have proved:

Lemma 4.3: If ω and ω' are two faithful normal states on N, then π_ω and $\pi_{\omega'}$ are unitarily equivalent.

In order to show Proposition 4.1 more explicitly, we need two additional lemmas.

Lemma 4.4: Given two Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$ and $W\subset\mathcal{B}(\mathcal{H}_1)$, let v be an isometric map of $\mathcal{H}_1$ onto $\mathcal{H}_2$ such that $W\to vWv^*$; then $(vWv^*)'=vW'v^*$.

Proof: Let $y\in W$, then $vyv^*\in vWv^*$. For $x\in(vWv^*)'$, $(vyv^*)x=x(vyv^*)$; hence $y(v^*xv)=(v^*xv)y$. It follows that $v^*xv\in W'$, and thus $x\in vW'v^*$.

On the other hand, let $x\in W'$ and $y\in W$; then $(vxv^*)(vyv^*)=vxyv^*=vyxv^*=(vyv^*)(vxv^*)$, which implies $vxv^*\in(vWv^*)'$, since $vyv^*\in vWv^*$. Hence, $vW'v^*\subseteq(vWv^*)'$.

Lemma 4.5: Given two Hilbert space $\mathcal{H}_1$, $\mathcal{H}_2$ and an isometry v of $\mathcal{H}_1$ onto $\mathcal{H}_2$ such that $v\mathcal{B}(\mathcal{H}_1)v^*=\mathcal{B}(\mathcal{H}_2)$. If M_1 and M_2 are two von Neumann subalgebras of $\mathcal{B}(\mathcal{H}_1)$, then $vM_1v^*\cap vM_2v^*=v(M_1\cap M_2)v^*$.

Proof: Let $x\in vM_1v^*\cap vM_2v^*$. For each $y\in v(M_1'\cup M_2')v^*$, $xy=yx$. However, $v(M_1'\cup M_2')v^*=v(M_1\cap M_2)'v^*=(v(M_1\cap M_2)v^*)'$; here the last equality is due to Lemma 4.4. Hence $x\in v(M_1\cap M_2)v^*$.

On the other hand, $vM_1v^*\cap vM_2v^*\supseteq v(M_1\cap M_2)v^*$ holds obviously.

Proof of Proposition 4.1: Let $\omega'=\omega\gamma_t$; then ω is also faithful and normal (Lemma 4.2); hence π_ω and $\pi_{\omega'}$ are equivalent (Lemma 4.3). Therefore, there is a isometry v of $\mathcal{H}_\omega$ onto $\mathcal{H}_{\omega'}$ such that $v\pi_\omega(N)v^*=\pi_{\omega'}(N)$ and $v\mathcal{U}_\omega(G)v^*=\mathcal{U}_{\omega'}(G)$; hence from Lemma 4.4

$$\pi_{\omega'}(N)'=(v\pi_\omega(N)v^*)'=v\pi_\omega(N)'v^*,$$

$$\mathcal{U}_{\omega'}(G)'=(v\mathcal{U}_\omega(G)v^*)'=v\mathcal{U}_\omega(G)'v^*$$

Therefore,

$$\begin{aligned}\pi_{\omega'}(N)'\cap\mathcal{U}_{\omega'}(G)'&=v\pi_\omega(N)'v^*\cap v\mathcal{U}_\omega(G)'v^*\\ &=v(\pi_\omega(N)'\cap\mathcal{U}_\omega(G)')v^*\quad\text{[by Lemma 4.5]}\\ &=v\{\lambda 1_\omega\}v^*\\ &=\{\lambda 1_{\omega'}\}.\end{aligned}$$

ACKNOWLEDGMENT

The author wishes to thank Professor G.W. Mackey for his kind hospitality at Harvard, where this paper is written.

*Supported in part by NSF Grant DID71-04010-A02.
Present address: Departamento de Matemática, Pontificia Universidade Católica, Rua Marquês de São Vicente, Rio de Janeiro, ZC-20, Brazil.
[1]N. G. van Kampen, in *Fundamental Problems in Statistical Mechanics I*, edited by E. G. D. Cohen (North-Holland, Amsterdam, 1962).
[2]C. George, I. Prigogine, and L. Rosenfeld, Mat. Fys. Medd. Dansk, Vidensk. Selsk. **38**, No. 12 (1972), and the references cited there.
[3]J. Tomiyama, Proc. Japan Acad. **33**, 608 (1957).
[4]M. Takesaki, J. Functional Anal. **9**, 306 (1972).
[5]M. Takesaki, in *Statistical Mechanics and Mathematical Problems*, edited by A. Lenard, Lecture Notes in Physics No. 20 (Springer-Verlag, Berlin, 1973).
[6]A. Connes, Ann. Scient. Eccole Normale Supérieure, 4éme Sér. **6**, 133 (1973).
[7]E. Nelsen, J. Functional Analys. **12**, 97 (1973).
[8]L. Accardi, "Non-Commutative Markof chains," (preprint).
[9]D. Ruelle, *Statistical Mechanics* (Benjamin, New York 1969).
[10]S. Sakai, *C^*-Algebras and W^*-Algebras* (Springer-Verlag, Berlin, 1971), p. 105.

Perfect fluids and symmetry mappings leading to conservation laws

D. R. Oliver, Jr. and W. R. Davis

Department of Physics, North Carolina State University, Raleigh, North Carolina 27607
(Received 10 May 1976)

Some of the results recently obtained by Glass relating to shear-free perfect fluids are extended and further interpreted. In particular, it is pointed out that certain symmetry methods are fundamental to this type of investigation. This approach of using symmetry mappings (e.g., characterizing vanishing shear) at the level of the matter tensor is seen to naturally lead to considerations of an important family of symmetry properties which include Ricci collineations as a special case. These considerations are used to obtain new conservation expressions holding for perfect fluids.

1. INTRODUCTION

Recently Glass[1] investigated several aspects of shear-free perfect fluids. Some of the new results that were obtained in this work led directly to the formulation of certain interesting local conservation expressions. The purpose of the present paper is essentially threefold: (1) to point out that certain symmetry methods are fundamental, if not essential, in this type of investigation; (2) to re-examine and further interpret some of the results that were found by Glass; and (3) to obtain and interpret, employing symmetry methods, additional conservation expressions that were not obtained by Glass.

A simple example of the relevance of symmetry considerations is immediately given by noting that vanishing shear[2] ($\sigma_{ij}=0$) is equivalent to $\mathcal{L}_u\gamma_{ij}=(2/3)\theta\gamma_{ij}$ [or $\mathcal{L}_{\varphi u}g_{ij}=2\{u_{(i}\partial_{j)}\varphi+\varphi a_{(i}u_{j)}+\varphi(1/3)\theta\gamma_{ij}\}$]. This expression which was given, for example, by Glass[1] clearly shows that vanishing shear can be equivalently regarded as an infinitesimal symmetry mapping ($x^i \to x^i+\epsilon\varphi u^i$) of the metric along the timelike congruence given by the $u^i(x)$ field.

In accord with the stated purpose of this paper, it will be shown that vanishing shear and other dynamic and kinematic conditions fundamentally relate to space–time symmetry properties defined at the level of the matter tensor. In particular, it will be seen that this type of symmetry investigation relates to a certain family of symmetry properties which are particularly interesting because of their relation to the matter tensor and a general conservation law generator. Ricci collineations are the most familiar members of this symmetry family which will be referred to as the family of contracted Ricci collineations[3] (see Fig. 1). Several investigations of Ricci collineations have been made in connection with conservation expressions concerning gravitational and electromagnetic radiation.[4] Recently, among other results, Shaha[5] obtained one special theorem, for a magnetofluid admitting Ricci collineations, that has some formal relationship to certain aspects of this paper.

Here we treat the case of perfect fluids which have matter tensors of the form

$$T_{ij}=\rho u_i u_j - p\gamma_{ij}, \tag{1.1}$$

where ρ and p are the density of total mass energy and the isotropic pressure as measured in the fluid elements rest frame, respectively. The familiar "dynamical" and "conservation" equations of the fluid follow from $\nabla_j T^{ij}=0$ and take the form

$$(\rho+p)\,a_i=\gamma_i^k\,\partial_k p \tag{1.2}$$

and

$$\dot{\rho}+(\rho+p)\,\theta=0. \tag{1.3}$$

In addition it is assumed that the Ricci tensor is related to the matter tensor by Einstein's field equations $R_{ij}-\frac{1}{2}g_{ij}R=\kappa T_{ij}$. Furthermore, the fluid is taken to be a thermodynamical perfect fluid.[6]

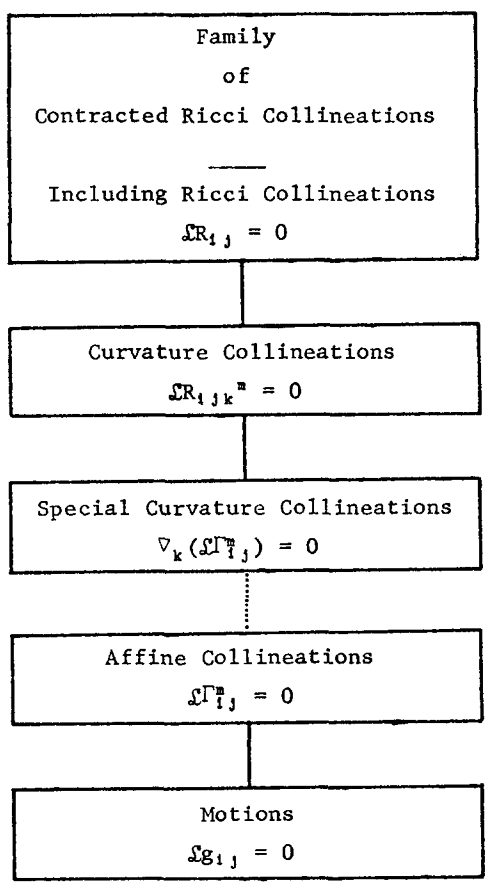


FIG. 1. Symmetry property inclusion diagram.

2. SYMMETRY PROPERTIES RELATING TO VANISHING SHEAR AND ISENTROPIC FLOW

In this section the relations between certain conditions on perfect fluids and corresponding symmetry properties at the level of the matter tensor are investigated. Thus space—time symmetries will be examined in terms of Lie deformations of the components of the matter tensor (i.e., $T_{ij}-\frac{1}{2}g_{ij}T$) or alternatively Lie deformations of R_{ij}. Clearly, symmetry demands (in the context of an assumed perfect fluid) can be made by requiring that $\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)=K_{ij}$, where K_{ij} is a symmetric tensor which is not by definition equal to $\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)$ for all perfect fluids.

First, we look at the symmetry family defined by requiring K_{ij} to be a trace-free tensor.[7] The importance of this family of symmetry properties is that each of its members satisfies the condition $g^{ij}\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)=0$ which is equivalent to the conservation law generator[8]

$$\nabla_i[\sqrt{-g}(T^i_j-\tfrac{1}{2}\delta^i_jT)\,\xi^j]=0. \tag{2.1}$$

Even though the general form of the conservation law is the same for all members of this family, it follows that the particular form the conservation law will take will depend upon the given K_{ij} and the particular ξ^i (assuming that the relevant symmetry is admitted).

In view of our particular interest in symmetry mappings along the timelike u^i congruence, we observe that $\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)$ in the case of a perfect fluid for $\xi^i=\varphi u^i$ may be expressed in the form

$$\pounds_\xi(T_{ij}-\tfrac{1}{2}g_{ij}T)=\psi(T_{ij}-\tfrac{1}{4}g_{ij}T)+(\rho+3p)\,u_{(i}\gamma^k_{j)}(a_k\varphi+\partial_k\varphi)+\varphi(p-\rho)\sigma_{ij}+\tfrac{1}{4}g_{ij}\nabla_k[(\rho+3p)\,\varphi u^k], \tag{2.2}$$

where it has been assumed $\rho+p\neq 0$ with $\psi=(\rho+p)^{-1}\{\nabla_k[(\rho+3p)\,\varphi u^k]-(4/3)(\rho+3p)^{1/2}\varphi\nabla_k[(\rho+3p)^{1/2}u^k]\}$. We observe that the first three terms on the right hand side of (2.2) are trace free while the last term is not trace free.[9]

We are now in a position to see in detail how two familiar conditions placed on perfect fluids lead to timelike symmetry demands at the level of the matter tensor.

Theorem 2.1: (a) A perfect fluid ($\rho\neq p$) is shear free if and only if $\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)=\psi(T_{ij}-\frac{1}{4}g_{ij}T)+(\rho+3p)\,u_{(i}\gamma^k_{j)}(a_k\varphi+\partial_k\varphi)+\frac{1}{4}g_{ij}\nabla_k[(\rho+3p)\,\varphi u^k]$ for all $\xi^i=\varphi u^i$. (b) A thermodynamical perfect fluid ($\rho+3p\neq 0$) is isentropic if and only if $\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)=\psi(T_{ij}-\frac{1}{4}g_{ij}T)+(p-\rho)\varphi\sigma_{ij}+\frac{1}{4}g_{ij}\nabla_k[(\rho+3p)\varphi u^k]$ for $\xi^i=u^i/f$, where f is the index of the fluid.[6]

Proof: The proof of part (a) follows from $p\neq\rho$ and (2.2). The proof of part (b) follows from $\rho+3p\neq 0$, (2.2) and the equivalence of isentropic flow to $a_k=\gamma^i_k(\partial_if)/f$ for a thermodynamical perfect fluid.

3. SYMMETRY CONSIDERATIONS IN CONNECTION WITH $\nabla_i(\sqrt{-g}\,r^{-1/3}f^2\omega^2u^i)=0$

Glass[1] has shown that for a shear-free isentropic perfect fluid with $\nabla_i(nu^i)=0$, for some function n, one obtains the conservation law $\nabla_i(n^{1/3}f\omega u^i)=0$ which is equivalent to $\nabla_i(\sqrt{-g}\,n^{-1/3}f^2\omega^2u^i)=0$. We shall see in this section how this conservation expression can be obtained in accord with the point of view provided by symmetry methods.

Theorem 3.1: For a thermodynamical perfect fluid with $\pounds_u(f\omega_{ij})=0$ and $\sigma_{ij}\omega^i\omega^j=0$ the conservation law

$$\nabla_i(\sqrt{-g}\,r^{-1/3}f^2\omega^2u^i)=0 \tag{3.1}$$

holds.

Proof: Note that $\pounds_uf\omega_{ij}=0$ implies $r^{-4/3}f\omega^{ij}\pounds_uf\omega_{ij}=0$. Using $\pounds_u(f\omega_{ij})=0$, $\sigma_{ij}\omega^i\omega^j=0$, and $\nabla_i(ru^i)=0$ one can show $f\omega_{ij}\pounds_u(r^{-4/3}f\omega^{ij})=f\omega_{ij}\pounds_u(r^{-4/3}g^{ik}g^{jm}f\omega_{km})=f^2\omega_{ij}\,\omega_{km}\pounds_u(r^{-4/3}g^{ik}g^{jm})=0$. This implies $\pounds_u(r^{-4/3}f^2\omega^2)=0$ which leads to $\nabla_i[\sqrt{-g}\,r^{-1/3}f^2\omega^2u^i]=0$. Alternatively, we can make a stronger statement as to the conditions for the conservation law (3.1) to hold.

Theorem 3.2: A thermodynamical perfect fluid with $\sigma_{ij}\omega^i\omega^j=0$ admits the conservation law (3.1) if and only if one of the following conditions is met: (i) $\pounds_\chi\rho=0$ where $\chi^i=\frac{1}{2}\eta^{ijkm}u_j\omega_ka_m$ ($\chi^i\neq 0$), (ii) $\omega_i=\lambda a_i$ (i.e., $\pounds_au_i=a^2u_i$), or (iii) $a_i=0$ (i.e., $\pounds_uu_i=0$).

Proof: Using the relation for $D\omega^2/ds$ in Ref. 2, $\sigma_{ij}\omega^i\omega^j=0$, and the perfect fluid assumption, we find that $D\omega^2/ds=-2[(2/3)\,\theta+\dot{p}/(\rho+p)]\,\omega^2-(\rho+p)^{-1}a_n\omega^{nm}\partial_m\rho$. With $\nabla_i(ru^i)=0$, $\omega^{nm}=\eta^{mnij}\omega_iu_j$, $f=(\rho+p)/r$ and (1.3), we have $D/ds(r^{-4/3}\omega^2f^2)=(2f/r^{7/3})\,\chi^m\partial_m\rho$. Thus $D/ds(r^{-4/3}\omega^2f^2)=0$ which is equivalent to $\nabla_i(\sqrt{-g}\,r^{-1/3}\omega^2f^2u^i)=0$ if and only if $\chi^m\partial_m\rho=0$. If $\chi^m\neq 0$, then $\chi^m\partial_m\rho=\pounds_\chi\rho=0$. When $\chi^m=0$, then either $\omega^i=\lambda a^i$ (i.e., $\pounds_au_i=a^2u_i$) or $a_i=0$ (i.e., $\pounds_uu_i=0$).

We now show how the conservation law (3.1) relates to a particular symmetry at the level of the matter tensor. For a symmetry vector $\xi^i=[r^{-1/3}f^2\omega^2/(\rho-\frac{1}{2}T)]u^i$ satisfying $g^{ij}\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)=0$, we obtain, with the help of (2.1), a conservation law equivalent to (3.1). We further observe that $\pounds_u(f\omega_{ij})=0$ implies $\pounds_{\varphi u}(f\omega_{ij})=0$ for all φ. Thus this symmetry property, which is a special timelike member of the family of contracted Ricci collineations, leads to the same conservation law as the symmetry demands $\pounds_\xi f\omega_{ij}=0$ and $\sigma_{ij}\omega^i\omega^j=0$.

4. ROTATION RELATED CONSERVATION EXPRESSION BASED ON A RICCI COLLINEATION

It has been shown in previous sections of this paper how symmetry properties are related to certain conditions that may be placed upon perfect fluids. Indeed, we looked for symmetry properties corresponding to given conditions. Here we require the symmetry $\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)=0$ for $\xi^i=\varphi\omega^i$ and explore some of the conditions that this Ricci collineation places on a perfect fluid.

Theorem 4.1: If a perfect fluid ($\rho\neq p$) admits the symmetry property $\pounds_\xi(T_{ij}-\frac{1}{2}g_{ij}T)=0$ for $\xi^i=\varphi\omega^i$, then (a) $\nabla_j[\alpha(\rho+3p)\omega^j]=0$ where $\omega^j\partial_j\alpha=0$, (b) $\omega^j\partial_j[(\rho+3p)(\rho-p)]=0$ and if in addition[10] $a_i\omega^i\neq 0$, then (c) $D\omega^2/ds=-[(4/3)\,\theta+2\dot{p}/(\rho+p)]\omega^2-2\sigma_{ij}\omega^i\omega^j$ or equivalently $a_n\omega^{nm}\partial_m\rho=-2\chi^m\partial_m\rho=0$.

Proof: (a) Using the perfect fluid matter tensor, the relations involving ω^i in Ref. 2 and (1.2), we find that

$$\mathcal{L}_\xi(T_{ij} - \tfrac{1}{2}g_{ij}T) = \varphi\nabla_k[(\rho+p)\omega^k]u_i u_j + \tfrac{1}{2}\varphi\omega^k\partial_k(p-\rho)g_{ij} + \tfrac{1}{2}(p-\rho)\mathcal{L}_\xi g_{ij}, \tag{4.1}$$

where $\xi^i = \varphi\omega^i$. Part (a) of this theorem then follows by noting that

$$u^i u^j \mathcal{L}_\xi(T_{ij} - \tfrac{1}{2}g_{ij}T) = 0 = \tfrac{1}{2}\nabla_j[(\rho+3p)\omega^j]. \tag{4.2}$$

(b) From (1.2), (4.2) and the relation involving $\nabla_i\omega^i$ in footnote 2 we have

$$\omega^j\partial_j(p-\rho) = (p-\rho)\nabla_j\omega^j. \tag{4.3}$$

Next making use of (4.2) and (4.3), part (b) of this theorem follows.

(c) The relation $u^i\omega^j\mathcal{L}_\xi(T_{ij} - \frac{1}{2}g_{ij}T) = 0$ can be written in the form

$$D\omega^2/ds = [-2(\dot\varphi/\varphi) + (2/3)\theta]\omega^2 - 2\sigma_{ij}\omega^i\omega^j. \tag{4.4}$$

From $u^i a^j\mathcal{L}_\xi(T_{ij} - \frac{1}{2}g_{ij}T) = 0$, one obtains

$$[\dot\varphi - \varphi\theta - (\varphi\dot p)/(\rho+p)]a_i\omega^i = 0. \tag{4.5}$$

Combining (4.4), (4.5), and $a_i\omega^i \neq 0$ proves the first portion of part (c). The second half of part (c) and the equivalence of the two halves follow because for a general perfect fluid we can prove

$$D\omega^2/ds = -[(4/3)\theta + 2\dot p/(\rho+p)]\omega^2 - [a_n\omega^{nm}(\partial_m\rho)/(\rho+p)] - 2\sigma_{ij}\omega^i\omega^j \tag{4.6}$$

with the help of results previously given.[2]

In view of Theorem (3.2) we see that for the special case where $\sigma_{ij}\omega^i\omega^j = 0$, part (c) of Theorem (4.1) implies the conservation law $\nabla_i[\sqrt{-g}\, r^{-1/3}f^2\omega^2u^i] = 0$. Thus in addition to the other new information contained in Theorem (4.1) (to be discussed) we have gained further insight relating to Theorems (3.1) and (3.2). This conservation law was shown to hold for isentropic shear-free perfect fluids by Glass.[1] Earlier an essentially equivalent expression for the special case of geodesic fluids was given, for example, by Ryan and Shepley[11] who interpreted it as an expression of "conservation of rotation."

We now show how part (a) of Theorem (4.1) can be interpreted with the help of some results obtained in a paper by Greenberg.[12] In particular, Greenberg[12] has shown

$$(1/A)(DA/d\tau) = -(1/\omega)(D\omega/d\tau) - (a_i\omega^i/\omega), \tag{4.7}$$

where A is the proper area subtended by the "vortex lines" as they pass through the "screen" which is the two surface dual to the surface formed by u_i and ω_i and where $DA/d\tau = (\omega^i/\omega)\partial_i A$. Now using $\nabla_i\omega^i + 2a_i\omega^i = 0$, part (a) of Theorem (4.1) and (4.7), we find that

$$D/d\tau[(\rho+3p)^{1/2}A\omega] = 0. \tag{4.8}$$

Thus $(\rho+3p)^{1/2}A\omega$ is a constant along the vortex tube. This is a generalization of the Kelvin–Helmholtz theorem of Newtonian theory where ωA is constant along the vortex tube.[12]

ACKNOWLEDGMENTS

The authors wish to acknowledge helpful discussions with W. M. Baker, L. H. Green, and L. K. Norris.

[1]E. N. Glass, J. Math. Phys. 16, 2361 (1975).

[2]In accord with the notations and definitions used by J. A. Schouten, *Ricci-Calculus* (Springer-Verlag, Berlin, 1954), here and throughout the paper we use (i) ∇_k for the operation of covariant differentiation, (ii) $\mathcal{L}_\xi$ for the operation of Lie differentiation with respect to the vector ξ^i, and (iii) round and square brackets on indices for the operations of symmetrization and antisymmetrization, respectively. The following definitions of quantities associated with a timelike congruence defined by the four-velocity u^i ($u^iu_i = 1$ with signature of metric -2) will be needed (i) acceleration $a^i = u^j\nabla_j u^i$; (ii) expansion, $\theta = \nabla_j u^j$; (iii) projection tensor, $\gamma_{ij} = g_{ij} - u_iu_j$; (iv) shear tensor, $\sigma_{ij} = \nabla_{(j}u_{i)} - a_{(i}u_{j)} - (1/3)\theta\gamma_{ij}$; (v) rotation tensor, $\omega_{ij} = \nabla_{[j}u_{i]} - a_{[i}u_{j]}$; (vi) rotation vector, $\omega^i = \frac{1}{2}\eta^{ijkm}u_j\nabla_m u_k$, where η^{ijkm} is the permutation tensor with $\eta^{0123} = -(-g)^{-1/2}$; (vii) dual of ω_{ij}, ${}^*\omega^{ji} = \frac{1}{2}\eta^{jikm}\omega_{km}$; (viii) shear scalar, $2\sigma^2 = \sigma_{ij}\sigma^{ij}$; and (ix) rotation scalar, $2\omega^2 = \omega^{ij}\omega_{ij} = -2\omega_i\omega^i$. We further define $(D/ds)B = \dot B = u^k\nabla_k B$, where B is a tensor of arbitrary rank. Also, the following identities will prove to be useful: (i) ${}^*\omega_{ij} = 2\omega_{[i}u_{j]}$, (ii) $\omega^{mn} = \eta^{mnij}\omega_i u_j$, (iii) $\sigma_{ij}\omega^i\omega^j = \sigma_{ij}\omega^{ki}\omega_k{}^j$, (iv) $\nabla_i\omega^i + 2a_i\omega^i = 0$, (v) $D\omega^i/ds = -\omega^j a_j u_i + \frac{1}{2}\eta^{ijkm}u_j\nabla_m a_k - (2/3)\theta\omega^i + \sigma^i{}_m\omega^m$, and (vi) $D\omega^2/ds = -(4/3)\theta\omega^2 - 2\sigma_{ij}\omega^i\omega^j + \omega^{ij}\nabla_j a_i$.

[3]See symmetry property inclusion diagram (Fig. 1). A symmetry property inclusion diagram of this type (without the family of contracted Ricci collineations) first appeared in Katzin–Levine–Davis[4] (1969) and later in a considerably expanded version (less family of contracted Ricci collineations and with minor corrections) in articles by G. H. Katzin and J. Levine [Colloq. Math. 26, 21 (1972)] and W. R. Davis ["Conservation Laws in Einstein's General Theory of Relativity" in *Lanczos Festschrift*, edited by B. K. P. Scaife (Academic, London, 1974), pp. 29–64].

[4]G. H. Katzin, J. Levine, and W. R. Davis, J. Math. Phys. 10, 617 (1969); Tensor N. S. 21, 52 (1970); and K. P. Singh and D. N. Sharma, J. Phys. A: Math. Gen. 8, 1875 (1975).

[5]R. R. Shaha, Ann. Inst. H. Poincaré 20, 189 (1974).

[6] A thermodynamical perfect fluid, following A. Lichnerowicz [*Relativistic Hydrodynamics and Magnetohydrodynamics* (Benjamin, New York, 1967), pp. 23–31], is defined by (i) $\rho = r(1+\epsilon)$, where r is the "proper material density" and ϵ the "specific internal energy"; (ii) $T\,ds = d\epsilon + p\,d(1/r)$, where T is the "proper temperature," s the "specific proper entropy" and d is the operator of exterior differentiation; (iii) $f = (p+\rho)/r$, the "index of the fluid"; and (iv) $\nabla_i(ru^i) = 0$.

[7]This symmetry family, which was referred to earlier in the Introduction and Ref. 3, is called the family of contracted Ricci collineations.

[8]This conservation law generator was first obtained by Katzin–Levine–Davis[4] (1969) (for space–times with $R=0$ admitting RC and CC). Somewhat later C. D. Collinson [Gen. Rel. Grav. 1, 137 (1970)] pointed out that this expression holds for any space–time ($R \neq 0$) admitting RC. Clearly, this conservation law generator may be regarded as a generalization of the familiar Trautman expression [A. Trautman, "Conservation Laws in General Relativity" in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), pp. 169–198] $\nabla_j(\sqrt{-g}\,T^j_i\xi^i) = 0$ which only holds, for general T_{ij}, if the ξ^i is a symmetry vector characterizing a group of motions (M) (isometries) admitted by the given space–time. Of course, here we have assumed Einstein's field equations $(R_{ij} - \frac{1}{2}g_{ij}R = \kappa T_{ij})$.

[9]In addition, it is noted that no sum of terms on the right-hand side of (2.2) can vanish unless each individual term in the given sum vanishes.

[10]With the assumptions as given, the following statements are equivalent: (i) $a_i\omega^i \neq 0$, (ii) $\omega^k\partial_k(p-\rho) \neq 0$, (iii) $\nabla_i\omega^i \neq 0$, and (iv) $\mathcal{L}_\xi g_{ij} \neq 0$ for $\xi = \varphi\omega^i$.

[11]M. Ryan and L. Shepley, *Homogeneous Relativistic Cosmologies* (Princeton U. P., Princeton, N. J., 1975), pp. 53–56.

[12]P. J. Greenberg, J. Math. Anal. Appl. 30, 128 (1970).

Principal null directions without spinors*

Ann K. Stehney

Department of Mathematics, Wellesley College, Wellesley, Massachusetts 02181
(Received 9 February 1976)

The method used by Petrov to obtain the first classification of vacuum space–times has since been overshadowed by methods which yield a finer classification based on principal null directions. It is shown that this finer classification can be expressed in Petrov's terms and a simple algorithm is obtained for finding all repeated principal null directions by matrix methods.

1. INTRODUCTION

In the recent literature, the Riemann tensor of a vacuum space—time, or more generally of an Einstein space, is usually classified in terms of the coincidence of its "principal null directions" rather than its eigenvalues and eigen-bivectors. That these null directions have physical importance themselves and also yield physical information about Riemann tensors of different types has contributed to the appeal of the tensor and spinor methods of classification,[1] to the extent that the matrix methods used by Petrov[2] in formulating the original classification are now widely overlooked. In addition, it is felt that the matrix methods are unnecessarily cumbersome.[3]

It is known that principal null directions can be described in Petrov's terms. A definition as elegant as those of the tensor- and spinor-theoretic approaches has been given by Thorpe[4] in the formalism of the Riemann tensor of an Einstein space as a symmetric transformation on a normed complex vector space. However, Petrov and Thorpe point out only three classes of such tensors, while there are six possible coincidence patterns of principal null directions.

It is our purpose to modify the Petrov—Thorpe scheme to derive the six classes and to obtain the repeated principal null directions of a Riemann tensor by matrix methods. The classification is determined by the minimal polynomial of a complex 3×3 matrix. Once this is known, the eigen-bivectors are easily found. The null directions then arise naturally from the eigenbivectors and their coincidence from the repetition of the corresponding eigenvalue.

The classification is described in Sec. 2. In Sec. 3, the canonical forms of the Riemann tensor are used to verify the coincidence of the principal null directions of each type. The algorithm for finding the repeated principal null directions is summarized in Sec. 4. It may be used without finding the basis with respect to which the matrix is in canonical form.

2. THE PETROV CLASSIFICATION

We shall first summarize the Petrov—Thorpe viewpoint, referring the reader to Thorpe[4] for details and proofs. At the end of this section we refine their methods to obtain six classes instead of three.

The classification assigned at each point to a manifold M with arbitrary Lorentz metric is that of the (Weyl) conformal curvature part of its (Riemann) curvature operator R. It therefore suffices to assume that the Ricci curvature of R vanishes, i.e., that the Einstein field equation for a vacuum is satisfied. The action of R on the tangent space to M at one point is abstracted to the following situation.

Let V be a four-dimensional vector space with inner product $\langle\ ,\ \rangle$ of Lorentz signature $(+++-)$. The bivector space $W=\Lambda^2(V)$ has an induced inner product, also denoted $\langle\ ,\ \rangle$, of signature $(+++---)$. We shall assume that V is oriented and $\{e_1, e_2, e_3, e_4\}$ is an orthonormal basis for V which is compatible with the orientation, where $\langle e_4, e_4\rangle=-1$. Following Ref. 4, we define the corresponding Lorentz basis for W to be

$$\{E_1,\ldots,E_6\}=\{e_1\wedge e_2, e_1\wedge e_3, e_1\wedge e_4, e_3\wedge e_4, e_4\wedge e_2, e_1\wedge e_4\}. \tag{1}$$

A curvature operator on V is a linear transformation $R: W\to W$, which is self-adjoint with respect to the inner product. Its Ricci curvature is the map $r: V\to V$ defined by $\langle r(e_i), e_j\rangle=\sum_k\langle R(e_i\wedge e_k), e_j\wedge e_k\rangle$. If $r=0$, then with respect to any Lorentz basis, R is represented by a matrix of the form

$$[R]=\begin{bmatrix} A & B\\ -B & A\end{bmatrix},$$

where 3×3 matrices A and B are symmetric and have trace zero. The canonical forms for such a matrix are of three types, namely:

Type I:

$$A=\begin{bmatrix} a_1 & 0 & 0\\ 0 & a_2 & 0\\ 0 & 0 & a_3\end{bmatrix} \quad\text{and}\quad B=\begin{bmatrix} b_1 & 0 & 0\\ 0 & b_2 & 0\\ 0 & 0 & b_3\end{bmatrix},$$

Type II:

$$A=\begin{bmatrix} a_1 & 0 & 0\\ 0 & a_2+1 & 0\\ 0 & 0 & a_2-1\end{bmatrix} \quad\text{and}\quad B=\begin{bmatrix} b_1 & 0 & 0\\ 0 & b_2 & 1\\ 0 & 1 & b_2\end{bmatrix},$$

Type III:

$$A=\begin{bmatrix} 0 & 1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \quad\text{and}\quad B=\begin{bmatrix} 0 & 0 & 0\\ 0 & 0 & -1\\ 0 & -1 & 0\end{bmatrix},$$

where in each case $\operatorname{tr}A=\operatorname{tr}B=0$.

Petrov based his classification on the Segre characteristic of the symmetric complex matrix $C=A+iB$.

TABLE I. Refined Petrov classification of curvature operators. The numbers $\lambda_j = a_j + ib_j$ are the characteristic roots of the matrix $C = A + iB$.

Class	Characteristic polynomial of $C = \det(\lambda I - A - iB)$	Minimal polynomial of C	Independent invariants
I. a. λ_j's distinct	$(\lambda-\lambda_1)(\lambda-\lambda_2)\times(\lambda-\lambda_3)$	$(\lambda-\lambda_1)(\lambda-\lambda_2)\times(\lambda-\lambda_3)$	a_1, b_1, a_2, b_2
b. $\lambda_3=\lambda_2\neq\lambda_1$	$(\lambda-\lambda_1)(\lambda-\lambda_2)^2$	$(\lambda-\lambda_1)(\lambda-\lambda_2)$	a_1, b_1
c. $\lambda_3=\lambda_2=\lambda_1=0$	λ_3	λ	---
II. a. $\lambda_2\neq\lambda_1$	$(\lambda-\lambda_1)(\lambda-\lambda_2)^2$	$(\lambda-\lambda_1)(\lambda-\lambda_2)^2$	a_1, b_1
b. $\lambda_2=\lambda_1=0$	λ^3	λ^2	---
III. As given	λ^3	λ^3	---

The numbers $\lambda_j = a_j + ib_j$ are the characteristic roots of the matrix C. The types correspond to the cases where its minimal polynomial (I) has distinct linear factors $(\lambda - \lambda_j)$, (II) has factor $(\lambda - \lambda_2)$ repeated twice, and (III) is λ^3.

Thorpe obtained the canonical forms by regarding W as a three-dimensional complex vector space and R as the complex linear transformation on W represented by the matrix C. Types I, II, and III have respectively 3, 2, and 1 independent eigenvectors.

These three classes may be refined by considering repetitions among the λ_j's. Taking into account the condition $\mathrm{tr}A = \mathrm{tr}B = 0$, we obtain the six cases listed in Table I. They correspond to the six possible minimal polynomials of complex 3×3 matrices with trace zero. The maximum number of invariants for each class is given in the last column. One number of each pair a_j, b_j may be zero. In class Ia, one pair a_j, b_j may be identically zero. Notice that the characteristic polynomial of C may not even distinguish between Types I, II, and III.

3. PRINCIPAL NULL DIRECTIONS

Consider a point m of a vacuum space—time M at which the curvature tensor R_{abcd} is not identically zero. The principal null directions (p.n.d.'s) at m are the directions along the nonzero null vectors k_a at m which satisfy

$$k_{[a}R_{b]ij[c}k_{d]}k^i k^j = 0. \tag{2}$$

This quartic equation has in general four independent solutions, but will have fewer solutions if any are "repeated" or "coincide," that is, if they satisfy a similar equation of lower degree. The possible coincidence patterns of the four solutions are just the partitions of four, and are summarized by the symbols [1111], [211], [22], [31], and [4] (see Pirani[1]). The case $R=0$ has no distinguished null directions and is included in this scheme with the symbol [-].

The p.n.d.'s may be defined via the transformational approach as follows. Let V be the oriented tangent space at m. Let $*$ denote the Hodge star operator, which assigns to any oriented subspace its oriented orthogonal complement. Its matrix with respect to any (orientation-respecting) Lorentz basis is

$$[*] = \begin{bmatrix} 0 & -I \\ -I & 0 \end{bmatrix},$$

where I is the 3×3 identity matrix. Then $\langle \xi, *\xi\rangle = 0$ if and only if ξ is a decomposable bivector: $\xi = x\wedge y$ for some $x, y \in V$. A common zero of $\langle\xi,\xi\rangle$ and $\langle\xi,*\xi\rangle$ thus determines a plane[5] which meets the null cone of V in a single null direction, i.e., the multiples of some nonzero null vector K.

Proposition 1 (*Thorpe*): A nonzero null vector K represents a principal null direction of R if and only if $\langle R\xi, \xi\rangle = 0$ and $\langle R\xi, *\xi\rangle = 0$ for some null plane ξ containing K.

Proof: Equation (2) is the tensorial form of the equation

$$\langle R(K\wedge[\langle K,x\rangle y - \langle K,y\rangle x]), K\wedge[\langle K,z\rangle w - \langle K,w\rangle z]\rangle = 0 \tag{3}$$

for all $x, y, z, w \in V$. Let $v(x,y)$ denote the vector $\langle K,x\rangle y - \langle K,y\rangle x$. Then $\langle K, v\rangle = 0$ and, unless v is a multiple of K, $\langle v, v\rangle > 0$. Conversely, every spacelike vector orthogonal to K is $v(x,y)$ for some x and y. Therefore, Eq. (3) is equivalent to

$$\langle R(K\wedge v), K\wedge v'\rangle = 0 \tag{3'}$$

for all spacelike vectors v and v' orthogonal to K.

Given a null plane ξ containing K, there are vectors X and Y so that $\langle K,X\rangle = \langle K,Y\rangle = \langle X,Y\rangle = 0$, $\xi = K\wedge X$, and $*\xi = K\wedge Y$. Now any vector v orthogonal to K is a linear combination of K, X, and Y, where X and Y arise as above from an arbitrary null plane ξ containing K. Therefore, writing $K\wedge v$ and $K\wedge v'$ as linear combinations of ξ and $*\xi$, (3′) becomes

$$\langle R\xi, \xi\rangle = \langle R\xi, *\xi\rangle = 0. \tag{3''}$$

Finally we remark that if this equation is satisfied by one such ξ, it is satisfied by all. ||

Similar arguments applied to the tensorial definitions of the repeated principal null directions prove the following result.

Proposition 2: A nonzero null vector K represents a principal null direction which is

(a) double if and only if $\langle R\xi, \eta\rangle = 0$,

(b) triple if and only if $R\xi = 0$, and

(c) quadruple if and only if $R\eta = 0$

for every null plane ξ containing K and/or every plane η containing K.

The conditions given in this proposition for triple and quadruple p.n.d.'s are linear in K. Condition (a) for double p.n.d.'s has a linear version, which is derived as follows. Given any null vector K, there is a unique timelike plane η which contains K and which is orthogonal to all null planes containing K. Given such a null plane ξ, any other such null plane is a linear combination of ξ and $*\xi$, and an arbitrary plane containing K is a linear combination of ξ, $*\xi$, and η. Condition (a) now becomes: There exist real numbers α and β such that, for all null planes ξ containing K, $R(\xi)=\alpha\eta+\beta*\xi$. It follows that there is a timelike plane ζ with $R(\zeta)=\alpha'\zeta+\beta'*\zeta$ for some α' and β'. We shall see that the coefficients in the linear combinations are the numerical invariants of the curvature operator.

Using these descriptions of repeated p.n.d.'s and the canonical forms for R given in Sec. 2, we shall show that the six classes of Table I correspond to coincident p.n.d.'s as follows:

$$\begin{array}{cccccc} \text{Ia} & \text{Ib} & \text{Ic} & \text{IIa} & \text{IIb} & \text{III} \\ [1111] & [22] & [-] & [211] & [4] & [31] \end{array} \tag{4}$$

In each case, let $\{e_1, e_2, e_3, e_4\}$ be a basis for V such that the matrix for R is in canonical form with respect to the corresponding Lorentz basis for W.

Type III: Any null plane containing $K=(e_2+e_4)$ is a linear combination of $\xi=e_1\wedge K$ and $*\xi=-e_3\wedge K$. Since $R(\xi)=R(*\xi)=0$ and $R(e_4\wedge e_2)\neq 0$, K is a triple but not a quadruple p.n.d.

Type IIb: $(a_j=b_j=0)$; Any plane containing $K=(e_3-e_4)$ is a linear combination of $\xi=e_1\wedge K$, $*\xi=-e_2\wedge K$, and $\eta=e_3\wedge e_4$. Since $R(\xi)=R(*\xi)=R(\eta)=0$, K is a quadruple p.n.d.

Type IIa: Let $K=(e_3-e_4)$ and ξ, η be as in IIb. Then $R(\xi)=a_2\xi+b_2*\xi$ and $R(*\xi)=a_2*\xi-b_2\xi=a_2*\xi+b_2**\xi$. Thus K is a double p.n.d., but not a triple p.n.d. since a_2 and b_2 are not both zero. Notice that $R(\eta)=a_1\eta+b_1*\eta$. It can easily be shown that for other values of α and β, the transformation $R-\alpha I-\beta*$ has trivial kernel. Hence R has no other repeated p.n.d.'s.

Type Ic: is clear.

Type Ib $(a_3=a_2=-\frac{1}{2}a_1,\ b_3=b_2=-\frac{1}{2}b_1)$: If $K^{\pm}=(e_3\pm e_4)$, then $\xi^{\pm}=e_1\wedge K$ and $*\xi^{\pm}=\pm e_2\wedge K^{\pm}$ are in the kernel of $(R-a_2I-b_2*)$, and so $K^{\pm}$ are each double p.n.d.'s. Notice that $\eta=e_3\wedge e_4$ is in the kernel of $(R-a_1I-b_1*)$.

Type Ia: The equation $(R-\alpha I-\beta*)\xi=0$ has solutions precisely for $\alpha,\beta=a_j, b_j$, but none of the decomposable solutions are null. Hence R has no repeated p.n.d.'s.

4. ALGORITHM FOR THE NULL DIRECTIONS

Although the canonical forms for vacuum curvature operators were used in the last section to verify the correspondence between the matrix and p.n.d. classification schemes, it is not necessary in practice to find explicitly the basis for V that yields the canonical form for R.

(i) The equation

$$\langle R(e_i\wedge e_j), e_k\wedge e_l\rangle = R_{ijkl}$$

relates R as a linear transformation to the components R_{ijkl} for R relative to any orthonormal basis $\{e_1, e_2, e_3, e_4\}$ for V. From Eq. (1), one obtains a symmetric matrix $Q_{IJ}=\langle R(E_I), E_J\rangle$, which is related to that of R by

$$[Q]=\begin{bmatrix} A & B \\ B & -A \end{bmatrix},\quad [R]=\begin{bmatrix} A & B \\ -B & A \end{bmatrix}.$$

(ii) Comparing Table I and line (4), the p.n.d. type of R is uniquely determined by the minimal polynomial of the complex 3×3 matrix $C=A+iB$.

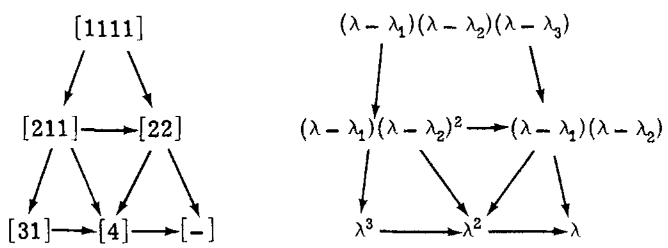
(iii) Once the eigenvalues $\lambda_j=a_j+ib_j$ of C are known, the repeated principal null directions of R may be found by solving only linear equations, as follows:

[22] *and* [211]: In these cases the characteristic polynomial of C is $(\lambda-\lambda_1)(\lambda-\lambda_2)^2$. Each solution to $(R-a_2I-b_2*)\xi=0$ is a plane which meets the null cone in a double p.n.d. Note that the matrices for I and $*$ are independent of the choice of Lorentz basis.

[31] *and* [4]: The characteristic polynomial of C is λ^3. The two independent null planes among the solutions to $R\xi=0$ each contain the repeated p.n.d. In type [4], there will be additional timelike and spacelike solutions, the former also containing the p.n.d.

Remark: There does not appear to be a similar linear condition for nonrepeated principal null directions. The conditions of Proposition 1, however, imply the following: Let K be a null vector, ξ a null plane containing K, and η the unique timelike plane orthogonal to both ξ and $*\xi$. Then K is a principal null direction for R if and only if $R\xi$ is a linear combination of ξ, $*\xi$, η, and $*\eta$. The coefficients depend quadratically on the real and imaginary parts of the eigenvalues for C.

Remark: The hierarchy of types of Riemann tensors suggested by Penrose's diagram is evident from the minimal polynomials of the types:



The arrows point to types whose minimal polynomials have lower degree and/or whose eigenvalues have more repetition.

ACKNOWLEDGMENTS

The author is indebted to the University of Chicago for its hospitality during 1974–75 and especially to R.P. Geroch and R.M. Wald for many valuable discussions during that year.

*Research undertaken at the Enrico Fermi Institute of the University of Chicago with the support of a Faculty Research Award from Wellesley College.

[1]For a discussion of these issues, see, for example, the articles by F.A.E. Pirani, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, London/New York, 1962) and *Letures on General Relativity*, *Vol. 1*, Brandeis Summer Institute 1964 (Prentice-Hall, Englewood Cliffs, N.J., 1965). Both papers give extensive references to the original research and to other survey articles. The tensor method is due to R. Debever [C.R. Acad. Sci. (Paris)

249, 1324–26 (1959)] and R. K. Sachs [Proc. Roy. Soc. (London) A 264, 309–38 (1961)] and the spinor method to R. Penrose [Ann. Phys. (N. Y.) 10, 171–201 (1960)].
[2]A. Z. Petrov, Sci. Not. Kazan State Univ. 114, 55–69 (1954). See also A. Z. Petrov, *Einstein Spaces* (Pergamon, London, 1964).
[3]As Pirani described the situation, "The spinor method is probably the shortest and most elegant route to the Petrov classification, while the tensor method is useful in many calculations. The matrix method suffers particularly from the fact that it does not bring out the hierarchy of the types... so well as other methods do; nor is it particularly convenient in calculations, but it does have some slight advantages in some of the simpler physical interpretations." (Pirani, Ref. 1, 1962, p. 216.)
[4]J. A. Thorpe, J. Math. Phys. 10, 1–7 (1969). This paper relates the Petrov classification to the critical point behavior of the sectional curvature function and thus provides a geometrical interpretation of the classification.
[5]A 2-plane $P \subset V$ is represented by any decomposable bivector $\xi = x \wedge y$ such that P is the linear span of x and y. This representation is not unique. On the other hand, a nonzero decomposable $\xi \in W$ represents a unique 2-plane. The plane P is called timelike, spacelike, or null according to $\langle \xi, \xi \rangle$ being <0, >0, or $=0$ for each representative ξ.

Extended inertial frames and Lorentz transformations. II

Edward H. Kerner

Department of Physics, University of Delaware, Newark, Delaware 19711
(Received 9 April 1976)

It is shown that there is a unique ten-parameter group of projective (fractional-linear) transformations of space–time preserving free-particle motion and containing the inhomogeneous Lorentz group as a limiting case when a characteristic fundamental length of the projective group (whose structure comes from the group of rotations in the five-dimensional space of homogeneous space–time coordinates) is made infinite. The basic differential geometry, free-particle dynamics, and Lie group algebra going with the projectively extended space–time structure are developed, and it is remarked that the Hamiltonian for free particles most generally will have eight branches, owing to the fundamental space–time metric being that for a Finsler space. The means to projectively extending electrodynamics are also briefly noted, and the need for characterizing electrodynamics by two intrinsic quantities (charge, and a purely electromagnetic length) are pointed out.

INTRODUCTION

"Hence the first part of physical science relates to the relative position and motion of bodies." This summation[1] of the base of physics has meant since Newton that, first of all, particles are the primary entities of physical study; secondly, that inertial frames are the foundation stones for the reckoning of particle motions, these being the frames with respect to which a presumptively free particle runs in uniform straight-line motion. Assuming a prototypical inertial frame given, the rest have generally been supposed derivable by linear transformations, which self-evidently send free-particle motion into free-particle motion; and the development of physics accordingly has had much to do with the exfoliations from the Galilean and the Lorentz transformations. We may call these linearly interconnected frames of Newtonian relativity or special relativity the *ordinary inertial frames* or OIF.

The present note is a recapitulation and expansion of a preliminary discussion[2] of a transformation group that preserves uniform rectilinear motion, but goes beyond the OIF into a set of *extended inertial frames* or EIF. Just as the Galilean group is embraced in the *ordinary Lorentz transformation* (OLT) group as a limiting case for $c \to \infty$, the Lorentz group is contained in the EIF group of *extended Lorentz transformations* (ELT) in the limit that a characteristic fundamental length parameter b of the latter is allowed to become infinite,

$$\text{extended Lorentz transformations} \xrightarrow[b\to\infty]{} \text{ordinary Lorentz transformations} \xrightarrow[c\to\infty]{} \text{Galilean transformations.}$$

The ELT to be described are in fact a specialized form of the projective or fractional-linear transformations, a group of 24 parameters, which have long been known to preserve free-particle motion. Now the (affine) group of linear transformations of 20 parameters is cogently narrowed for physical purposes, in either Galilean or Lorentz cases, to ten parameters, associated to the covariance of physical statements under time- and space-displacements, frame changes to relatively moving frames (boosts), and space rotations. In the Galilean case the parameters may be grouped according to

$$(1)+(3)+(3)+\tfrac{1}{2}(3)(2), \qquad \text{(Galilean)}$$

describing the separateness of translations and of boosts and of space rotations; while in the Lorentz group they fall according to

$$(1)+(3)+\tfrac{1}{2}(4)(3), \qquad \text{(OLT)}$$

expressing the separateness of translations and the connectedness of space rotations and boosts. Similarly, we shall narrow the projective group from 24 to a physically fundamental set of ten parameters; this allows room for exactly the one more universal constant b besides c. In this reduced projective scheme it will appear that all ten of the group parameters fall together in a simple and unified way into a type of single over all rotation,

$$\tfrac{1}{2}(5)(4), \qquad \text{(ELT)}$$

the rotation in question being one in the five-dimensional space of homogeneous coordinates.

The projective transformations have, of course, a long history in geometry and to some degree in physics, as may be seen in the works of Cartan, Einstein and Mayer, Hlavaty, Hoffman, Kaluza, Klein, Pauli and Solomon, Schouten, Van Dantzig, Veblen, Weyl, and others, and are referenced in Schouten's[3] and Veblen's[4] monographs. The physical thrust, starting from Kaluza,[5] was toward an unification of electromagnetism and gravitation. The projective-geometrical considerations have, however, seemed excessively general mathematically, while the physical interpretations have sometimes been opaque, and without recognition and development, so far as we have seen, of projective transformations as instruments for describing *ab initio* wider types of inertial frames (aside from questions of unification). The present discussion is wholly concerned with the latter point.

EXTENSION OF INERTIAL FRAMES

In an inertial frame in one space-dimension a free particle travels according to

$$x = x_0 + vt.$$

Very simply and obviously we have that the fractional-linear transformation

$$x=\frac{a_1'+A_{11}'x'+A_{10}'t'}{1+\alpha_1'x'+\alpha_0't'}, \quad t=\frac{a_0'+A_{01}'x'+A_{00}'t'}{1+\alpha_1'x'+\alpha_0't'},$$

is sending x,t free-motion into x',t' free-motion, owing to the common denominator $1+\alpha_1'x'+\alpha_0't'$. We must understand then that the frame x',t', an EIF, has to be reckoned as a genuine inertial frame, under the sole and sufficient criterion of the principle of inertia, that uniform straight-line motion be preserved.

Calculating v' from the above indicated $x'=x'(t')$,

$$v'=-\frac{vA_{00}'+x_0\alpha_0'-A_{10}'}{vA_{01}'+x_0\alpha_1'-A_{11}'}.$$

It is to say that the EIF particle velocity depends not merely on the OIF velocity v but also on the origin x_0 of the OIF motion. Consequently, two parallel straights $x=x_0+vt$ and $\bar{x}=\bar{x}_0+vt$ in the x,t plane go over into EIF straights, but not parallel ones, in the x',t' plane. It means that the notion of "rigidity" as ordinarily understood from affine thinking, where the preservation $x-\bar{x}=x_0-\bar{x}_0$ is carried over into $x'-\bar{x}'=x_0'-\bar{x}_0'$, is being set aside (in some slight degree at least).

From the inverse to the preceding transformation,

$$x'=\frac{a_1+A_{11}x+A_{10}t}{1+\alpha_1x+\alpha_0t}, \quad t'=\frac{a_0+A_{01}x+A_{00}t}{1+\alpha_1x+\alpha_0t},$$

we may otherwise write the kinematical transformation rule for velocity, and also acceleration,

$$v'=\frac{A_{11}v+A_{10}-x'(\alpha_1v+\alpha_0)}{A_{01}v+A_{00}-t'(\alpha_1v+\alpha_0)},$$

$$a'=\frac{A_{11}-v'A_{01}-\alpha_1(x'-v't')}{A_{01}v+A_{00}-t'(\alpha_1v+\alpha_0)}\ \frac{dt}{dt'}\ a.$$

We can then rephrase the nonrigidity of the EIF transformation in a different physical metaphor: The EIF velocity at the world point x',t' in general depends explicitly on the location of that world point. There isn't any single relative OIF—EIF velocity, for the space of EIF is not in uniform motion *en masse* with respect to OIF, but is like a fluid flowing nonuniformly—v' is describing a velocity field, as in hydrodynamics; the affine "rigidity" has gotten replaced by a "fluidity." This situation is of course symmetrical, one can interchange the roles of OIF and EIF in these, and in all statements. All the same, $a=0$ means $a'=0$ and vice versa, so both EIF and OIF observers must say, the one about the other, that the other is inertial if the one is so. The designation of the starting frame x,t as OIF, and of x',t' as EIF is in fact gratuitous.

Let us recall also the simple geometric meaning of the EIF or fractional-linear transformation. Take two intersecting planes and set up x,t coordinates in one and x',t' coordinates in the other in any convenient manner. Draw any straight in one and then project it into the other, using rays from any point (center of projection) external to both for the projection. Then the straight projects into a straight, and the coordinates of corresponding points (those linked by a ray from the center of projection) are connected by fractional-linear transformation; an example is given below, see Fig. 1. Parallel straights in one plane, however, go over to oblique straights in the other—the book held near the candle casts a shadow on the desk that is a lopsided quadrilateral (but notice too how quickly this subsides into a very good parallelogram as the book is taken further away). The essential parametrization of the transformation is: one parameter for the angle between the planes; two for the location of the center of projection; two for the origin-location of x,t coordinate axes, and one for the orientation of these axes; and another similar three for x',t' axes. The eight a, A, α are built from these, and the vanishing of the denominator $1+\alpha_1x+\alpha_0t$ above is expressing simply that x' and t' have gone over to the usual projective-geometric "line at infinity," owing to the rays from the center of projection being parallel to the x',t' plane.

The OIF-origin-dependence of EIF-velocity, or the kinematical "fluidity," or the parallelism going over to obliquity, are of course all expressing the same thing about the EIF transformation. The point has been not only to characterize in simple ways the differences between OIF and EIF, affine and projective, but to draw sharply the issue that *inertiality of frames on the one hand, and "rigidity" or "fluidity" of the transformation rules on the other hand, are separate and distinct physical propositions*. It is hard to imagine doing without inertial frames so far as can be seen. But is rigidity, after all, compulsory too? There can be no answer short of examining the alternative of fluidity. And since the latter includes the former and can be made as little different from it as desired, the issue is more exactly, not whether but in how far physical observation may be requiring rigidity or allowing fluidity.

The transformation law

$$x_i'=\frac{a_i+A_{i\beta}x_\beta}{1+\alpha_\beta x_\beta} \quad (i=1,2,3,0)$$

is the general one now of EIF type that is being allowed to take us from Cartesians $x_1,x_2,x_3\equiv\mathbf{r}$ and time $t\equiv x_0$ of one inertial frame to another (summation on repeated Greek index β from 0 to 3), involving in all 24 parameters a_i, A_{ij}, α_i, with obvious restrictions for making $\mathbf{r}'$ and t' be three-vector and three-scalar. These general projective transformations moreover exhaust those preserving uniform straight-line motion, as is clear from a discussion of Veblen and Thomas[6] and a particularly stimulating one due to Fock,[7] which, however, entail complex calculations; a sketch of a very simple proof of the important fact of exhaustiveness is given in Appendix A. The inverse transformation is similarly the ratio of linear forms with a common denominator throughout. The parameters a,A are dimensionally like those for ordinary linear transformations, e.g., $a_0=$ (time) and $a_{1,2,3}=$ (length), etc., making $\mathbf{r}'$ and t' length and time if $\mathbf{r}$ and t are. But the denominator elements call for some new characteristic length, so that for instance

$$\alpha_\beta x_\beta=\frac{\boldsymbol{\sigma}\cdot\mathbf{r}+\sigma_0ct}{\text{characteristic length}},$$

with dimensionless $\boldsymbol{\sigma},\sigma_0$.

In an obvious condensed notation we can write $\mathbf{x}'=[A;a;\alpha]\mathbf{x}$. Then if $\mathbf{x}''=[D;d;\delta]\mathbf{x}'$ is a second EIF transformation, the rule of composition of the EIF or projective group is $\mathbf{x}''=[E;e;\epsilon]\,\mathbf{x}$ with

$$[E;e;\epsilon]=[D;d;\delta][A;a;\alpha] =\left[\frac{DA+(d\alpha)}{1+\delta\cdot a};\frac{D\cdot a+d}{1+\delta\cdot a};\frac{\alpha+\delta\cdot A}{1+\delta\cdot a}\right],$$

where $(d\alpha)_{ij}$ is $d_i\alpha_j$. Notice that if $\alpha=-\delta\cdot A$, the final transformation is $[E;e;0]$ and $\mathbf{x}''$ is a linear function of $\mathbf{x}$: That is, a pair of EIF transformations arranged in this way can be thought of as providing a kind of interpolative step in conducting a linear transformation; or, we can go from one OIF to another OIF via OIF to an EIF to a second EIF.

It is noticeable too that the *homogeneous* type of EIF transformations, those with the additive terms a_i, d_i, e_i omitted in the numerators, separately form a group. These are much simpler than the inhomogeneous type and give the starting point for extending Lorentz transformations.

THE HOMOGENEOUS ELT

First recall the ordinary homogeneous Lorentz transformation (excluding spatial rotation),

$$\mathbf{r}'=\varphi_v\cdot\mathbf{r}+\gamma\mathbf{v}t,\qquad t'=\gamma(t+\mathbf{v}\cdot\mathbf{r}/c^2),$$

$$\gamma\equiv(1-\mathbf{v}^2/c^2)^{-1/2},\quad \varphi_v\equiv I+[(\gamma-1)/\mathbf{v}^2]\mathbf{vv}\equiv I+\beta\mathbf{vv},$$

with three-velocity parameter $\mathbf{v}$. Now if $\mathbf{v}$ alone is to characterize the corresponding EIF boost, one expects to represent it by

$$\mathbf{r}'=\frac{\varphi_v\cdot\mathbf{r}+\gamma\mathbf{v}t}{1+h(\mathbf{v}\cdot\mathbf{r})/cb+k(ct)b},$$

$$t'=\frac{\gamma(t+\mathbf{v}\cdot\mathbf{r}/c^2)}{1+h(\mathbf{v}\cdot\mathbf{r})/cb+k(ct)/b},$$

or, say, $\mathbf{r}'$, $t'=\boldsymbol{F}(\mathbf{r},t;\mathbf{v})$, where h and k are dimensionless constants depending at most on $\mathbf{v}^2/c^2$, and where b is some universal length on the same standing as the universal c. Introducing a second boost $\mathbf{r}''$, t'' as $F(\mathbf{r}',t';\mathbf{v}')$ we find the composition

$$\mathbf{r}''=\frac{(\varphi_{v'}\cdot\varphi_v+\gamma'\gamma\mathbf{v}'\mathbf{v}/c^2)\cdot\mathbf{r}+(\varphi_{v'}\cdot\gamma\mathbf{v}+\gamma'\gamma\mathbf{v}')\,t}{1+(h\mathbf{v}+h'\mathbf{v}'\cdot\varphi_v+k'\gamma\mathbf{v})\,\mathbf{r}/cb+(k+h'\gamma\mathbf{v}'\cdot\mathbf{v}/c^2+k'\gamma)\,ct/b},$$

$$t''=\frac{\gamma'\gamma(1+\mathbf{v}'\cdot\mathbf{v}/c^2)\,t+(\gamma'\gamma\mathbf{v}/c^2+\gamma'\mathbf{v}'\cdot\varphi_v/c^2)\cdot\mathbf{r}}{1+(h\,\mathbf{v}+h'\mathbf{v}'\cdot\varphi_v+k'\gamma\mathbf{v})\,\mathbf{r}/cb+(k+h'\gamma\mathbf{v}'\cdot\mathbf{v}/c^2+k'\gamma)\,ct/b}.$$

The numerators give the usual composition rule for pure Lorentz boosts,

$$\varphi_{v'}\cdot\varphi_v+\gamma'\gamma\mathbf{v}'\mathbf{v}/c^2=\mathcal{R}\varphi_{v''},\quad \varphi_{v'}\cdot\gamma\mathbf{v}+\gamma'\gamma\mathbf{v}'=\mathcal{R}\gamma''\mathbf{v}'',$$

$$\gamma''\mathbf{v}''=\gamma'\gamma\mathbf{v}+\gamma'\mathbf{v}'\cdot\varphi_v,\qquad \gamma''=\gamma'\gamma(1+\mathbf{v}'\cdot\mathbf{v}/c^2),$$

with $\mathcal{R}$ being the usual space rotation attending the combination of such boosts. We are now to require as well

$$h''\mathbf{v}''=(h+k'\gamma)\,\mathbf{v}+h'\mathbf{v}'\cdot\varphi_v,$$

$$k''=k+k'\gamma+h'\gamma\mathbf{v}\cdot\mathbf{v}'/c^2,$$

if the EIF boosts are to compose properly into a group. The unique h and k ensuring this are simply

$$h=\gamma,\quad k=\gamma-1,$$

for this choice alone is compatible with the composition for $\gamma''\mathbf{v}''$ and γ''.

The homogeneous ELT then are

$$\mathbf{r}'=\frac{\varphi_v\cdot\mathbf{r}+\gamma\mathbf{v}t}{1+\gamma(\mathbf{v}\cdot\mathbf{r})/cb+(\gamma-1)ct/b}$$

$$t'=\frac{\gamma(t+\mathbf{v}\cdot\mathbf{r}/c^2)}{1+\gamma(\mathbf{v}\cdot\mathbf{r})/cb+(\gamma-1)ct/b},$$

forming a group in similar fashion to the homogeneous OLT and reducing to the latter for $b\to\infty$. The group altogether is comprising boosts and, inseparably, space rotations, with the projective denominator a rotational invariant.

The inverse transformation follows from interchanging $\mathbf{r}'$ and $\mathbf{r}$, and t' and t while replacing $\mathbf{v}$ by $-\mathbf{v}$, as in the case of OLT. For the transformation rule for velocities one finds now

$$\frac{d\mathbf{r}'}{dt'}=\left(\frac{\varphi_v\cdot d\mathbf{r}}{dt}+\gamma\mathbf{v}\right)-\frac{\mathbf{r}'c}{b}\left(\frac{\gamma\mathbf{v}\cdot d\mathbf{r}/dt}{c^2}+\gamma-1\right) \times\left[\gamma\left(1+\frac{\mathbf{v}\cdot d\mathbf{r}/dt}{c^2}\right)-\frac{ct'}{b}\left(\frac{\gamma\mathbf{v}\cdot d\mathbf{r}/dt}{c^2}+\gamma-1\right)\right]^{-1}.$$

As remarked before, for fixed $d\mathbf{r}/dt$ this is not any fixed velocity $d\mathbf{r}'/dt'$ at all, but rather a field of velocities depending on world point $\mathbf{r}'$, t' location; yet again $d^2\mathbf{r}'/dt'^2$ vanishes if $d^2\mathbf{r}/dt^2$ does. The EIF particle velocity $d\mathbf{r}'/dt'$ can therefore assume any value whatever, unrestricted by c.

To the question what then is c, one must answer now that it has turned from being an electromagnetic constant to being a universal scale factor for velocity, as indeed b stands for a universal scale factor for length. The question clearly begets the counter-question, what is the extended structure of electrodynamics—viz., how are Maxwell's equations to be widened so that they are (inhomogeneous-) ELT covariant instead of OLT covariant? This will be discussed briefly below and more fully in a later report. So long as the extended inertial frames are at all admitted, electrodynamics as well as gravitation and other physical statements have to recognize both c and b from the start. An extension of space—time structure cannot but modify the way physical statements within that structure may be formulated. Of course b has evidently to be understood as very large, so that the b corrections to physical statements are small. Parenthetically, as regards Einsteinian gravitation, not only would the central idea of elemental flatness of space—time by free fall into a local inertial frame have to be reconsidered, but also the possibility that gravitation and inertia could

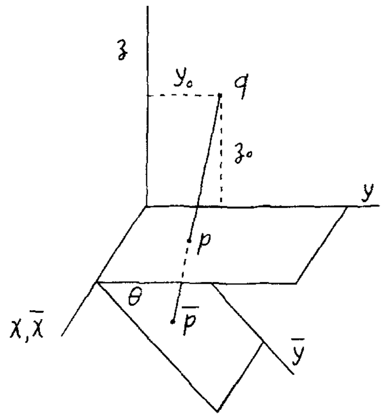


FIG. 1. Projective connection between q, p, $\bar{p}$, for the geometrical description of homographic space and time variables.

be linked in a novel way, owing to the fact that b could be replaced by $b\times$(a function of b/r_0), with r_0 the gravitational radius $G\Sigma m/c^2$ of all the masses Σm of a complete self-contained system. Of course r_0 could be taken otherwise, but the universality of gravitation suggests that a total gravitational radius may have a preferred place while also giving an embodiment to Mach's principle.

The distinctiveness of the common origin of $\mathbf{r}, t$ and $\mathbf{r}', t'$ frames above shows itself in that $\mathbf{r}=\mathbf{V}t$ implies $\mathbf{r}'=\mathbf{V}'t'$ with

$$\mathbf{V}'=\frac{\varphi_v\cdot\mathbf{V}+\gamma\mathbf{v}}{\gamma(1+\mathbf{v}\cdot\mathbf{V}/c^2)},$$

i.e., the usual Lorentz velocity composition rule. In this case $|\mathbf{V}|=c$ means $|\mathbf{V}'|=c$, so this much of special relativity holds for rays passing through the origin. The transition to the Galilean limit goes either by $b\to\infty$ and then $c\to\infty$, or else directly in one step by $c\to\infty$. Reversely, there is no working up from the Galilean limit without the avail of both c and b measures.

A simple geometric meaning for the homogeneous ELT comes from the intermediate homographic transformations

$$\bar{\mathbf{r}}=\frac{\mathbf{r}}{1-ct/b},\quad \bar{t}=\frac{t}{1-ct/b},$$

$$\bar{\mathbf{r}}'=\frac{\mathbf{r}'}{1-ct'/b},\quad \bar{t}'=\frac{t'}{1-ct'/b},$$

which bring the homogeneous ELT to linear Lorentz-like form

$$\bar{\mathbf{r}}'=\varphi_v\cdot\bar{\mathbf{r}}+\gamma\mathbf{v}\bar{t},\quad \bar{t}'=\gamma(\bar{t}+\mathbf{v}\cdot\bar{\mathbf{r}}/c^2).$$

Let us work in one space dimension for simplicity, $\bar{x}=x/(1-ct/b)$, etc. Now in Fig. 1 we illustrate a projective relation between points $p(x,y)$ and $\bar{p}(\bar{x},\bar{y})$ in intersecting planes, done by rays emanating from a point $q(y_0,z_0)$ external to both. The transformation rule set by the projective link between p, $\bar{p}$, and q is readily calculated to be

$$\bar{x}=\frac{x}{1-y\tan\theta(y_0\tan\theta+z_0)^{-1}},$$

$$\bar{y}=\frac{yz_0/\cos\theta(y_0\tan\theta+z_0)}{1-y\tan\theta(y_0\tan\theta+z_0)^{-1}}.$$

Place q at (b,b) and take $\theta=\frac{1}{2}\pi$ as convenient (not necessary or unique) choices and rename y as ct, and we have a geometrical picture of the homographies $\bar{x}=x/(1-ct/b)$, $\bar{t}=t/(1-ct/b)$. Thence in Fig. 2 we have the substance of the geometrical meaning of the homogeneous ELT: They are ordinary Lorentz transformations of homographic space and time variables $\bar{x}, \bar{t}$, as shown.

The invariant line element

$$(d\bar{s})^2=(d\bar{\mathbf{r}})^2-c^2(d\bar{t})^2,$$

and D'Alembertian

$$\bar{\Box}^2=\frac{\partial^2}{\partial\bar{\mathbf{r}}^2}-\frac{1}{c^2}\frac{\partial^2}{\partial\bar{t}^2}$$

are easily written out in $\mathbf{r}, t$ as pointed out in Paper I.[2] Let us observe further for comparison with later results the structure of the homogeneous-ELT-invariant free-particle action in these primitive variables,

$$L\,dt=\frac{-mc^2}{(1-ct/b)^2}\left(1-\frac{[\mathbf{v}(1-ct/b)+\mathbf{r}c/b]^2}{c^2}\right)^{1/2}dt,$$

leading to the Hamiltonian

$$H=\frac{1}{(1-ct/b)^2}\,[m^2c^4+c^2\mathbf{p}^2(1-ct/b)^2]^{1/2}-\frac{\mathbf{p}\cdot\mathbf{r}c/b}{1-ct/b},$$

or by canonical transformation back to $\bar{\mathbf{r}}=\mathbf{r}/(1-ct/b)$, and to $\bar{\mathbf{p}}=\mathbf{p}(1-ct/b)$, to the action principle

$$\delta\int\bar{\mathbf{p}}\cdot d\bar{\mathbf{r}}-(m^2c^4+c^2\bar{\mathbf{p}}^2)^{1/2}\,d\bar{t}=0.$$

In going to the inhomogeneous ELT one gets the complete perspective of these statements, seeing in particular how the flatness of the $\bar{\mathbf{r}}, \bar{t}$ space—time of homogene-

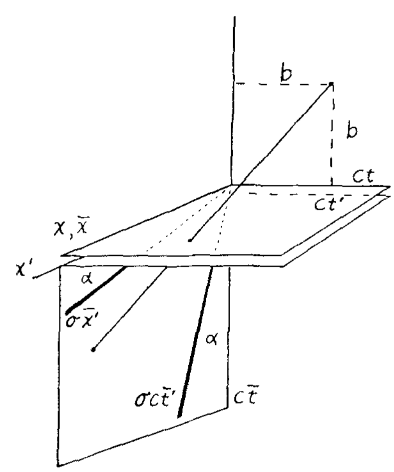


FIG. 2. Diagram showing the geometry of homogeneous ELT. The homographic space and time $\bar{x}$ and $c\bar{t}$ are first obtained projectively as in Fig. 1, with $\theta=\pi/2$ and y called ct. Then in the $\bar{x}$, $c\bar{t}$ plane a Lorentz transformation is conducted to oblique axes $\bar{x}'$ and $c\bar{t}'$ scaled in the usual Lorentz fashion by a scale factor $\sigma=(1+v^2/c^2)^{1/2}/(1-v^2/c^2)^{1/2}$ and with the angle $\alpha=\tan^{-1}v/c$ between $c\bar{t}$, $\sigma c\bar{t}$ and between the $\bar{x}$, $\sigma\bar{x}'$ axes, so that $\bar{x}'=\gamma(\bar{x}-v\bar{t})$ and $\bar{t}'=\gamma(\bar{t}-v\bar{x}/c^2)$. As the center of projection moves off from (b,b) to (∞,∞) the rays it sends off come in parallel and strike the horizontal and vertical planes symmetrically, so that these planes become replicas of one another; then the oblique axes in the vertical plane express the usual OLT.

ous ELT comes about as a limiting case of a more general situation.

THE INHOMOGENEOUS ELT

The nature of the homogeneous ELT is best seen by going to homogeneous coordinates

$$\mathbf{r}=\frac{\mathbf{R}}{U},\quad t=\frac{T}{U};\quad \mathbf{r}'=\frac{\mathbf{R}'}{U'},\quad t'=\frac{T'}{U'},$$

and further introducing

$$\mathbf{R}\equiv X_1, X_2, X_3;\quad icT\equiv X_4;\quad ibU\equiv X_5.$$

Then the homogeneous ELT is

$$\mu\begin{bmatrix}\mathbf{R}'\\ X_4'\\ X_5'\end{bmatrix}=\begin{bmatrix}I+\beta\mathbf{vv} & -i\gamma\mathbf{v} & 0\\ i\gamma\mathbf{v} & \gamma & 0\\ i\gamma\mathbf{v} & \gamma-1 & 1\end{bmatrix}\begin{bmatrix}\mathbf{R}\\ X_4\\ X_5\end{bmatrix}\equiv\Phi\begin{bmatrix}\mathbf{R}\\ X_4\\ X_5\end{bmatrix},$$

where now $\mathbf{v}$ is an abbreviation for $\mathbf{v}/c$, and I is the unit dyadic. The factor μ is an arbitrary nonzero quantity whose significance is merely to instruct that it is the ratios $X_{1,2,3,4}/X_5$ and $X'_{1,2,3,4}/X'_5$ that count—one returns to the fundamental fractional-linear forms by writing

$$\mu X'_a=\Phi_{a\omega}X_\omega\quad (a,\omega=1,\ldots,5),$$

and then μ falls away in

$$\frac{X'_i}{X'_5}=\frac{\Phi_{i\omega}X_\omega}{\Phi_{5\omega}X_\omega}=\frac{\Phi_{i\gamma}X_\gamma/X_5+\Phi_{i5}}{\Phi_{5\gamma}X_\gamma/X_5+\Phi_{55}}\quad (i,\gamma=1,\ldots,4).$$

The split notation above is convenient and unambiguous, and means for instance

$$\mu\mathbf{R}'=(I+\beta\mathbf{vv})\cdot\mathbf{R}-i\gamma\mathbf{v}X_4+0X_5$$

$$\mu X_4'=i\gamma\mathbf{v}\cdot\mathbf{R}+\gamma X_4+0X_5$$

$$\mu X_5'=i\gamma\mathbf{v}\cdot\mathbf{R}+(\gamma-1)X_4+1X_5.$$

The essential features of homogeneous ELT are at once visible from $\mu X_4'$ and $\mu X_5'$ and from the Lorentz rotation nesting in the upper left 4×4 corner of Φ,

$$\mu(X_4'-X_5')=X_4-X_5$$

$$\mu^2(\mathbf{R}'^2+X_4'^2)=\mathbf{R}^2+X_4^2.$$

It is to say that, taking $\mu^2=1$,

$$Q\equiv X_1^2+X_2^2+X_3^2+X_4^2+f^2(X_4-X_5)^2=\text{invariant}$$

under homogeneous ELT, where f^2 is (so far) an arbitrary pure numeric, either positive or negative, whose meaning will appear below. Hereafter we will drop μ.

The rotation hiding here is brought out by a simple similarity transformation

$$\Phi=\begin{bmatrix}I & 0 & 0\\ 0 & 1 & 0\\ 0 & 1 & f^{-1}\end{bmatrix}\begin{bmatrix}I+\beta\mathbf{vv} & -i\gamma\mathbf{v} & 0\\ i\gamma\mathbf{v} & \gamma & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}I & 0 & 0\\ 0 & 1 & 0\\ 0 & -f & f\end{bmatrix}$$

$$\equiv N^{-1}\Phi_L N.$$

That is, $N\mathbf{X}'=\Phi_L N\mathbf{X}$ with Φ_L a Lorentz-like rotation. Hence $(N\mathbf{X})\circ(N\mathbf{X})$ is invariant. This is Q above.

To go to inhomogeneous ELT is but a step along the same path: We take it as a fundamental hypothesis that the generality of ELT is defined through

$$\mathbf{X}'=N^{-1}\{\text{general five-dimensional rotation}\}N\mathbf{X}$$

$$=N^{-1}\mathcal{R}_5 N\mathbf{X}.$$

This interjects into the projective geometry an element ordinarily considered foreign to it, namely a metric. But there is nothing which on principle forbids this, and it is clearly in point physically. The $\mathcal{R}_5$ will have ten free parameters in it owing to the orthonormality of its rows; and it will be evident that, on returning to the primitive fractional-linear transformations in x_α, the inhomogeneous elements therein (corresponding to the space- and time-translations of OLT) become an integral part of the five-dimensional rotation group, not separable from the rest as in OLT.

The five-dimensional rotation group on quite another basis, stemming from de Sitter's cosmological model,[8] has otherwise come under study in both cosmological and elementary particle contexts.[9] As will be seen below, when we consider the differential geometry associated to ELT in ordinary space–time variables x_α as expressed for convenience in a suitable natural set of auxiliary variables $\xi_1(x)$, $\xi_2(x)$, $\xi_3(x)$, $\xi_4(x)$, we obtain a four-space of constant Gaussian curvature that devolves from a flat five-space of metric proportional to

$$d\xi_1^2+d\xi_2^2+d\xi_3^2-|d\xi_4|^2\pm|d\xi_5|^2,$$

$$\xi_5\equiv\xi_5(\xi_1,\xi_2,\xi_3,\xi_4).$$

That is, we have to do with a de Sitter or so-called anti-de Sitter space signed by $(1,1,1,-1,1)$ or by $(1,1,1,-1,-1)$, respectively. In a word, the de Sitter spaces of both kinds may be considered to be undergirded by the projective group of ELT bearing the three intrinsic universals c, b, f and connecting to an extended class of transformations of free-particle motions. The undergirding will also show that the metric of space–time is best understood to be that for a Finsler rather than a Riemannian space, due to an inherent sign-ambiguity when one goes from X_1, X_2, X_3, X_4, X_5 to ξ_1, ξ_2, ξ_3, ξ_4. The still larger group of projective transformations of 24 parameters, containing the inhomogeneous ELT as a subgroup, of course goes beyond any five-dimensional rotation group, offering the most general basis for studying free particles. This will be deferred until ELT has been more fully charted.

Returning to the construction of $\mathcal{R}_5$, a second similarity transformation simplifies its explicit representation,

$$\mathcal{R}_5=\begin{bmatrix}I & 0 & 0\\ 0 & 0 & 1\\ 0 & -1 & 0\end{bmatrix}\mathcal{R}_5^*\begin{bmatrix}I & 0 & 0\\ 0 & 0 & -1\\ 0 & 1 & 0\end{bmatrix}\equiv S^{-1}\mathcal{R}_5^* S.$$

For example, the Lorentzian Φ_L above is

$$\Phi_L=S^{-1}\begin{bmatrix}I+\beta\mathbf{vv} & 0 & -i\gamma\mathbf{v}\\ 0 & 1 & 0\\ i\gamma\mathbf{v} & 0 & \gamma\end{bmatrix}S\equiv S^{-1}\Phi_L^* S.$$

The reason for introducing S, $\mathcal{R}_5^*$, Φ_L^* is as follows. One basic (four-parameter) type of rotation in five-space is

$$\mathcal{R}_5^*(u)=\begin{bmatrix} I+B\mathbf{uu} & B\mathbf{u}u_0 & -iG\mathbf{u}\\ B\mathbf{u}u_0 & 1+Bu_0^2 & iGu_0\\ iG\mathbf{u} & -iGu_0 & G\end{bmatrix},$$

$$G\equiv(1-u_\alpha u_\alpha)^{-1/2},\quad B\equiv(G-1)/u_\alpha u_\alpha\quad(\alpha=0,\ldots,3).$$

It is a five-dimensional counterpart of a Lorentz rotation

$$\begin{bmatrix} I+\beta\mathbf{vv} & -i\gamma\mathbf{v}\\ i\gamma\mathbf{v} & \gamma\end{bmatrix}$$

in ordinary Minkowski four-space. The point is that Φ_L^* above fits into the framework of $\mathcal{R}_5^*(u)$ (with $\mathbf{u}_0, u_0 = \mathbf{v}, 0$), while Φ_L does not. The rotations in the X_4, X_5 plane by $\frac{1}{2}\pi$ and $-\frac{1}{2}\pi$ in S^{-1} and S are needed to get this fit right.

We may now take $\mathcal{R}_5^*(u)$ with $\mathbf{u}, u_0=\mathbf{v},\ i\eta$ to be the first generalization of $\mathcal{R}_5^*$, with η being an abbreviation for $\eta/b=$ (new length parameter)$/b$. In unabbreviated notation

$$G=(1-\mathbf{v}^2/c^2+\eta^2/b^2)^{-1/2}.$$

To get the rest, recall how in four-space the Lorentz transformation is broadened to account for both space rotations and boosts,

$$\mathcal{R}_4=\begin{matrix}\text{General four-rotation}\\ \text{in Minkowski space}\end{matrix}=\begin{bmatrix}\mathcal{R}_3 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix} I+\beta\mathbf{vv} & -i\gamma\mathbf{v}\\ i\gamma\mathbf{v} & \gamma\end{bmatrix}$$

where $\mathcal{R}_3$ is an ordinary three-space rotation, specified, for instance, by Euler angles θ, φ, ψ or say Θ for short. In similar but extended fashion we have to adjoin to $\mathcal{R}_5^*(u)$ a general rotation in four-space. But we already have this in $\mathcal{R}_4$ excepting for replacing $\mathbf{v}$ therein by a new vector triplet,

$$\mathcal{R}_5^*(\Theta;\boldsymbol{\zeta};\mathbf{v},i\eta)=\begin{bmatrix}\mathcal{R}_3 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix} I+A\boldsymbol{\zeta\zeta} & -iF\boldsymbol{\zeta} & 0\\ iF\boldsymbol{\zeta} & F & 0\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix} I+B\mathbf{vv} & iB\mathbf{v}\eta & -iG\mathbf{v}\\ iB\mathbf{v}\eta & 1-B\eta^2 & -G\eta\\ iG\mathbf{v} & G\eta & G\end{bmatrix}.$$

The $\boldsymbol{\zeta}$ here is an abbreviation for (new three-vector length parameter $\boldsymbol{\zeta})/b$ and

$$A=(1-\boldsymbol{\zeta}^2)^{-1/2},\quad F=(A-1)/\boldsymbol{\zeta}^2.$$

Putting everything together, *the general ten-parameter ELT in inhomogeneous coordinates is*

$$\mathbf{X}'=N^{-1}S^{-1}\mathcal{R}_5^*(\Theta;\boldsymbol{\zeta};\mathbf{v},i\eta)\,SN\mathbf{X}.$$

The construction has been done so as to recover the homogeneous ELT for $\mathcal{R}_3=I$ and $\boldsymbol{\zeta},\eta=0$, and so as to produce the *inhomogeneous* ordinary Lorentz transformations when $\mathbf{r}', t'$ are finally expressed fractional-linearly in $\mathbf{r}, t$ and the limit $b\to\infty$ taken. The details and an elaboration of infinitesimal general ELT are set out in Appendix B.

GEOMETRY, DYNAMICS, GROUP ALGEBRA

Having introduced the homogenizing fifth coordinate U we must now get rid of it so as to work in the physical coordinates x_α. The simplest way to do this is to divide the fundamental quadratic form Q by itself (or minus itself),

$$\frac{\mathbf{R}^2-c^2T^2-f^2(bU-cT)^2}{c^2T^2+f^2(bU-cT)^2-\mathbf{R}^2}$$

$$=\text{invariant}=\left[\frac{\mathbf{r}/b}{[ct/b+f^2(1-ct/b)^2-(\mathbf{r}/b)^2]^{1/2}}\right]^2-\left[\frac{ct/b}{[\;]^{1/2}}\right]^2-\left[\frac{f(1-ct/b)}{[\;]^{1/2}}\right]^2$$

or multiplying by f^2b^2 and introducing

$$\boldsymbol{\rho}\equiv\frac{\mathbf{r}/b}{[\;]^{1/2}},\quad \tau\equiv\frac{ct/b}{[\;]^{1/2}},\quad \sigma\equiv\frac{f(1-ct/b)}{[\;]^{1/2}},$$

we have

$$f^2b^2(\rho^2-\tau^2-\sigma^2)=-f^2b^2=\text{invariant}.$$

Also, differentially, it is seen that

$$ds^2=f^2b^2(d\rho^2-d\tau^2-d\sigma^2)$$

is the fundamental metric interval invariant under ELT.

We have arranged things so that for $b\to\infty$ the ds^2 collapses to the line element

$$(ds^2)_{\rm OLT}=dx^2+dy^2+dz^2-c^2\,dt^2$$

of ordinary special relativity. In the limit $f^2\to\infty$ we obtain

$$(d\bar{s})^2=\left[d\left(\frac{\mathbf{r}}{1-ct/b}\right)\right]^2-\left[d\left(\frac{ct}{1-ct/b}\right)\right]^2,$$

which returns us to homogeneous ELT. In this case, but not generally, the space is flattenable as shown.

Keeping to the general case, a Minkowskian $\xi_4\equiv i\tau$ and $\boldsymbol{\rho}\equiv\xi_1, \xi_2, \xi_3$ brings

$$ds^2=f^2b^2(d\xi_1^2+d\xi_2^2+d\xi_3^2+d\xi_4^2-d\sigma^2)$$

or owing to

$$\sigma=(1+\xi_1^2+\xi_2^2+\xi_3^2+\xi_4^2)^{1/2},$$

we have in the space $\xi_1, \xi_2, \xi_3, \xi_4$, which is now the projective counterpart of the ordinary Minkowskian $x_1, x_2, x_3, x_4=ict$, the Riemannian metric

$$ds^2=g_{\alpha\beta}\,d\xi_\alpha\,d\xi_\beta,$$

$$g_{ij}=(\delta_{ij}-\xi_i\xi_j/\sigma^2)f^2b^2,$$

where now repeated Greek indices are summed from 1 to 4. Thence we find

$$g^{ij}=(\delta_{ij}+\xi_i\xi_j)/f^2b^2,$$

$$|g^{\alpha\beta}|=(\sigma/f^4b^4)^2,$$

$$\begin{bmatrix} ij\\ k\end{bmatrix}=\left(\frac{\xi_i\xi_j\xi_k}{\sigma^4}-\frac{\delta_{ij}\xi_k}{\sigma^2}\right)f^2b^2,$$

$$\begin{Bmatrix} ij\\ k\end{Bmatrix}=\frac{\xi_i\xi_j\xi_k}{\sigma^2}-\delta_{ij}\xi_k.$$

For the D'Alembertian or wave operator there follows

$$\boxdot^2 = g^{\alpha\beta}\left(\frac{\partial^2}{\partial\xi_\alpha\partial\xi_\beta} - \begin{Bmatrix}\alpha\beta\\ \gamma\end{Bmatrix}\frac{\partial}{\partial\xi_\gamma}\right)$$
$$= \frac{1}{f^2b^2}\left[(\delta_{\alpha\beta}+\xi_\alpha\xi_\beta)\frac{\partial^2}{\partial\xi_\alpha\partial\xi_\beta} + 4\xi_\gamma\frac{\partial}{\partial\xi_\gamma}\right].$$

The covariant curvature tensor is

$$R_{mijk} = f^2b^2\left[\delta_{mk}\delta_{ij} - \delta_{mj}\delta_{ik} + \frac{\xi_i\xi_k}{\sigma^2}\delta_{mj} - \frac{\xi_i\xi_j}{\sigma^2}\delta_{mk} + \frac{\xi_m\xi_j}{\sigma^2}\delta_{ik} - \frac{\xi_m\xi_k}{\sigma^2}\delta_{ij}\right],$$

and is the same as

$$-[g_{mj}g_{ik} - g_{mk}g_{ij}]/f^2b^2.$$

Hence the space is of constant Gaussian curvature $K = -1/f^2b^2$, as was clear from the beginning, the space being simply the surface of a type of five-dimensional sphere (going flat for $f^2 \to \infty$). We can in fact introduce polar coordinates $\theta, \psi, \chi, \varphi$ in the five-space to hold us to this surface,

$$\xi_1 \sim \cos\theta,$$
$$\xi_2 \sim \sin\theta\cos\psi,$$
$$\xi_3 \sim \sin\theta\sin\psi\cos\chi,$$
$$\xi_4 \sim \sin\theta\sin\psi\sin\chi\cos\varphi,$$
$$\xi_5 \equiv i\sigma \sim \sin\theta\sin\psi\sin\chi\sin\varphi,$$

and obtain (χ being imaginary)

$$ds^2 = f^2b^2(d\theta^2 + \sin^2\theta\,d\psi^2 + \sin^2\theta\sin^2\psi\,d\chi^2 + \sin^2\theta\sin^2\psi\sin^2\chi\,d\varphi^2).$$

The geodesics follow from

$$\delta\int ds \sim \delta\int\left[1 - \left(\frac{d\boldsymbol{\rho}}{d\tau}\right)^2 + \left(\frac{d\sigma}{d\tau}\right)^2\right]^{1/2} d\tau$$
$$\equiv \delta\int \Lambda\,d\tau = 0,$$
$$\frac{d\sigma}{d\tau} \equiv \dot\sigma = \frac{1}{\sigma}\left(\boldsymbol{\rho}\cdot\frac{d\boldsymbol{\rho}}{d\tau} - \tau\right),$$

which in dynamical terms are the extremals for the Lagrangian

$$L = -m\Lambda,$$

$m \equiv$ reduced mass parameter $\equiv$ (particle mass) $\cdot\, cfb$.

It is surprisingly cumbersome to bring out the rudimentary fact that they are the straights $d^2\mathbf{r}/dt^2 = 0$. A simpler way is to write

$$t^* = \alpha\tau + \gamma\sigma$$

such that

$$(dt^*)^2 = d\tau^2 + d\sigma^2$$

under

$$\alpha^2 = 1, \quad \gamma^2 = 1, \quad \alpha\gamma + \gamma\alpha = 0,$$

viz., t^* is a little spinor, a sort of compound or twisting time. Then immediately

$$\frac{d^2\boldsymbol{\rho}}{dt^{*2}} = 0,$$

$$\frac{d\boldsymbol{\rho}}{dt^*} = \mathbf{k}^* = \alpha\mathbf{k}_1 + \gamma\mathbf{k}_2,$$
$$\boldsymbol{\rho} = \tfrac{1}{2}(\mathbf{k}^*t^* + t^*\mathbf{k}^*) = \mathbf{k}_1\tau + \mathbf{k}_2\sigma,$$
$$\mathbf{r} = \mathbf{k}_1 ct + \mathbf{k}_2 f(b - ct) = \mathbf{K}_1 t + \mathbf{K}_2,$$

where in the third line we naturally have had to take the symmetrized sum, as the integration gives ambiguously $\mathbf{k}^*t^*$ or $t^*\mathbf{k}^*$.

The null lines similarly appear,

$$\left(\frac{d\boldsymbol{\rho}}{dt^*}\right)^2 = 1,$$
$$\frac{d\boldsymbol{\rho}}{dt^*} = \mathbf{n}^*,$$
$$\boldsymbol{\rho} = \tfrac{1}{2}(\mathbf{n}^*t^* + t^*\mathbf{n}^*) = \mathbf{n}_1\tau + \mathbf{n}_2\sigma,$$
$$\mathbf{r} = \mathbf{n}_1 ct + \mathbf{n}_2 f(b - ct) = (\mathbf{n}_1 - f\mathbf{n}_2)ct + \mathbf{n}_2 fb,$$
$$(\mathbf{n}_1^2 + \mathbf{n}_2^2 = 1).$$

Writing

$$\mathbf{n}_1 = \hat{a}_1\cos\delta, \quad \mathbf{n}_2 = \hat{a}_2\sin\delta,$$

we have that the null rays are

$$\mathbf{r} = (\hat{a}_1\cos\delta - f\sin\delta\,\hat{a}_2)ct + \hat{a}_2 fb\sin\delta,$$

and travel with speed

$$\mathbf{u}^2 \equiv \left(\frac{d\mathbf{r}}{dt}\right)^2 = c^2[1 + \sin^2\delta(f^2 - 1) - 2f\sin\delta\cos\delta\cos\theta],$$

where θ is the angle between $\hat{a}_1$ and $\hat{a}_2$. The rays therefore are in anisotropic and inhomogeneous flight, knowing in their speed the starting point $\hat{a}_2 fb\sin\delta \equiv \hat{a}_2 A_2$ and their orientation with respect to this starting point. For f^2 negative, δ must be taken imaginary. In the OIF limit, $b \to \infty$, we must understand that finite arbitrary starting points $A_2\hat{a}_2$ have to be allowed, requiring $\delta \to 0$, so that $\mathbf{u}^2 \to c^2$.

Let us go back to dynamical language and calculate the Hamiltonian consequent from L. Calling $\mathbf{w}$ the velocity $d\boldsymbol{\rho}/d\tau$, the momentum conjugate to $\boldsymbol{\rho}$ is

$$\mathbf{p} = \frac{\partial L}{\partial\mathbf{w}} = \frac{m}{\Lambda}\left(\mathbf{w} - \frac{\dot\sigma}{\sigma}\boldsymbol{\rho}\right).$$

Inversely,

$$\mathbf{w} = \frac{\Lambda}{m}\left(\mathbf{p} + \boldsymbol{\rho}\frac{\mathbf{p}\cdot\boldsymbol{\rho}}{1 - \tau^2}\right) - \frac{\tau}{1 - \tau^2}\boldsymbol{\rho},$$

$$\Lambda = [1 - \mathbf{w}^2 + \dot\sigma^2]^{1/2}$$
$$= [1 - \tau^2 + (1 - \tau^2)\mathbf{p}^2/m^2 + (\mathbf{p}\cdot\boldsymbol{\rho})^2/m^2]^{-1/2},$$

and in the ordinary way the Hamiltonian works out to be

$$H = \frac{m}{1 - \tau^2}[1 - \tau^2 + (1 - \tau^2)\mathbf{p}^2/m^2 + (\mathbf{p}\cdot\boldsymbol{\rho})^2/m^2]^{1/2} - \frac{\tau}{1 - \tau^2}\mathbf{p}\cdot\boldsymbol{\rho}.$$

Under the canonical transformation

$$\bar{\boldsymbol{\rho}} = \frac{\boldsymbol{\rho}}{(1 - \tau^2)^{1/2}}, \quad \bar{\mathbf{p}} = \mathbf{p}(1 - \tau^2)^{1/2},$$

we obtain the new Hamiltonian

$$\bar{H}=\frac{m}{1-\tau^2}[1-\tau^2+\bar{\mathbf{p}}^2/m^2+(\bar{\mathbf{p}}\cdot\bar{\rho})^2/m^2]^{1/2},$$

or with $\tau=\tanh\bar{\tau}$ we have

$$\delta\int\bar{\mathbf{p}}\cdot d\rho-\bar{H}'\,d\bar{\tau}=0,$$

$$\bar{H}'=[\bar{\omega}^2+\bar{\mathbf{p}}^2+(\bar{\mathbf{p}}\cdot\bar{\rho})^2]^{1/2},$$

$$\bar{\omega}\equiv m\,\mathrm{sech}\bar{\tau}.$$

Now the infinitesimal transformations in ρ,τ variables of the ELT group are for infinitesimal ϵ, λ, ν, θ:

time-translation type

$$\tau\to\tau+\epsilon\sigma,$$

$$\rho\to\rho\quad(\sigma\to\sigma-\epsilon\tau),$$

space-translation type

$$\tau\to\tau,$$

$$\rho\to\rho+\lambda\sigma\quad(\sigma\to\sigma+\rho\cdot\lambda),$$

Lorentz-rotation type

$$\tau\to\tau+\nu\cdot\rho,$$

$$\rho\to\rho+\nu\tau\quad(\sigma\to\sigma),$$

space-rotation type

$$\tau\to\tau,$$

$$\rho\to\rho-\theta\cdot(I\times\rho)\quad(\sigma\to\sigma),$$

as follows directly from the primitive infinitesimal transformations in $\mathbf{r},t$ variables given in Appendix B. The critical difference from OLT lies in the infinitesimal addenda $\epsilon\sigma$, $\lambda\sigma$ to the translations, which in OLT are, instead of type ϵ, λ; the five-rotational character of the transformations is reflected in the transformations for σ as well as ρ and τ. All these transformations of course leave $L\,d\tau$ invariant, and accordingly by Noether's theorem we find the corresponding conservation laws:

$$E=\frac{m}{\Lambda}(\sigma-\tau\dot{\sigma})=\sigma H,$$

$$\mathbf{P}=\frac{m}{\Lambda}(\sigma\mathbf{w}-\dot{\sigma}\rho)=\sigma\mathbf{p},$$

$$\mathbf{K}=\frac{m}{\Lambda}(\rho-\mathbf{w}\tau)=H\rho-\tau\mathbf{p},$$

$$\mathbf{L}=\rho\times\frac{m}{\Lambda}\left(\mathbf{w}-\frac{\dot{\sigma}}{\sigma}\rho\right)=\rho\times\mathbf{p}.$$

This brings us also to the group algebra. The Lie generators corresponding to the infinitesimal ELT are

$$\mathcal{T}=\sigma\frac{\partial}{\partial\tau},$$

$$\mathcal{P}_i=\sigma\frac{\partial}{\partial\xi_i},$$

$$\mathcal{K}_i=\tau\frac{\partial}{\partial\xi_i}+\xi_i\frac{\partial}{\partial\tau},$$

$$\mathcal{L}_i=-\left(\rho\times\frac{\partial}{\partial\rho}\right)_i\quad(i=1,2,3).$$

Then the Lie commutator brackets come out,

$$(\mathcal{T},\mathcal{K}_i)=\mathcal{P}_i,\quad(\mathcal{K}_i,\mathcal{K}_j)=-\epsilon_{ijk}\mathcal{L}_k,$$

$$(\mathcal{T},\mathcal{P}_i)=-\mathcal{K}_i,\quad(\mathcal{K}_i,\mathcal{P}_j)=-\mathcal{T}\delta_{ij},$$

$$(\mathcal{T},\mathcal{L}_i)=0,\quad(\mathcal{K}_i,\mathcal{L}_j)=\epsilon_{ijk}\mathcal{K}_k,$$

$$(\mathcal{P}_i,\mathcal{P}_j)=-\epsilon_{ijk}\mathcal{L}_k,$$

$$(\mathcal{P}_i,\mathcal{L}_j)=\epsilon_{ijk}\mathcal{P}_k,$$

$$(\mathcal{L}_i,\mathcal{L}_j)=\epsilon_{ijk}\mathcal{L}_k,$$

and are distinguished from OLT in the failure of commutability of any pairs of $\mathcal{T}$, $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$, again reflecting that the translations are a species of rotation. An important consequence is that the conserved momentum components $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$ cannot be made the canonical momenta in any Hamiltonian framework starting from ρ as canonical, as is otherwise evident from $\mathbf{P}=\sigma\mathbf{p}$.

Let us return to the fundamental differential invariant ds^2, involving $d\rho=d(\mathbf{r}/b)/[Z]^{1/2}$, etc. We see that when

$$Z(x,y,z,t)\equiv\left(\frac{ct}{b}\right)^2+f^2\left(1-\frac{ct}{b}\right)^2-\left(\frac{\mathbf{r}}{b}\right)^2$$

$$\equiv u^2+f^2(1-u)^2-\left(\frac{\mathbf{r}}{b}\right)^2$$

goes from positive to negative values, ds^2 does so too. The geodesics, the free particles, however are all the straights $d^2\mathbf{r}/dt^2=0$ and ignorantly pass through the surface $Z=0$ without hindrance; all of space—time is traversed by them. The quadric *separator surface* $Z=0$ may be rewritten

$$\left(\frac{r}{b}\right)^2-\frac{f^2}{1+f^2}=(1+f^2)\left(u-\frac{f^2}{1+f^2}\right)^2,$$

and we may call the "interior" or "exterior" the regions $Z>0$ or $Z<0$, correspondingly replacing the equal sign in the preceding equation by $<$ or $>$. The separator is entirely controlled by f^2 as shown in Fig. 3.

The only simple way to avoid the jump in ds^2 from (+) to (−) in, say, an infinitesimal step from the interior to the exterior across the separator, is to square ds^2,

$$ds^4=f^4b^4(d\rho^2-d\tau^2-d\sigma^2)^2.$$

This means we have to work in a Finsler space[10] rather than a Riemannian space, but a very simple one, loosely the "square" of a Riemannian space.

In dynamical terms we now have

$$\delta\int L_F\,d\tau=0,$$

$$L_F\equiv-m[(1-\mathbf{w}^2+\dot{\sigma}^2)^2]^{1/4}\equiv-m[\Omega^2]^{1/4},$$

whereupon

$$\mathbf{p}\equiv\frac{\partial L_F}{\partial\mathbf{w}}=\frac{m(\mathbf{w}-(\dot{\sigma}/\sigma)\rho)\Omega}{(\Omega^2)^{3/4}},$$

and then

$$\Omega=\{1-\tau^2\pm[(1-\tau^2)\mathbf{p}^2/m^2+(\mathbf{p}\cdot\rho)^2/m^2]\}^{-1},$$

so that the Hamiltonian becomes

$$H_F=\frac{m}{1-\tau^2}\left\{\left[1-\tau^2\pm\left((1-\tau^2)\frac{\mathbf{p}^2}{m^2}+\frac{(\mathbf{p}\cdot\rho)^2}{m^2}\right)\right]^2\right\}^{1/4}-\frac{\tau}{1-\tau^2}\mathbf{p}\cdot\rho.$$

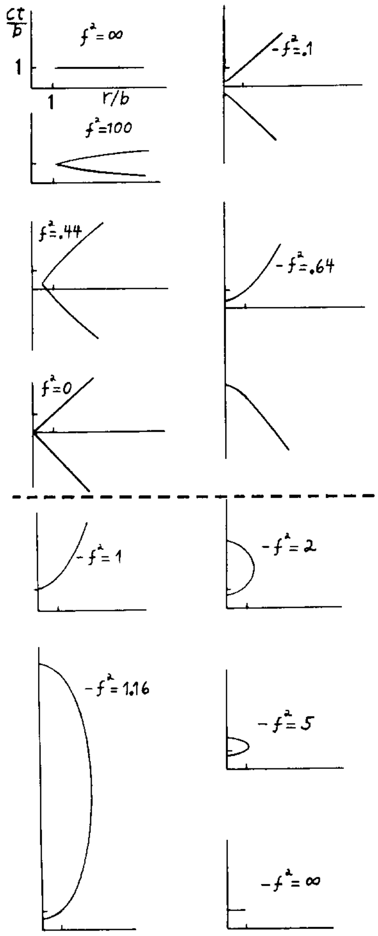


FIG. 3. Geometry of the separator surface for diffierent f^2. Swing each figure around the vertical (ct/b) axis for a representation of $Z(x,y,t)=0$. For $f^2 \geq 0$ the values of f^2 are written in the "interior" region, while for $f^2<0$ they are written in the "exterior" region. The intersecting hyperboloidal sheets for $f^2>0$ are collapsed into a disc having a hole through it, for $f^2 \to \infty$, and go over with decreasing f^2 to form a cone for $f^2=0$. Then as f^2 turns negative, separate upper and lower hyperboloidal sheets are formed, with the lower sheet sinking away to $-\infty$ as $f^2 \to -1$, while the upper sheet becomes a paraboloid. As f^2 falls below -1, the upper sheet bends round to close and become a prolate spheroid, then a sphere, and then an oblate spheroid which flattens finally for $f^2 \to -\infty$ to a disc occupying the hole in the starting $f^2=\infty$ disc.

Simplifying by the same canonical transformation to $\bar{\rho}, \bar{\mathbf{p}}$ as before, and again introducing $\bar{\tau}$, $\bar{\omega}$,

$$\bar{H}_F = \{(\bar{\omega}^2 \pm [\bar{\mathbf{p}}^2 + (\bar{\mathbf{p}}\cdot\bar{\rho})^2])^2\}^{1/4}.$$

That is, $\bar{H}_F$ has eight branches

$$\bar{H}_F = \pm (\bar{\omega}^2 \pm [\bar{\mathbf{p}}^2 + (\bar{\mathbf{p}}\cdot\bar{\rho}^2])^{1/2} \pm i(\bar{\omega}^2 \pm [\bar{\mathbf{p}}^2 + (\bar{\mathbf{p}}\cdot\bar{\rho})^2])^{1/2},$$

being the roots of the biquartic

$$\{\bar{H}_F^4 - \bar{\omega}^4 - [\bar{\mathbf{p}}^2 + (\bar{\mathbf{p}}\cdot\bar{\rho})^2]^2\}^2 = 4\bar{\omega}^4[\bar{\mathbf{p}}^2 + (\bar{\mathbf{p}}\cdot\bar{\rho})^2]^2.$$

The more extended and simplified theory of a free particle based on the Finsler-type metric will be discussed in a forthcoming paper.

CONCLUSION

In summary, it has been shown that there is a unique extension of the Lorentz group from being a ten-parameter sector of the affine transformations to being (in one-to-one correspondence) a ten-parameter sector of the projective transformations, each sector preserving equally the uniform straight line motion of a free particle, with the projective extension of Lorentz transformations containing besides c a new universal length constant b and going over to the Lorentz group for $b \to \infty$, and with the projective extension coming from the group of rotations in the five-dimensional space of homogeneous space—time coordinates.

A main question, on which hinges possible observational tests, plainly is the extension of the structure of field physics, especially electrodynamics, so as to make them extended-Lorentz covariant. The general outlines toward this are already clearly imprinted in the extended space—time structure itself. In the case of electrodynamics, one naturally starts with a projective type of five-potential $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$, $\mathcal{A}_4$, $\mathcal{A}_5$ and current density J_1, J_2, J_3, J_4, J_5 that are posited to transform as do the homogeneous coordinates X_1, X_2, X_3, X_4, X_5 of space—time $[(N\mathcal{A})^2$ and $(N\mathbf{J})^2$ invariant$]$ and proceeds to get rid of the homogenizing $\mathcal{A}_5$ and to go to $\mathbf{A}$, $i\varphi$ plus a nonindependent fifth $i\psi$ or $\bar{A}_1$, $\bar{A}_2$, $\bar{A}_3$, $\bar{A}_4$, $\bar{A}_5$ with $\Sigma \bar{A}_i^2$ fixed, as in the transition earlier to $\rho, i\tau, i\sigma = \xi_1, \xi_2, \xi_3, \xi_4, \xi_5$ with $\Sigma \xi_i^2$ fixed. Similarly, J_5 is disposed of by going to a nonindependent quintet $\bar{j}_1$, $\bar{j}_2$, $\bar{j}_3$, $\bar{j}_4$, $\bar{j}_5$:

$$\bar{\mathbf{A}}, \bar{A}_4, \bar{A}_5 = \left\{\frac{\mathbf{A}}{[\varphi^2 + f^2(\kappa b - \varphi)^2 - \mathbf{A}^2]^{1/2}}, \frac{i\varphi}{[\;]^{1/2}}, \frac{if(\kappa b - \varphi)}{[\;]^{1/2}}\right\} \times f\kappa b$$

$$\bar{\mathbf{j}}, \bar{j}_4, \bar{j}_5 = \left\{\frac{\mathbf{j}}{[(c\rho)^2 + f^2(\nu b - c\rho)^2 - \mathbf{j}^2]^{1/2}}, \frac{ic\rho}{[\;]^{1/2}}, \frac{if(\nu b - c\rho)}{[\;]^{1/2}}\right\} \times f\nu b.$$

Besides the nonlinear composition of field and source quantities, it is noticeable that one more specifically electromagnetic constant (apart from c and b) in addition to charge is required, for the sake of integrity of physical dimensions. Namely, we have written $\mathcal{A}_5 = ib\kappa V$, $J_5 = ib\nu W$ to dehomogenize to

$$\frac{\mathcal{A}}{V}, \quad \frac{\mathcal{A}_4}{V} \equiv \mathbf{A}, \qquad i\varphi = (\text{charge/length}),$$

$$\frac{\mathbf{J}}{W}, \quad \frac{J_4}{W} \equiv \mathbf{j}, \qquad ic\rho = (\text{charge/length}^2 \cdot \text{time}),$$

and we must accordingly require

$$\kappa = (\text{charge/length}^2),$$

$$\nu = (\text{charge/length}^3 \cdot \text{time}).$$

Here the (length) cannot be taken to be b as it is necessary that κ, ν remain of nonzero size for $b \to \infty$, in order that extended electrodynamics fall back into ordinary Maxwell theory in this limit. Hence, one further intrinsic electromagnetic quantity besides charge must enter, say an electromagnetic length Δ, so that κ, ν can, for instance, be written as numerical multiples of e/Δ^2, ec/Δ^4. The pure electromagnetic field, corresponding to that of the source-free Maxwell equations, recognizes interiorly the existence of charge, whereas ordinary source-free electrodynamics has internally to do exclusively with c. That is, besides c and b inherent in space—time, the field knows κ and the source ν, and electrodynamics rests on two essential internal parameters.

The ELT covariant electrodynamics may be formulated on the base of constrained ξ_i, $\bar{A}_i$, $\bar{j}_i$, and in other ways, as will be discussed in a report in preparation.

ACKNOWLEDGMENT

My gratitude goes to Professor Mark Kac for his steady encouragement in this work.

APPENDIX A

To see the kernel of the simple proof that the projective transformations are *necessary* for preservation of straights, let us consider the one-dimensional case and first allow an undetermined functional connection between x', t' and x, t,

$$x' = F(x,t), \quad t' = G(x,t).$$

Then

$$v' = \frac{dx'}{dt'} = \frac{DF}{DG},$$

$$D \equiv \dot{x}\,\frac{\partial}{\partial x} \equiv v\,\frac{\partial}{\partial x} + \frac{\partial}{\partial t},$$

and secondly

$$\frac{dv'}{dt'} = \left[\frac{dv}{dt}\,\frac{\partial}{\partial v}\left(\frac{DF}{DG}\right) + D\left(\frac{DF}{DG}\right)\right]\frac{dt}{dt'}.$$

For $dv/dt = 0$ to imply $dv'/dt' = 0$ and reversely, means requiring

$$D\left(\frac{DF}{DG}\right) = 0,$$

where in D the velocity parameter v is constant but arbitrary. The general first integral is

$$\frac{DF}{DG} \equiv \frac{vF_x + F_t}{vG_x + G_t} = M(x - vt, v).$$

Since the left side is fractional-linear in v, M must be so too,

$$M = \frac{\alpha_1(x - vt) + \alpha_2 v + \alpha_3}{a_1(x - vt) + a_2 v + a_3} = \frac{(\alpha_2 - \alpha_1 t)\,v + \alpha_1 x + \alpha_3}{(a_2 - a_1 t)\,v + a_1 x + a_3} \equiv \frac{M_1}{M_2}.$$

Moreover, the Jacobian $F_x G_t - F_t G_x$ of the transformation is not to vanish, so neither must

$$(\alpha_2 - \alpha_1 t)(a_1 x + a_3) - (a_2 - a_1 t)(\alpha_1 x + \alpha_3).$$

Going to

$$DF = MDG$$

and integrating again,

$$F - MG = E(x - vt, v).$$

Once more the left side is fractional-linear in v so E has to be of the form

$$E = \frac{\beta_1(x - vt) + \beta_2 v + \beta_3}{b_1(x - vt) + b_2 v + b_3} = \frac{(\beta_2 - \beta_1 t)\,v + \beta_1 x + \beta_3}{(b_2 - b_1 t)\,v + b_1 x + b_3} = \frac{E_1}{E_2}.$$

Thence,

$$\frac{FM_2 - GM_1}{M_2} = \frac{E_1}{E_2}.$$

On cross multiplying and equating coefficients of like powers of the arbitrary v, one sees after a short calculation that F and G are wholly restricted to being general fractional-linear forms with common denominators. The proof in three dimensions goes the same way.

APPENDIX B

Dropping the three-space rotation, we find by multiplying $N^{-1}S^{-1}R_4(\boldsymbol{\zeta})R_5^*(u)\,SN$, and returning to $\mathbf{r}', t'$, $\mathbf{r}, t$ variables, the inhomogeneous ELT in fractional-linear form,

$$\begin{aligned}
\mathbf{r}' = \{&[\varphi_\zeta \cdot \varphi_v + BF\eta\boldsymbol{\zeta}\mathbf{v}] \cdot \mathbf{r} \\
&+ [(G - fB\eta)\,\varphi_\zeta \cdot \mathbf{v} + (fF(1 - B\eta^2) - FG)\,\boldsymbol{\zeta}]\,ct \\
&+ [B\eta\varphi_v \cdot \mathbf{v} - F(1 - B\eta^2)]fb\}/\Delta,
\end{aligned}$$

$$t' = \{G(1 + f\eta)\,t + G\mathbf{v}\cdot\mathbf{r}/c - fbG\eta/c\}/\Delta,$$

$$\begin{aligned}
\Delta \equiv &[F(1 - B\eta^2) - FB\eta\mathbf{v}\cdot\boldsymbol{\zeta} - fG\eta] + [(fG - FB\eta)\,\mathbf{v} \\
&- F\boldsymbol{\zeta}\cdot\varphi_v]\cdot\mathbf{r}/fb + [G(1 + f\eta) - F(1 - B\eta^2) \\
&+ FB\eta\mathbf{v}\cdot\boldsymbol{\zeta} - f^{-1}FG(\boldsymbol{\zeta}\cdot\mathbf{v} - \eta)]\,ct/b
\end{aligned}$$

$$(\varphi_\zeta \equiv 1 + A\,\boldsymbol{\zeta}\boldsymbol{\zeta}, \quad \varphi_v \equiv 1 + B\mathbf{v}\mathbf{v}),$$

where $\mathbf{v}$, $\boldsymbol{\zeta}$, η are the abbreviated ones standing for $\mathbf{v}/c$, $\boldsymbol{\zeta}/b$, η/b. When f is imaginary, $\boldsymbol{\zeta}$ and η are to be taken imaginary.

For $b \to \infty$ we obtain (β and γ having their OLT meanings)

$$\mathbf{r}' = \varphi_v \cdot \mathbf{r} + \gamma\mathbf{v}t + f\eta\beta c\mathbf{v} - f\boldsymbol{\zeta},$$

$$t' = \gamma(t + \mathbf{v}\cdot\mathbf{r}/c^2) - f\gamma\eta/c,$$

here (and below) in *nonabbreviated* quantities, $\mathbf{v}$ = velocity, $\boldsymbol{\zeta}$ = displacement, η = length. Hence the recovery of the inhomogeneous Lorentz transformations.

The separate infinitesimal transformations are:

time-translational type (η infinitesimal, $\boldsymbol{\zeta}$ and $\mathbf{v}$ zero)

$$\mathbf{r}' = \mathbf{r} - \mathbf{r}\,\frac{\eta}{bf}\,[(1 + f^2)\,u - f^2],$$

$$u' = u - \frac{\eta}{bf}\,[f^2(u - 1)^2 + u^2];$$

space-translational type ($\boldsymbol{\zeta}$ infinitesimal, η and $\mathbf{v}$ zero)

$$\mathbf{r}' = \mathbf{r} + \mathbf{r}\,f^{-1}\,\frac{\boldsymbol{\zeta}\cdot\mathbf{r}}{b^2} + f\,\boldsymbol{\zeta}(u - 1),$$

$$u' = u + uf^{-1}\,\frac{\boldsymbol{\zeta}\cdot\mathbf{r}}{b^2};$$

Lorentz-rotation type ($\mathbf{v}$ infinitesimal, $\boldsymbol{\zeta}$ and η zero)

$$\mathbf{r}' = \mathbf{r} - \frac{1}{bc}\,\mathbf{r}\,\mathbf{v}\cdot\mathbf{r} + \mathbf{v}\,\frac{b}{c}\,u,$$

$$u' = u - \frac{\mathbf{v}\cdot\mathbf{r}}{bc}\,(u - 1),$$

where u stands for ct/b. The infinitesimals ϵ, $\boldsymbol{\lambda}$, $\boldsymbol{\nu}$ in the text are $-\eta/b$, $-\boldsymbol{\zeta}/b$, $\mathbf{v}/c$.

[1]James Clerk Maxwell, *Matter and Motion* (Macmillan, New York, 1920).
[2]E.H. Kerner, Proc. Natl. Acad. Sci. **73**, 1418 (1976), referred to as Paper I.
[3]J.A. Schouten, *Ricci Calculus* (Springer, Berlin, 1954).
[4]O. Veblen, *Projektive Relativitätstheorie* (Ergebnisse der mathematik und ihrer grenzgebiete) (Springer, Berlin, 1933).
[5]Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss. Berlin **54**, 966 (1921).
[6]O. Veblen and J.M. Thomas, Ann. Math. **27**, 279 (1926).
[7]V. Fock, *The Theory of Space, Time, and Gravitation* (Pergamon, Oxford, 1964), Chapter 1 and Appendix A.
[8]W. de Sitter, Mon. Not. R. Astron. Soc. **78**, 3 1917; H.P. Robertson and T.W. Noonan, *Relativity and Cosmology* (Saunders, Philadelphia, 1968).
[9]T.O. Philips and E.P. Wigner, in *Group Theory and Its Applications*, edited by E.M. Loebl (Academic, New York, 1968); J. Patera, P. Winternitz, and H. Zassenhaus, J. Math. Phys. **17**, 717 (1976).
[10]H. Rund, *The Differential Geometry of Finsler Spaces* (Springer, Berlin, 1959), It is interesting that Riemann in his 1854 Habilitätionsvortrag already considered $ds^4 = dx_1^4 + dx_2^4$ (see Chap. 1).

The high conductivity limit in mean field electrodynamics

F. Krause and P. H. Roberts

Zentralinstitut für Astrophysik, Potsdam, Germany and University of Newcastle upon Tyne, England
(Received 15 October 1975)

An attempt is made to resolve conflicting views on the reliability of first order smoothing theory when applied to electromagnetic induction of magnetic field by a turbulently moving conductor, in the astrophysically most interesting case of large microscale magnetic Reynolds numbers.

The ensemble mean, $\langle\mathbf{B}\rangle$, of the magnetic field, $\mathbf{B}$, in a uniform conducting fluid moving with the nonrelativistic velocity, $\mathbf{u}'$, is governed by the mean induction equation[1]

$$\partial\langle\mathbf{B}\rangle/\partial t=\nabla\times\mathcal{E}+\eta\nabla^2\langle\mathbf{B}\rangle, \tag{1}$$

where $\mathcal{E}=\langle\mathbf{u}'\times\mathbf{B}'\rangle$ and $\mathbf{B}'=\mathbf{B}-\langle\mathbf{B}\rangle$ is the fluctuating part of $\mathbf{B}$. It is governed by

$$\partial\mathbf{B}'/\partial t=\nabla\times(\mathbf{u}'\times\langle\mathbf{B}\rangle+\mathbf{G}')+\eta\nabla^2\mathbf{B}', \tag{2}$$

where $\mathbf{G}'=\mathbf{u}'\times\mathbf{B}'-\mathcal{E}$. It is the objective of mean field electrodynamics to use (2) to express $\mathcal{E}$ as a linear functional of $\langle\mathbf{B}\rangle$. Then (1) is a closed equation for the mean field, which may be studied in isolation from $\mathbf{B}'$.

The computation of $\mathcal{E}$ is simplest when $\mathbf{G}'=0$, an approximation sometimes called "first order smoothing theory," or FOST for short. Roberts and Soward recently attempted to delineate parameter ranges in which FOST might be safely used.[2-4] The four terms in (2) are, respectively, of order B'/T, $U\langle B\rangle/L$, UB'/L, $\eta B'/L^2$, where L and T are length and time scales typical of $\mathbf{B}'$, and U is the rms turbulent velocity. It might at first sight be thought that $G'(\doteq UB'/L)$ would be negligible if[5,6]

$$\epsilon_1\equiv UL/\eta\ll 1,\quad\text{or } \epsilon_2\equiv UT/L\ll 1, \tag{3}$$

(or both), and that $B'/\langle B\rangle$ would then be of order ϵ_1 or ϵ_2, respectively. In particular FOST would seem to be valid in the astrophysically interesting case of large microscale magnetic Reynolds number $U^2T/\eta(=\epsilon_1\epsilon_2)$ provided $\epsilon_2\ll 1$. Roberts and Soward saw however, that in this case $B'/\langle B\rangle$ would, no matter how small initially, be ultimately[7] of order $\epsilon_1\epsilon_2(\gg 1)$, and that the contribution to (2) of $\mathbf{G}'$ would then not be negligible compared with that of $\mathbf{u}'\times\langle\mathbf{B}\rangle$. They argued that FOST was justified[8] only when $\epsilon_1\ll 1$ and $\epsilon_1\epsilon_2\ll 1$, and implied that the excellent account given by FOST of, for example, solar electrodynamics was fortuitous (see, for instance, Refs. 9 and 10).

It seems to us that Roberts and Soward were mistaken, and for a somewhat subtle reason, adumbrated earlier by Steenbeck and Krause.[9] It is true[7] that, when $\epsilon_2\ll 1$ and $\epsilon_1\epsilon_2\gg 1$, the rms magnitude of $\mathbf{B}'$ is large compared with $\langle\mathbf{B}\rangle$, but this is irrelevant. In computing $\mathcal{E}$ it is only $\mathbf{B}'_{\rm cor}$, the part of $\mathbf{B}'$ correlated with $\mathbf{u}'$, that is required, and this is of order $\epsilon_2\langle\mathbf{B}\rangle$, i.e., it is small compared with $\langle\mathbf{B}\rangle$. Thus FOST is reinstated for ranges (3), broader than those envisaged by Roberts and Soward. As an unintentional side effect, their area of disagreement with Lerche and Parker is exacerbated.[2-4] The main issue is of course the astrophysical one. The remainder of this note presents the argument, in a more complete form than Steenbeck and Krause.[9]

In the interests of economy, the space arguments will be suppressed, and $\mathbf{x}$, ξ, $\mathbf{x}+\xi$, $\xi-\xi'$, etc. should be understood below to be respectively implied by time arguments t, τ, $t+\tau$, $\tau-\tau'$, etc. We will use the summation convention, and the abbreviations

$$\Delta=\partial/\partial\tau-\eta\nabla^2,\quad \nabla_i=\partial/\partial\xi_i.$$

We wish to contrast (2) with two equations obtained from it:

$$\Delta\langle u_i'(t)B_j'(t+\tau)\rangle-\epsilon_{jkl}\epsilon_{lmn}\nabla_k\langle u_i'(t)u_m'(t+\tau)B_n'(t+\tau)\rangle$$
$$=\epsilon_{jkl}\epsilon_{lmn}\nabla_k[\langle u_i'(t)u_m'(t+\tau)\rangle\langle B_n(t+\tau)\rangle], \tag{4}$$

$$\Delta\langle B_i'(t)B_j'(t+\tau)\rangle-\epsilon_{jkl}\epsilon_{lmn}\nabla_k\langle B_i'(t)u_m'(t+\tau)B_n'(t+\tau)\rangle$$
$$=\epsilon_{jkl}\epsilon_{lmn}\nabla_k[\langle B_i'(t)u_m'(t+\tau)\rangle\langle B_n(t+\tau)\rangle]. \tag{5}$$

The advantage of (4) over (2) is clear. While (4) is an equation that can yield directly the quantity, $\mathcal{E}$, of interest, (2) governs $\mathbf{B}'$, a quantity not directly relevant to (1). Supposing that (3) is true, and justifies the neglect of the triple correlation in (4), we may solve that equation without difficulty. Since $\langle u_i'(t)B_j'(t+\tau)\rangle\to 0$ as $\tau\to-\infty$, we obtain

$$\langle u_i'(t)B_j'(t+\tau)\rangle$$
$$=\int_{-\infty}^{t}d\tau'\int d\xi'\,G(\tau-\tau')\,\epsilon_{jkl}\epsilon_{lmn}\nabla_k'[Q_{im}(\tau')\langle B_n(t+\tau')\rangle], \tag{6}$$

where $Q_{im}(\tau)=\langle u_i(t)\,u_m(t+\tau)\rangle$ is the two-point two-time velocity correlation tensor and $G(\tau)=(4\pi\eta\tau)^{-3/2}\exp(-\xi^2/4\eta\tau)$ is the Greens function of the diffusion operator, Δ.

The velocity correlation $Q_{im}(\tau')$ in (6) is negligible when $|\tau'|\gtrsim T$, and when $|\xi'|\gtrsim L$. Thus, $\langle u_i'B_j'\rangle$ is effectively zero for $\tau<-T$, and for larger τ we may, with little error, replace the limits of τ' integration in (6) by $-T$ and $\min(\tau,T)$, while confining the ξ' integration to the interior of $|\xi'|=L$. Specializing now to the high-conductivity limit $L^2\gg\eta T$, we see that if $0<\tau\lesssim L^2/\eta$ the Greens function $G(\tau-\tau')$ is effectively $\delta(\xi-\xi')$. Equation (6) therefore shows that, for $\xi\lesssim L$,

$$\langle u_i'(t)\,B_j'(t+\tau)\rangle\approx(UT/L)\,U\langle B\rangle, \tag{7}$$

and

$$\Delta\langle u_i'(t)\,B_j'(t+\tau)\rangle\approx(\partial/\partial\tau)\langle u_i'(t)\,B_j'(t+\tau)\rangle$$
$$\approx\langle u_i'(t)\,B_j'(t+\tau)\rangle/T, \tag{8}$$

these quantities being essentially zero for $\xi\gtrsim L$. Thus, when $UT/L\ll 1$, the correlation $\langle u_i'B_j'\rangle$ is small compared with $U\langle B\rangle$, despite the fact (see below) that, when $U^2T/\eta\gg 1$, $(\langle\mathbf{B}'^2\rangle)^{1/2}$ is large compared with $\langle B\rangle$. Estimate (7) shows that

$$|\langle u_i'(t)u_m'(t+\tau)B_n'(t+\tau)\rangle| < U|\langle u_m'(t+\tau)B_n'(t+\tau)\rangle| \approx (U^3T/L)\langle B\rangle, \tag{9}$$

so that by (8) the term involving the triple correlation in (4) is indeed much smaller than the first term on the left-hand side of (4), so justifying *a posteriori* the application of FOST. We may also observe that, for $\tau \gg L^2/\eta$, $G(\tau-\tau')$ and therefore $\langle u_i'B_j'\rangle$ will decrease as $\tau^{-3/2}$. The correlation (7) has the magnitude there indicated during an entire Ohmic decay time L^2/η, defined by the correlation length.

We may apply very similar arguments to (5) and establish in a parallel fashion that the term involving the triple correlation $\langle B_i'(t)u_m'(t+\tau)B_n'(t+\tau)\rangle$ is likewise much smaller than $\Delta\langle B_i'(t)B_j'(t+\tau)\rangle \approx (\partial/\partial\tau)\langle B_i'(t)B_j'(t+\tau)\rangle$. However, the consequences of (5) for $\langle B_i'(t)B_j'(t+\tau)\rangle$ are then surprisingly different from those of (4) for $\langle u_i'(t)B_j'(t+\tau)\rangle$. Analogously to (6), we obtain

$$\langle B_i'(t)B_j'(t+\tau)\rangle = \int_{-\infty}^{\tau} d\tau' \int d\xi'\, G(\tau-\tau')\epsilon_{jkl}\epsilon_{lmn} \times \nabla_k'[\langle B_i'(t)u_m'(t+\tau')\rangle\langle B_n(t+\tau')\rangle]. \tag{10}$$

The crucial difference in the estimation of the integrals (6) and (10) arises from the difference in behavior of the two-point two-time correlations. Whereas $Q_{im}(\tau')$ in (6) is effectively nonzero only for $-T \leq \tau' \leq T$, $\langle B_i'u_m'\rangle$ is, as we have seen, nonzero in $-T \leq \tau' \leq L^2/\eta$, an interval large compared with T in the high conductivity limit. Using (7) and following arguments similar to those given below (6) we find that, for $|\xi| \lesssim L$ and $|\tau| \lesssim L^2/\eta$, the integrand of (10) is of order $U^2T\langle B\rangle^2/L^2$ over a τ' interval of order L^2/η. In particular, for $\xi=\tau=0$ we obtain

$$\langle \mathbf{B}'^2\rangle \approx (U^2T/\eta)\langle B\rangle^2. \tag{11}$$

This is the basis of the statement made earlier that, when $U^2T/\eta \gg 1$, $\langle \mathbf{B}'^2\rangle$ will be large compared with $\langle B\rangle^2$. Again, an *a posteriori* comparison of the terms in (5) suggests that the application of FOST to (5) was justified even though, paradoxically, (11) confirms[2] that that approximation could not be made in (2).

ACKNOWLEDGMENTS

We are grateful to A. M. Soward for clarifying discussions, and to the Science Research Council of Great Britain for the award of a grant (B/RG 8532) which enabled one of us (F. K.) to visit the other.

[1]As is usual in magnetokinematics, $\mathbf{u}'$ is assumed known in all its statistical properties. It will here be supposed for simplicity that the turbulence is statistically steady, homogeneous and of zero mean; η is the magnetic diffusivity.

[2]P. H. Roberts and A. M. Soward, J. Math. Phys. **16**, 609 (1975). According to Refs. 3 and 4, there are no finite ranges of parameters in which FOST may be used.
[3]I. Lerche and E. N. Parker, J. Math. Phys. **14**, 1949 (1973).
[4]I. Lerche and E. N. Parker, J. Math. Phys. **16**, 1838 (1975).
[5]See K.-H. Rädler, Z. Naturforsch. A **23**, 1841 (1968), translated in Ref. 6.
[6]P. H. Roberts and M. Stix, NCAR Technical note IA-60 (1971).
[7]H. Bräuer and F. Krause, Astron. Nachr. **294**, 179 (1973); **295**, 223 (1974).

[8]Justified in the sense that no gross, order of magnitude, error would be incurred. They showed[2] that, even when (3) holds, small numerical errors can arise through the unjustified neglect of G′ in the high-velocity tail of the velocity distribution, where $u' \gg \eta/L$ and L/T.

[9]M. Steenbeck and F. Krause, Astron. Nachr. **291**, 49 (1968), translated in Ref. 6.
[10]F. Krause and G. Rüdiger, Solar Phys. **42**, 107 (1975).

On the analyticity of certain stationary nonvacuum Einstein space–times*

Lee Lindblom

Department of Physics and Astronomy, University of Maryland, College Park, Maryland 20742
(Received 22 March 1976)

It is shown that certain nonvacuum solutions of Einstein's general relativistic field equations are analytic space–times, i.e., an analytic atlas exists with respect to which the components of the metric tensor, and all material fields are analytic functions. The two specific cases discussed here are interesting from an astrophysical point of view. The first is the class of space–times containing a source free electromagnetic field: the exterior of a charged black hole, for example. The second is the class of space–times filled with a rigidly moving perfect fluid, often used to describe the interior of a rotating star.

1. INTRODUCTION

Müller zum Hagen[1,2] has shown that every stationary vacuum solution of the Einstein field equations is analytic, i.e., there exists an analytic atlas with respect to which the components of the metric tensor field are analytic functions. His result has proven to be a useful tool for the study of space–times of astrophysical interest, e.g., the exteriors of rotating stars or black holes (see Refs. 3 and 4). The purpose of this paper is to extend those results to certain nonvacuum space–times which have possible astrophysical interpretations. The case of a space–time which contains a source-free electromagnetic field includes the charged black hole solutions. Hawking,[4] in his proof that stationary black holes are axisymmetric, uses the analyticity of the metric tensor. The result presented here, therefore, makes his argument rigorous for the case where electromagnetic fields are present. The other case presented here, space–times containing a rigidly moving barytropic fluid, is often used to model the interiors of rotating stars.

2. BACKGROUND

Analyticity of these space–times is demonstrated by showing that the functions which describe the geometry and the configuration of the matter satisfy systems of elliptic partial differential equations. A theorem of Morrey[5] is then recalled which guarantees the analyticity of such functions. Several definitions and results implied from previous work will be required to effect these proofs; they are simply listed in this section. Discussion and proofs of these points may be found in the references.[1,2,5]

Definition: A coordinate chart in a stationary space–time M is said to be stationary and harmonic if (a) the components of the timelike Killing vector are given by $\eta^\alpha = \delta^\alpha_0$; and if (b) the Christoffel connection satisfies $\Gamma^\alpha_{\mu\nu} g^{\mu\nu} = 0$, $(\alpha, \beta, \cdots = 0, 1, 2, 3)$.

Definition: A function $f(x)$ is said to be Hölder continuous of order $0 < \mu < 1$ (C^μ), on some domain D, if $\exists$ a constant K, such that $\forall\ x, y \in D$, $|f(x) - f(y)| < K|x-y|^\mu$.

Lemma 1 (*Müller zum Hagen*): Assumptions: A space–time M is C^{n+2} for integer $n \geq 2$. It contains a globally timelike vector field, η^α, which is C^{n+1}. The metric tensor is C^n. Assertions: In a neighborhood of each point $x \in M$ there exists a stationary harmonic coordinate chart which is $C^{n+\mu}$, $0 < \mu < 1$, related to the C^{n+2} charts on M.

Lemma 2 (*Müller zum Hagen*): Consider a stationary space–time M in which the components of the metric tensor are analytic functions of the stationary harmonic coordinate systems at each point. The the stationary harmonic coordinate charts form a basis for an analytic atlas on M.

Definition: A system of second order partial differential equations, $\Phi^A(x^\alpha, f^B, \partial_\alpha\partial_\beta f^B) = 0$ $(A, B = 1, 2 \ldots, N)$ is said to be elliptic in some domain D if $\forall\, x^\alpha \in D$ and $\forall$ vectors $\lambda^\alpha \neq 0$,

$$0 \neq \det\left\{\sum_{\alpha,\beta} \lambda^\alpha\lambda^\beta \left[\frac{\partial}{\partial y^B_{\alpha\beta}}\, \Phi^A(x^\gamma, y^C, y^C_\gamma, y^C_{\gamma\epsilon})\right]\right\},$$

evaluated at $y^B = f^B$, $y^B_\alpha = \partial_\alpha f^B$, and $y^B_{\alpha\beta} = \partial_\alpha\partial_\beta f^B$.

Theorem (*Morrey*): Assumptions: f^B is a function which is the solution of the system of elliptic differential equations, $\Phi^A(x^\alpha, f^B, \partial_\alpha f^B, \partial_\alpha\partial_\beta f^B) = 0$, $(A, B = 1, 2 \ldots, N)$ on some domain D. f^B is of class $C^{2+\mu}$, $0 < \mu < 1$. The functions $\Phi^A(x^\alpha, y^B, y^B_\alpha, y^B_{\alpha\beta})$ are analytic in the variables $(x^\alpha, y^B, y^B_\alpha, y^B_{\alpha\beta})$. Assertion: The functions f^B are analytic on the domain D.

3. ELECTROMAGNETISM

The electromagnetic field is described by the vector potential A^α. In a source-free space–time, the field equations which govern A^α are

$$\nabla_\alpha \nabla^\alpha A^\beta + R^\beta_\alpha A^\alpha = 0. \tag{1}$$

(The Lorentz gauge condition has been adopted, $\nabla_\alpha A^\alpha = 0$.) These fields are themselves sources for the gravitational field, via the Einstein equations

$$R_{\alpha\beta} = (2g^{\nu\xi}\delta^\mu_\alpha\delta^\sigma_\beta - \tfrac{1}{2}g_{\alpha\beta}g^{\mu\sigma}g^{\nu\xi})(\nabla_\mu A_\nu - \nabla_\nu A_\mu)(\nabla_\sigma A_\xi - \nabla_\xi A_\sigma). \tag{2}$$

These space–times are called stationary if there exists a globally timelike vector field η^α which satisfies

$$L_\eta A^\alpha = \eta^\mu \nabla_\mu A^\alpha - A^\mu \nabla_\mu \eta^\alpha = 0. \tag{3}$$

$$L_\eta g_{\alpha\beta} = \nabla_\alpha \eta_\beta + \nabla_\beta \eta_\alpha = 0. \tag{4}$$

For space–times described by Eqs. (1)–(3) we will derive the following:

Proposition 1: A stationary space–time M is assumed to contain a source-free electromagnetic field, described by Eqs. (1)–(3). If M is C^6, the components of the Killing vector field η^α are C^5, the metric tensor $g_{\alpha\beta}$ is C^4, and the vector potential A^α is C^3, then M admits an analytic atlas with respect to which $g_{\alpha\beta}$ and A^α are analytic functions.

Proof: In a harmonic coordinate system the components of the Ricci tensor, and the D'Alembertian of a vector field may be expressed as,

$$R_{\alpha\beta}=\tfrac{1}{2}g^{\mu\nu}\partial_\mu\partial_\nu g_{\alpha\beta}+B_{\alpha\beta}(g,\partial g), \tag{5}$$

$$\nabla^\mu\nabla_\mu A^\alpha=g^{\mu\nu}\partial_\mu\partial_\nu A^\alpha+R^\alpha_\mu A^\mu+C^\alpha(A,\partial A,g,\partial g). \tag{6}$$

The functions $B_{\alpha\beta}$ and C^α are functions only of the metric $g_{\alpha\beta}$, the vector field A^α, and their first derivatives. Equations (1) and (2) may be rewritten in harmonic coordinates [using Eqs. (5) and (6)] to obtain,

$$g^{\mu\nu}\partial_\mu\partial_\nu g_{\alpha\beta}=B'_{\alpha\beta}(g,\partial g,A,\partial A), \tag{7}$$

$$g^{\mu\nu}\partial_\mu\partial_\nu A^\alpha=C'^\alpha(g,\partial g,A,\partial A). \tag{8}$$

When a stationary and harmonic coordinate system (existence is guaranteed by Lemma 1) is assumed, Eqs. (3) and (4) may be rewritten as

$$\partial_0 A^\alpha=0,\quad \partial_0 g_{\alpha\beta}=0.$$

In these coordinates, the operator $g^{\mu\nu}\partial_\mu\partial_\nu$ may be replaced by $g^{ij}\partial_i\partial_j$, with $i,j=1,2,3$, in Eqs. (7) and (8). Since g^{ij} is a positive definite matrix, Eqs. (7) and (8) are elliptic systems of differential equations for A^α and $g_{\alpha\beta}$. Morrey's theorem guarantees the analyticity of A^α and $g_{\alpha\beta}$ with respect to the stationary harmonic coordinate charts. Lemma 2 guarantees the existence of an analytic atlas for M. ■

4. FLUIDS

Perfect fluids are described via the Einstein equations,

$$R_{\alpha\beta}=8\pi[(\rho+p)u_\alpha u_\beta+\tfrac{1}{2}(\rho-p)g_{\alpha\beta}]. \tag{9}$$

The energy density of the fluid is ρ, the pressure is p, and u^α ($u^\alpha u_\alpha=-1$) is the four-velocity, tangent to the world lines of the fluid. The fluid under consideration here is assumed to have an analytic barytropic equation of state, i.e. $\rho(p)$ is an analytic function of the pressure. Also, the fluid is assumed to be moving rigidly; this condition is given by,

$$(\delta^\mu_\alpha+u_\alpha u^\mu)(\delta^\nu_\beta+u_\beta u^\nu)(\nabla_\mu u_\nu+\nabla_\nu u_\mu)=0. \tag{10}$$

The stationarity of these space–times is expressed by the existence of a timelike vector field η^α, satisfying Eq. (4) and,

$$L_\eta u^\alpha=\eta^\mu\nabla_\mu u^\alpha-u^\mu\nabla_\mu\eta^\alpha=0, \tag{11}$$

$$L_\eta p=\eta^\mu\nabla_\mu p=0. \tag{12}$$

For these space–times, the following proposition holds:

Proposition 2: A stationary space–time M is assumed to contain a rigidly moving barytropic fluid (with analytic equation of state).[6] If M is C^7, the components of the Killing vector field η^α are C^6, the metric $g_{\alpha\beta}$ is C^5 and, the pressure p and the four-velocity u^α are C^3, then M admits an analytic atlas with respect to which $g_{\alpha\beta}$, p, and u^α are analytic functions.

Proof: Equations (9) and (10) and the fact that the fluid is barytropic imply that the following relationships are satisfied:

$$\nabla^\alpha\nabla_\alpha p=-\nabla^\alpha(\rho+p)u^\beta\nabla_\beta u_\alpha+(\rho+p)(u^\alpha u^\beta R_{\alpha\beta}-\nabla_\alpha u_\beta\nabla^\beta u^\alpha), \tag{13}$$

$$\nabla^\alpha\nabla_\alpha u^\beta=(u^\mu\nabla_\mu u_\alpha)(\nabla^\beta u^\alpha-\nabla^\alpha u^\beta)-u^\beta(\nabla_\mu u_\nu)(\nabla^\nu u^\mu). \tag{14}$$

Equations (9), (13), and (14) form a system of second order differential equations for the functions $g_{\alpha\beta}$, u^α, and p. These equations may be written in a stationary harmonic coordinate system, in analogy with Eqs. (7) and (8),

$$g^{ij}\partial_i\partial_j p=A(p,g,\partial g,u,\partial u),$$

$$g^{ij}\partial_i\partial_j u^\alpha=B^\alpha(p,g,\partial g,u,\partial u),$$

$$g^{ij}\partial_i\partial_j g_{\alpha\beta}=C_{\alpha\beta}(p,g,\partial g,u).$$

These equations form an elliptic system, thus the theorem of Morrey and Müller zum Hagen's lemma can be applied to complete the proof. ■

ACKNOWLEDGMENT

I would like to thank D. Brill and P. Yasskin for helpful comments.

*This research was supported by the National Science Foundation grant GP-25548.

[1]H. Müller zum Hagen, Proc. Cambridge Philos. Soc. 67, 415 (1970).
[2]H. Müller zum Hagen, Proc. Cambridge Philos. Soc. 68, 199 (1970).
[3]L. Lindblom, "Stationary Stars are Axisymmetric" to be published in Ap. J.
[4]S. W. Hawking, Commun. Math. Phys. 25, 152 (1972).
[5]C. B. Morrey, Am. J. Math. 80, 198 (1958).
[6]It has been shown elsewhere (Ref. 3) that a stationary, viscous, heat conducting fluid is necessarily rigidly moving and barytropic. Thus any real, truly stationary, fluid would be expected to satisfy these criteria with the possible exception of the analyticity of the equation of state.

Steady, one-dimensional multigroup neutron transport with anisotropic scattering*

Edward W. Larsen†

Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012

Paul F. Zweifel‡

The Rockefeller University, New York, New York 10021
(Received 14 February 1975; revised manuscripts received 14 July 1975 and 10 March 1976)

The solution of steady, one-dimensional half-space multigroup transport problems with degenerate anisotropic scattering is obtained for L_1 sources and incident distributions. The solution is expressed in terms of contour integrals of the resolvent operator $(\lambda I - K)^{-1}$, where K is the "separated" transport operator. The connection between this method and the singular eigenfunction method is briefly discussed, and the half-space albedo problem is treated in detail. This problem reduces to obtaining the Wiener–Hopf factorization of the dispersion matrix, hence to solving two coupled nonlinear, nonsingular matrix integral equations.

I. INTRODUCTION

The one-dimensional multigroup transport equation, which has been studied extensively in the past decade, has recently been solved[1,2] by the application of resolvent operator techniques first applied to the one-speed equation.[3] Although Refs. 1 consider only isotropic scattering, Sancaktar[2] has extended the analysis to include scattering matrices $C(\mu, s)$ which are separable as a matrix product $C(\mu)L(s)$.

The analysis of Refs. 1–3 leads to singular eigenfunction expansions and is restricted to the case that the interior source and incident distribution are Hölder continuous in the angular variable μ. However, we expect that these expansions can be extended to the much larger class of L_p functions for $p > 1$.[4] In a related paper,[5] the solution of the one-speed equation is obtained in the physically natural L_1 space by employing a contour integral method rather than the singular eigenfunction formalism.

In the present paper, we extend the analysis of Ref. 5 to one-dimensional multigroup transport problems in L_1 with an anisotropic scattering matrix of the form

$$C(\mu, s) = \sum_{n=1}^{M} A_n(\mu) B_n(s). \qquad (1.1)$$

We are able to exploit the technique of Ref. 2, the separable kernel case, by noting that this degenerate sum is in fact a separable matrix product of nonsquare matrices $A(\mu) \cdot B(s)$ (cf. Sec. III). However, we shall use a contour integral method to obtain L_1 solutions rather than construct singular eigenfunction expansions of solutions for Hölder-continuous interior sources and incident distributions. We do this for two reasons: First, L_1 is the natural space in which to solve transport problems; and, second, the singular eigenfunction formalism is notationally very awkward for the present problem.

Aside from the fact that L_1 solutions to multigroup transport problems have not previously been constructed, the degenerate scattering matrix C, defined in (1.1), has hitherto been analyzed with only partial success.[6] This is because previous results were expressed in terms of a matrix Riemann–Hilbert problem which, except for very sepcial cases, has never been solved. But the contour integral method employed in this paper depends upon the solution of a different problem, namely the Wiener–Hopf factorization of the dispersion matrix $\underline{\underline{\Lambda}}(z)$. This factorization has been derived by Mullikin,[7] and it is this recent result which has sparked the current flurry of activity in multigroup neutron transport theory.

The plan of the present paper is as follows. In Sec. II we express the solution of a half-space transport problem in terms of a contour integral of the resolvent operator $(\lambda I - K)^{-1}$ around the spectrum of the "separated" transport operator K. Then in Sec. III we construct the resolvent operator explicitly. In Sec. IV we introduce the analysis necessary to the solution of half-space problems, i.e., the extension of a function $h(\mu)$ defined on $(0, 1]$ to $Eh(\mu)$, defined on $[-1, 1]$, in such a way that $(\lambda I - K)^{-1}Eh$ is analytic in λ for $\mathrm{Re}\lambda < 0$. (This is just the extension operator E of Refs. 1–3. In fact E has a physical interpretation as the "reflection" operator.) It is in this section that the Mullikin factorization, referred to above, is used. This factorization is presented in detail in Appendix A; it is expressed in terms of the (numerical) solution of two coupled, nonlinear, nonsingular matrix integral equations.

In Sec. V we treat the half-space albedo problem in detail. In particular, we take the contour integral solution of Secs. 2–4 and cast it in a form which explicitly indicates the dependence of the solution on the continuous and the discrete spectrum of K. We then consider this expression for the special case of one-speed, isotropic scattering, and show that it reduces to a form which was derived previously for the solution of this problem.[5] The simple connection between this form and the singular eigenfunction form of the solution is then briefly discussed.

In Appendix B, we fully describe the domain of the transport operator defined in Sec. II; this domain is a dense subspace of the L_1 space which is physically appropriate for neutron transport problems.

The analysis presented here cannot be considered completely rigorous from a strict mathematical point of view, since many lengthy technical details involving interchange of orders of integration and differentiation, etc., are omitted. The full analysis, if included, might double the length of this paper without enhancing its physical content.

II. CONTOUR INTEGRAL SOLUTION

We consider the following neutron transport problem:

$$\mu \frac{\partial \psi}{\partial x}(x, \mu) + \Sigma\psi(x, \mu) - \int_{-1}^{1} C(\mu, s)\psi(x, s)\, ds$$

$$= q(x, \mu), \quad x > 0, \quad -1 \le \mu \le 1, \tag{2.1}$$

$$\psi(0, \mu) = \psi_0(\mu), \quad 0 < \mu \le 1, \tag{2.2}$$

$$0 = \lim_{x\to+\infty} q(x, \mu) = \lim_{x\to\infty} \psi(x, \mu). \tag{2.3}$$

In these equations ψ is an $N\times 1$ vector whose components represent the neutron angular density in N velocity groups; Σ is an $N\times N$ diagonal matrix whose diagonal elements σ_i satisfy: $1 = \sigma_1 \le \sigma_2 \le \cdots \le \sigma_N$, with σ_i representing the total cross section in velocity group "i"; $C(\mu, s)$ is the $N\times N$ scattering matrix which is Hölder continuous in μ and s; and ψ_0 and q are nonnegative $N\times 1$ vectors representing the incident distribution and sources respectively.

We require

$$\sum_{i=1}^{N} \int_0^1 \mu\psi_{0,i}(\mu)\, d\mu < \infty;$$

then the total number of neutrons per unit cross sectional area which enter the half-space through its boundary $x = 0$ in a unit time interval is finite. For a similar reason, we require

$$\sum_{i=1}^{N} \int_0^\infty \int_{-1}^{1} q(x, \mu)\, d\mu\, dx < \infty.$$

We seek a solution ψ of problem (2.1)—(2.3) which lies in X, the Banach space of $N\times 1$ vectors $f(x, \mu)$, defined for $x \ge 0$ and $|\mu| \le 1$, and satisfying

$$\|f\| \equiv 2\pi \sum_{i=1}^{N} \int_0^\infty \int_{-1}^{1} |f_i(x, \mu)|\, d\mu\, dx < \infty.$$

Among all the L_p spaces, this is the physically natural space for the transport operator since $\|\psi\|$ represents the total number of neutrons per unit cross sectional area of the system. We note, by the conditions imposed in the previous paragraph, that $q \in X$.

We require the half-space $x \ge 0$ to be subcritical. Thus there cannot exist a solution of (2.1) which is independent of x for $q = 0$.

We proceed by rewriting equation (2.1) as

$$\mu \frac{\partial \psi}{\partial x}(x, \mu) + L\psi(x, \mu) = \mu q_0(x, \mu), \tag{2.4}$$

where

$$q_0(x, \mu) = \mu^{-1} q(x, \mu) \tag{2.5}$$

and where $L: X \to X$ is the bounded operator defined by

$$L\psi(x, \mu) = \Sigma\psi(x, \mu) - \int_{-1}^{1} C(\mu, s)\psi(x, s)\, ds. \tag{2.6}$$

In Sec. III we shall show that for a degenerate kernel C of the form (1.1), for which the half-space $x \ge 0$ is subcritical, the operator $L^{-1}: X \to X$ exists and is bounded. Thus, we may define an operator K by

$$Kg(x, \mu) = L^{-1}\mu g(x, \mu). \tag{2.7}$$

The domain and range of K are not X, but instead the larger Banach space X_1 defined as follows:

$$X_1 = \{g(x, \mu) \mid \mu g(x, \mu) \in X\},$$

$$\|g\|_1 = 2\pi \sum_{i=1}^{N} \int_0^\infty \int_{-1}^{1} |\mu g_i(x, \mu)|\, d\mu\, dx.$$

It is trivial to verify that $K: X_1 \to X_1$ is a bounded operator. Also, by (2.5), $q_0 \in X_1$ and so (2.4) may be written in terms of K as

$$K \frac{\partial \psi}{\partial x}(x, \mu) + \psi(x, \mu) = Kq_0(x, \mu). \tag{2.8}$$

In Sec. III we shall prove that the spectrum of K consists of the line segment $[-1, 1]$ plus certain discrete eigenvalues. We require that none of these eigenvalues be purely imaginary. (Kuscer and Vidav[8] have shown that for a moderator with continuous energy dependence, the reciprocity relation implies that the space eigenvalues—i.e., the point eigenvalues of K—are all real; see Fig. 1.)

Now we shall construct a solution of the transport problem (2.8), (2.2), and (2.3) using a method which is suggested by the operational calculus for operators in a Banach space.[9] Although the procedure may seem cumbersome, it is actually straightforward. The key idea is to express the solution as an integral, around the spectrum of K, of the resolvent operator $(\lambda I - K)^{-1}$ acting on an undetermined function $f(x, \mu, \lambda)$. We then solve for f and express it in terms of the problem data. Following this procedure, we write

$$\psi(x, \mu) = \frac{1}{2\pi i} \int_\Gamma (\lambda I - K)^{-1} f(x, \mu, \lambda)\, d\lambda. \tag{2.9}$$

Here $\Gamma = \Gamma^- \cup \Gamma^+$ is a closed curve enclosing the left and right halves of the spectrum of K, as indicated in Fig. 1. (Γ intersects the Reλ axis at an angle between 0 and $\pi/2$.) We require f, which is to be determined, to be an analytic vector-valued function of λ for λ inside Γ, and a continuous vector-valued function of λ for $\lambda \in \Gamma$, with vector values in X_1. Also, to satisfy Eq. (2.3) and to avoid certain problems at the point $\lambda = 0$, where Γ intersects the spectrum of K, we shall require

$$\lim_{x\to\infty} f(x, \mu, \lambda) = 0, \quad \lambda \in \Gamma,$$
$$\lim_{\lambda\to 0} f(x, \mu, \lambda) = 0, \quad \lambda \in \Gamma. \tag{2.10}$$

To determine f, we require ψ to satisfy (2.8). This yields

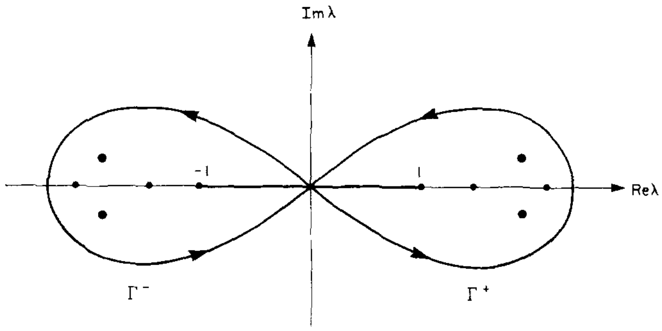


FIG. 1. The contours Γ^- and Γ^+. (The "continuous" spectrum of K is represented by the heavy line, $-1 \le \lambda \le 1$, while the point spectrum is denoted by dots. If the results of Ref. 8 apply, then the point spectrum is real.)

$$Kq_0 = \frac{1}{2\pi i}\int_\Gamma (\lambda I - K)^{-1}\left(K\frac{\partial}{\partial x} + I\right) f\, d\lambda$$
$$= \frac{1}{2\pi i}\int_\Gamma (\lambda I - K)^{-1}\left\{[(K-\lambda I) + \lambda I]\frac{\partial}{\partial x} f + f\right\} d\lambda$$
$$= -\frac{1}{2\pi i}\int_\Gamma \frac{\partial}{\partial x} f\, d\lambda + \frac{1}{2\pi i}\int_\Gamma (\lambda I - K)^{-1}\left[\lambda\frac{\partial}{\partial x} + I\right] f\, d\lambda. \tag{2.11}$$

We shall show that this equation is satisfied if f obeys the boundary conditions (2.10) and the o.d.e.

$$\left(\lambda\frac{\partial}{\partial x} + I\right) f(x,\mu,\lambda) = \lambda q_0(x,\mu). \tag{2.12}$$

The general solution of (2.12) subject to (2.10) can be expressed in the form

$$f(x,\mu,\lambda) = \begin{cases} \exp(-x/\lambda) g(\mu) + \int_0^x \exp[(t-x)/\lambda] q_0(t,\mu)\, dt, & \mathrm{Re}\lambda > 0, \\ \int_{+\infty}^x \exp[(t-x)/\lambda] q_0(t,\mu)\, dt, & \mathrm{Re}\lambda < 0. \end{cases} \tag{2.13}$$

Here $g(\mu)$ is an arbitrary function which will be determined by the boundary condition at $x=0$. [We note that $f(x,\mu,0)=0$ can be obtained from either of the above expressions by taking the limit $\lambda \to 0$.]

Before determining g, we shall show that we have indeed satisfied Eq. (2.11). To do this, we insert (2.12) into (2.11) and obtain

$$Kq_0 = \frac{1}{2\pi i}\int_\Gamma \left[\frac{1}{\lambda} f - q_0\right] d\lambda + \frac{1}{2\pi i}\int_\Gamma (\lambda I - K)^{-1}\lambda q_0\, d\lambda,$$

or

$$0 = \frac{1}{2\pi i}\int_\Gamma \frac{1}{\lambda} f\, d\lambda. \tag{2.14}$$

If f, as given by (2.13), did not satisfy this equation, the representation of our solution as a contour integral would not be correct. However, f does satisfy this equation for any g, as we shall verify at the end of this section.

We now return to Eq. (2.13) in order to determine the arbitrary function g, and hence to complete the construction of the solution. We introduce f, from Eq. (2.13) into (2.9), and obtain the decomposition

$$\psi(x,\mu) = \psi_1(x,\mu) + \psi_2(x,\mu), \tag{2.15}$$

where

$$\psi_1(x,\mu) = \frac{1}{2\pi i}\int_{\Gamma^+} \exp(-x/\lambda)(\lambda I - K)^{-1} g(\mu)\, d\lambda \tag{2.16}$$

and

$$\psi_2(x,\mu) = \frac{1}{2\pi i}\int_{\Gamma^-} (\lambda I - K)^{-1}\int_{+\infty}^x \exp[(t-x)/\lambda] q_0(t,\mu)\, dt\, d\lambda + \frac{1}{2\pi i}\int_{\Gamma^+} (\lambda I - K)^{-1}\int_0^x \exp[(t-x)/\lambda] q_0(t,\mu)\, dt\, d\lambda. \tag{2.17}$$

At this point, ψ satisfies Eqs. (2.8) and (2.3), but not (2.2). This last condition is the one which will determine g. We set $x=0$ in (2.15)–(2.17) and use (2.2) to get

$$\psi_0(\mu) - \psi_2(0,\mu) = \frac{1}{2\pi i}\int_{\Gamma^+} (\lambda I - K)^{-1} g(\mu)\, d\lambda, \quad 0 < \mu \le 1. \tag{2.18}$$

If the range of integration in this equation extended over $\Gamma^+ \cup \Gamma^-$, then this equation would be identically satisfied by any g defined on $-1 \le \mu \le 1$, such that

$$g(\mu) = \psi_0(\mu) - \psi_2(0,\mu), \quad 0 < \mu \le 1. \tag{2.19}$$

Let us therefore require g to satisfy (2.19), and, in addition,

$$0 = \frac{1}{2\pi i}\int_{\Gamma^-} (\lambda I - K)^{-1} g(\mu)\, d\lambda, \quad -1 \le \mu \le 1. \tag{2.20}$$

To satisfy (2.20), we shall extend $g(\mu)$ as defined by (2.19), to $-1 \le \mu < 0$ in such a way that $(\lambda I - K)^{-1} g$ is analytic in λ for $\mathrm{Re}\lambda < 0$. It is customary to express g as

$$g(\mu) = E[\psi_0(\mu) - \psi_2(0,\mu)], \quad -1 \le \mu \le 1, \tag{2.21}$$

where E is the well-known "extension" operator. (See Refs. 1–3.) In Sec. IV we compute E, finally completing the specification of g and hence of the solution to the transport problem (2.1)–(2.3).

To recapitulate, given a "modified" source q_0 and an incident distribution ψ_0 on a half-space, we calculate $(\lambda I - K)^{-1}$ (Sec. III), and $\psi_2(0,\mu)$ from (2.17). Then we define $g(\mu)$, $0 < \mu \le 1$, by (2.19) and appropriately extend g to $-1 \le \mu \le 1$ [Eq. (2.21) and Sec. IV]. The solution ψ is then given by (2.15)–(2.17).

It remains to verify that the conditions imposed earlier on f are actually satisfied. These conditions are:

(i) f is an analytic vector-valued function of λ for λ inside Γ, and a continuous vector-valued function of λ for $\lambda \in \Gamma$, with vector values in X_1;

(ii) f satisfies (2.14);

(iii) ψ, defined by (2.9), is an element of X.

To verify (1), we use (2.13) to obtain the inequality

$$2\pi\sum_{i=1}^{N}\int_0^\infty\int_1^1 |\mu f_i(x,\mu,\lambda)|\,d\mu\,dx \le \frac{|\lambda|^2}{|\mathrm{Re}\lambda|}\left\{2\pi\sum_{i=1}^{N}\int_{-1}^{1}|\mu E g_i(\mu)|\,d\mu+\|q\|\right\}. \tag{2.22}$$

Since the right side of this inequality is finite for each λ inside and on Γ, and since the left side is just $\|f\|_1$, then $f\in X_1$ for each λ inside and on Γ. f is obviously analytic in λ for λ inside Γ, and since $f\to 0$ as $\lambda\to 0$ with $\lambda\in\Gamma$, f is continuous in λ for $\lambda\in\Gamma$. This proves (i).

To prove (ii), we observe from (2.22) that we may write

$$f(x,\mu,\lambda)=\lambda F(x,\mu,\lambda),$$

where F is a vector-valued analytic function of λ for λ inside Γ, and a bounded, piecewise continuous function of λ for $\lambda\in\Gamma$. Therefore,

$$\int_\Gamma \frac{1}{\lambda} f\,d\lambda=\int_\Gamma F\,d\lambda=0,$$

and so f satisfies (2.14). This verifies (ii).

Condition (iii) must be verified because the integral in Eq. (2.9) exists in the context of the Banach space X_1, not the physically correct space X. In Sec. V we shall show that (iii) holds for $q=0$; the proof for $q\neq 0$ is similar, but for simplicity we shall not present it in this paper.

III. CONSTRUCTION OF THE RESOLVENT OPERATOR

In Eq. (2.1), we take the scattering kernel C to be degenerate:

$$C(\mu,s)=\sum_{n=1}^{M}A_n(\mu)B_n(s). \tag{3.1}$$

Here $A_n(\mu)$ and $B_n(s)$ are $N\times N$ matrices. Since $C(\mu,s)$ was assumed to be Hölder continuous in μ and s, it follows that $A_n(\mu)$ and $B_n(s)$ are Hölder continuous in μ and s respectively. We define the $N\times NM$ matrix **A** and the $NM\times N$ matrix **B** by

$$\mathbf{A}=(A_1A_2\cdots A_M)$$

and

$$\mathbf{B}^T=(B_1^TB_2^T\cdots B_M^T),$$

where the superscript T denotes transpose.

Then

$$C(\mu,s)=\mathbf{A}(\mu)\cdot\mathbf{B}(s), \tag{3.2}$$

and so by (2.6), L has the form

$$Lf(x,\mu)=\Sigma f(x,\mu)-\mathbf{A}(\mu)\cdot\textstyle\int_{-1}^{1}\mathbf{B}(s)f(x,s)\,ds. \tag{3.3}$$

To determine L^{-1}, we set $Lf=g$ and multiply by Σ^{-1} to obtain

$$\Sigma^{-1}g(x,\mu)=f(x,\mu)-\Sigma^{-1}\mathbf{A}(\mu)\cdot\textstyle\int_{-1}^{1}\mathbf{B}(s)f(x,s)\,ds. \tag{3.4}$$

Now we multiply on the left by $\mathbf{B}(\mu)$ and integrate over μ to get

$$\textstyle\int_{-1}^{1}\mathbf{B}(s)\,\Sigma^{-1}g(x,s)\,ds=\mathbf{J}\cdot\int_{-1}^{1}\mathbf{B}(s)f(x,s)\,ds, \tag{3.5}$$

where $\mathbf{J}$ is the $NM\times NM$ matrix defined by

$$\mathbf{J}=\mathbf{I}-\textstyle\int_{-1}^{1}\mathbf{B}(s)\,\Sigma^{-1}\mathbf{A}(s)\,ds. \tag{3.6}$$

We solve (3.5) for $\int_{-1}^{1}\mathbf{B}(s)f(x,s)\,ds$ and insert the result into (3.4), which gives the operator L^{-1} as

$$f(x,\mu)=L^{-1}g(x,\mu)=\Sigma^{-1}g(x,\mu)+\Sigma^{-1}\mathbf{A}(\mu)\cdot\mathbf{J}^{-1}\cdot\textstyle\int_{-1}^{1}\mathbf{B}(s)\Sigma^{-1}g(x,s)\,ds. \tag{3.7}$$

Of course, to verify this procedure, we must show that $\det\mathbf{J}=|\mathbf{J}|\neq 0$.

To do this, we assume $|\mathbf{J}|=0$. Then there exists a constant $NM\times 1$ vector **V** such that $\mathbf{J}\cdot\mathbf{V}=0$ and, by (3.3),

$$L\left\{\Sigma^{-1}\mathbf{A}(\mu)\cdot\mathbf{V}\right\}=\mathbf{A}(\mu)\cdot\mathbf{J}\cdot\mathbf{V}=0.$$

Thus $\psi(\mu)=\Sigma^{-1}\mathbf{A}(\mu)\cdot\mathbf{V}$ is a space-independent solution of the homogeneous transport equation (2.4), and so the half-space $x>0$ is critical. Since we have assumed subcriticality, we have a contradiction; hence $|\mathbf{J}|\neq 0$.

The operator K is now defined by (2.7) and (3.7), i.e.,

$$Kg(x,\mu)=\Sigma^{-1}\mu g(x,\mu)+\Sigma^{-1}\mathbf{A}(\mu)\cdot\mathbf{J}^{-1}\cdot\textstyle\int_{-1}^{1}\mathbf{B}(s)\times\Sigma^{-1}sg(x,s)\,ds. \tag{3.8}$$

The resolvent operator $(\lambda I-K)^{-1}\colon X\to X$ exists as a bounded operator for all λ not in $\sigma(K)$, the spectrum of K.

We now compute a formal expression for the resolvent operator $(\lambda I-K)^{-1}$, and then determine $\sigma(K)$ from its singularities. [The same result for $\sigma(K)$ could be obtained by a direct but tedious spectral analysis.]

We proceed by examining

$$h=(\lambda I-K)g(x,\mu)=(\lambda I-\Sigma^{-1}\mu)g(x,\mu)-\Sigma^{-1}\mathbf{A}(\mu)\cdot\mathbf{J}^{-1}\cdot\textstyle\int_{-1}^{1}\mathbf{B}(s)\,\Sigma^{-1}sg(x,s)\,ds. \tag{3.9}$$

We introduce the convenient notation

$$D(\lambda,\mu)=(\lambda I-\Sigma^{-1}\mu)^{-1}, \tag{3.10}$$

and we multiply Eq. (3.9) by $D(\lambda,\mu)$ to obtain

$$D(\lambda,\mu)h(x,\mu)=g(x,\mu)-D(\lambda,\mu)\Sigma^{-1}\mathbf{A}(\mu)\cdot\mathbf{J}^{-1}\cdot\textstyle\int_{-1}^{1}\mathbf{B}(s)\Sigma^{-1}sg(x,s)\,ds. \tag{3.11}$$

We then multiply on the left by $\mathbf{B}(\mu)\Sigma^{-1}\mu$ and integrate over μ to get

$$\textstyle\int_{-1}^{1}\mathbf{B}(s)\Sigma^{-1}sD(\lambda,s)h(x,s)\,ds=\underline{\underline{\Lambda}}(\lambda):\mathbf{J}^{-1}\cdot\int_{-1}^{1}\mathbf{B}(s)\,\Sigma^{-1}sg(x,s)\,ds, \tag{3.12}$$

where $\underline{\underline{\Lambda}}$ is defined by

$$\underline{\underline{\Lambda}}(\lambda)=\mathbf{J}-\textstyle\int_{-1}^{1}\mathbf{B}(s)\Sigma^{-1}sD(\lambda,s)\Sigma^{-1}\mathbf{A}(s)\,ds \tag{3.13}$$

$$= \mathbf{I} - \lambda \int_{-1}^{1} \mathbf{B}(s)\Sigma^{-1} D(\lambda, s)\mathbf{A}(s)\, ds.$$

$\underline{\underline{\Delta}}$ is analytic in λ off the cut $[-1, 1]$, and $|\underline{\underline{\Delta}}(\lambda)| \equiv \Omega(\lambda)$ satisfies $\Omega(\infty) = |\mathbf{J}| \neq 0$. Equations (3.11) and (3.12) can be solved for $g = (\lambda I - K)^{-1}h$, yielding

$$(\lambda I - K)^{-1}h(x, \mu) = D(\lambda, \mu)[h(x, \mu) + \Sigma^{-1}\mathbf{A}(\mu) \cdot \underline{\underline{\Delta}}^{-1}(\lambda) \cdot \mathbf{T}(\lambda, x)], \tag{3.14}$$

where

$$\mathbf{T}(\lambda, x) = \int_{-1}^{1} \mathbf{B}(s)\Sigma^{-1} s D(\lambda, s) h(x, s)\, ds. \tag{3.15}$$

From Eq. (3.14), we observe that $\sigma(K) = [-1, 1] \cup \{\lambda \mid \Omega(\lambda) = 0\}$. A simple application of Gohberg's theorem[10] shows that the set $\{\lambda \mid \Omega(\lambda) = 0\}$ consists of discrete, isolated points. This is well known anyway, since our $\underline{\underline{\Delta}}$ is identical to the dispersion matrix previously calculated for this degenerate scattering kernel.[11]

IV. HALF-RANGE THEORY

As was explained in Sec. II, in order to solve transport problems in a half-space it is necessary to form a function $h(\mu)$, $0 < \mu \leq 1$, from the given source and incoming distribution, and then to extend it to the function $Eh(\mu) = g(\mu)$ defined on the full-range $-1 \leq \mu \leq 1$; the extension operator E is required to satisfy

(i) $Eh(\mu) = h(\mu)$, $0 < \mu \leq 1$,

and

(ii) $(\lambda I - K)^{-1} Eh(\mu)$ is analytic in λ for $\mathrm{Re}\lambda < 0$.

Since h and Eh are independent of x, we cannot interpret K as an operator mapping $X_1 - X_1$. Thus, we introduce the space X_1^0 as the space of all $N \times 1$ vectors $g(\mu)$, defined for $-1 \leq \mu \leq 1$ and satisfying

$$\|g\|_1^0 \equiv 2\pi \sum_{i=1}^{N} \int_{-1}^{1} |\mu g_i(\mu)|\, d\mu < \infty.$$

Since K is independent of and does not act upon x, we may consider K as an operator mapping $X_1^0 - X_1^0$.

In this section we derive the operator E. We begin by introducing the $N \times N$ matrices I_i, $1 \leq i \leq N$, for which the element in the ith row and ith column is 1 and all other elements are 0. Then by (3.10),

$$D(\lambda, \mu) = \sum_{i=1}^{N} \left(\lambda - \frac{\mu}{\sigma_i}\right)^{-1} I_i. \tag{4.1}$$

Thus with $\mathbf{T}(\lambda) = \int_{-1}^{1} \mathbf{B}(s)\Sigma^{-1} s D(\lambda, s) Eh(s)\, ds$ we have, by (3.14),

$$(\lambda I - K)^{-1} Eh(\mu) = \sum_{i=1}^{N} \left(\lambda - \frac{\mu}{\sigma_i}\right)^{-1} \times \left\{ I_i Eh(\mu) + \frac{I_i}{\sigma_i}\mathbf{A}(\mu) \cdot \underline{\underline{\Delta}}^{-1}(\lambda) \cdot \mathbf{T}(\lambda) \right\}. \tag{4.2}$$

Then, in order that $(\lambda I - K)^{-1} Eh$ be analytic for $\mathrm{Re}\lambda < 0$, E must be chosen such that the following three conditions are satisfied:

(a) $[\underline{\underline{\Delta}}^{-1}(\mu)]^{+} \cdot \mathbf{T}^{+}(\mu) = [\underline{\underline{\Delta}}^{-1}(\mu)]^{-} \cdot \mathbf{T}^{-}(\mu) \equiv \mathbf{H}(\mu)$, $-1 \leq \mu < 0$,

(b) $0 = I_i Eh(\mu) + \dfrac{I_i}{\sigma_i}\mathbf{A}(\mu) \cdot \mathbf{H}\left(\dfrac{\mu}{\sigma_i}\right)$, $1 \leq i \leq N$, $-1 \leq \mu < 0$,

and

(c) $\underline{\underline{\Delta}}^{-1}(\lambda) \cdot \mathbf{T}(\lambda)$ is analytic at the points $-\nu_n$, where $|\Omega(-\nu_n)| = 0$ and $\mathrm{Re}\nu_n > 0$. [Here we use the standard notation $F^{\pm}(\mu) = \lim_{\epsilon \downarrow 0} F(\mu \pm i\epsilon)$.]

Condition (a) implies that $(\lambda I - K)^{-1} Eh$ has the same limit as λ approaches the negative branch cut from above and below. Condition (b) implies that this limiting value is finite, and condition (c) implies that $(\lambda I - K)^{-1} Eh$ has no poles with negative real part.

In order that the limits in (a) and (b) exist, it is sufficient that $Eh(\mu)$ be Hölder continuous for $-1 \leq \mu < 0$. We shall verify this below. [See Eqs. (4.10) and (4.11).]

To meet conditions (a)—(c), we introduce the $NM \times NM$ matrices $\mathbf{X}$ and $\mathbf{Y}$, which factor $\underline{\underline{\Delta}}$ as follows:

$$\underline{\underline{\Delta}}(\lambda) = \mathbf{X}(\lambda) : \mathbf{Y}(-\lambda). \tag{4.3}$$

Here $\mathbf{X}(\lambda)$ and $\mathbf{Y}(\lambda)$ are analytic in the entire plane except on the cut $[-1, 0]$, and nonsingular (i.e., invertible) except at the points $-\nu_n$ described in (c). [We note that $\underline{\underline{\Delta}}(\lambda) = \underline{\underline{\Delta}}(-\lambda)$.] Furthermore, $\mathbf{X}$ and $\mathbf{Y}$ have constant (nonzero) limits at $\lambda = \infty$. The existence and construction of $\mathbf{X}$ and $\mathbf{Y}$ are discussed in the Appendix.

Now we assume (and shall verify immediately below) that $\mathbf{T}$ itself may be factored as:

$$\mathbf{T}(\lambda) = \mathbf{X}(\lambda) \cdot \mathbf{L}(\lambda), \tag{4.4}$$

where $\mathbf{L}(\lambda)$ is analytic off the cut $[0, 1]$. Furthermore, since $\mathbf{T}(\infty) = 0$ and $\mathbf{X}(\infty) \neq 0$, $\mathbf{L}(\infty) = 0$.

Equations (4.3) and (4.4) lead to

$$\underline{\underline{\Delta}}^{-1}(\lambda) \cdot \mathbf{T}(\lambda) = \mathbf{Y}^{-1}(-\lambda) \cdot \mathbf{L}(\lambda).$$

Then, since $\mathbf{Y}^{-1}(-\lambda)$ and $\mathbf{L}(\lambda)$ are analytic except on $[0, 1]$ and ν_n, the above conditions (a) and (c) are satisfied. We can now write (b) as

$$I_i Eh(\mu) = -\frac{I_i}{\sigma_i}\mathbf{A}(\mu) \cdot \mathbf{Y}^{-1}\left(-\frac{\mu}{\sigma_i}\right) \cdot \mathbf{L}\left(\frac{\mu}{\sigma_i}\right), \quad 1 \leq i \leq N, \ -1 \leq \mu < 0. \tag{4.5}$$

This defines $Eh(\mu)$ for $-1 \leq \mu < 0$ in terms of $\mathbf{L}$. It remains to determine $\mathbf{L}$ and to verify that the factorization (4.4) is possible and consistent with (4.5).

The factorization (4.4) is valid if there exists a matrix function $\mathbf{L}(\lambda)$ analytic off the cut $[0, 1]$, vanishing at ∞, and satisfying

$$\frac{1}{2\pi i}[\mathbf{T}^{+}(\mu) - \mathbf{T}^{-}(\mu)] = \frac{1}{2\pi i}[\mathbf{X}^{+}(\mu) - \mathbf{X}^{-}(\mu)] \cdot \mathbf{L}(\mu), \quad -1 < \mu < 0, \tag{4.6}$$

and

$$\frac{1}{2\pi i}[\mathbf{T}^{+}(\mu) - \mathbf{T}^{-}(\mu)] = \mathbf{X}(\mu) \cdot \frac{1}{2\pi i}[L^{+}(\mu) - L^{-}(\mu)], \quad 0 < \mu < 1. \tag{4.7}$$

From (3.15) and (4.1),

$$\frac{1}{2\pi i}[\mathbf{T}^+(\mu) - \mathbf{T}^-(\mu)] = \sum_{i=1}^{N} \mathbf{B}(\sigma_i\mu)\sigma_i\mu I_i E g(\sigma_i\mu),$$
$$-1 < \mu < 1. \tag{4.8}$$

[In Eq. (4.8) and in several others to follow, we adopt the notation

$$0 = \mathbf{B}(\sigma_i\mu) = \mathbf{A}(\sigma_i\mu) = h(\sigma_i\mu) \quad \text{for } |\sigma_i\mu| > 1.]$$

Combining Eqs. (4.7) and (4.8), and using $\mathbf{L}(\infty) = 0$, we obtain from Cauchy's theorem

$$\mathbf{L}(\lambda) = \sum_{i=1}^{N} \int_0^1 \mathbf{X}^{-1}\left(\frac{s}{\sigma_i}\right) \cdot \mathbf{B}(s) I_i h(s) \frac{s}{\lambda\sigma_i - s} ds. \tag{4.9}$$

Thus $\mathbf{L}$ is defined in terms of h for the half range $0 < \mu < 1$. It remains to show that (4.6) and (4.5) are consistent. To do this, we observe from (4.3), (3.12), and (4.1) that

$$\frac{1}{2\pi i}[\mathbf{X}^+(\mu) - \mathbf{X}^-(\mu)]$$
$$= \frac{1}{2\pi i}[\Lambda^+(\mu) - \Lambda^-(\mu)] : \mathbf{Y}^{-1}(-\mu)$$
$$= -\sum_{i=1}^{N} \mathbf{B}(\mu\sigma_i) I_i \mu \mathbf{A}(\mu\sigma_i) \cdot \mathbf{Y}^{-1}(-\mu), \quad -1 \le \mu < 0.$$

Using this and (4.8), we may rewrite (4.6) as

$$0 = \sum_{i=1}^{N} \mathbf{B}(\sigma_i\mu)\sigma_i\mu \Big[I_i E h(\sigma_i\mu) + \frac{I_i}{\sigma_i}\mathbf{A}(\sigma_i\mu) \cdot \mathbf{Y}^{-1}(-\mu) \cdot \mathbf{L}(\mu)\Big], \quad -1 \le \mu < 0.$$

But by (4.5), each of the bracketed terms is zero, and so (4.5) and (4.6) are indeed consistent.

Therefore, the factorization (4.4) is valid, $\mathbf{L}$ is defined by (4.9), and the extension Eh is defined in terms of $\mathbf{L}$ by (4.5). Solving for Eh explicitly, we find

$$Eh(\mu) = \begin{cases} h(\mu) & 0 < \mu \le 1, \\ \int_0^1 \mathcal{E}(\mu, s) h(s)\, ds, & -1 \le \mu < 0, \end{cases} \tag{4.10}$$

where $\mathcal{E}$ is the $N \times N$ matrix defined by

$$\mathcal{E}(\mu, s) = \sum_{i,j=1}^{N} \left[I_i \mathbf{A}(\mu) \cdot \mathbf{Y}^{-1}\left(-\frac{\mu}{\sigma_i}\right) : \mathbf{X}^{-1}\left(\frac{s}{\sigma_j}\right) \cdot \mathbf{B}(s) I_j \right] \frac{s}{s\sigma_i - \mu\sigma_j},$$
$$-1 \le \mu < 0, \quad 0 < s \le 1. \tag{4.11}$$

Since $Y^{-1}(-\mu)$ is analytic on $[-1, 0]$ and $\mathbf{A}$ is Hölder continuous (this requirement is stated in Sec. III), it follows that $Eh(\mu)$ is Hölder continuous for $-1 \le \mu < 0$. This property of E was required earlier in the present section.

This completes the construction of the operator E and hence of the solution of the transport problem (2.1)—(2.3). The explicit role which E plays in the solution of half space problems is described in Sec. II.

We note that for a sourceless half-space with incident distribution ψ_0, the neutron density for $x \ge 0$ is given by (2.15) with $g = E\psi_0$. For $x = 0$ and $-1 \le \mu < 0$, this equation reduces to

$$\psi(0, \mu) = E\psi_0(\mu), \quad -1 \le \mu < 0.$$

Thus the operator E physically describes the reflection of a beam incident upon a sourceless half-space.

V. THE HALF-SPACE ALBEDO PROBLEM

In order to clarify the contour integral techniques introduced in Sec. II, we sketch below the solution of the half-space albedo problem.

This problem is defined by Eqs. (2.1)—(2.3) with $q = 0$. The solution is given by Eq. (2.16) with $g = E\psi_0$; $(\lambda I - K)^{-1}$ is expressed in (3.14), (3.15), and the operator E is defined by (4.10), (4.11).

Using the explicit form for $(\lambda I - K)^{-1}$ in (2.16), we obtain

$$\psi(x, \mu) = \frac{1}{2\pi i} \int_{\Gamma^+} \exp(-x/\lambda) D(\lambda, \mu)\Big\{E\psi_0(\mu) + \Sigma^{-1}\mathbf{A}(\mu) \cdot \underline{\underline{\Lambda}}^{-1}(\lambda) \cdot \int_{-1}^{1} \mathbf{B}(s)\Sigma^{-1} s D(\lambda, s) E\psi_0(s)\, ds\Big\}\, d\lambda. \tag{5.1}$$

However, using (4.1) and

$$d(x, \mu) = \begin{cases} x/\mu, & \mu > 0, \\ +\infty, & \mu \le 0, \end{cases}$$

we get

$$\frac{1}{2\pi i} \int_{\Gamma^+} \exp(-x/\lambda) D(\lambda, \mu) E\psi_0(\mu) = \exp[-\Sigma d(x, \mu)]\psi_0(\mu).$$

Applying this to the first term in (5.1) and interchanging the integrations in the second term, we get

$$\psi(x, \mu) = \exp[-\Sigma d(x, \mu)]\psi_0(\mu) + \int_{-1}^{1} F(x, \mu; s) E\psi_0(s)\, ds. \tag{5.2}$$

The first term in this equation represents the uncollided neutrons. In the integral term, representing the collided flux, the kernel F is defined by

$$F(x, \mu; s) = \frac{1}{2\pi i} \int_{\Gamma^+} \exp(-x/\lambda) D(\lambda, \mu) \Sigma^{-1}\mathbf{A}(\mu) \cdot \underline{\underline{\Lambda}}^{-1}(\lambda) \cdot \mathbf{B}(s)\Sigma^{-1} D(\lambda, s) s\, d\lambda. \tag{5.3}$$

To express F explicitly as an integral over the positive spectrum of K, let us make the assumption that the eigenvalues $\nu_0, \ldots, \nu_p$ of K in the right half-plane are simple and do not lie on the cut $[0, 1]$. Then F may be written as

$$F(x, \mu; s) = \sum_{k=0}^{p} \exp(-x/\nu_k) F_{\nu_k}(\mu, s) + \int_0^1 \exp(-x/\nu) F_\nu(\mu, s)\, d\nu. \tag{5.4}$$

The terms $F_{\nu_k}(\mu, s)$ arise from the residues of the integrand in (5.3) at the (simple) poles ν_k. If we recall that $\Omega = |\underline{\underline{\Lambda}}|$ and define the cofactor matrix $\mathbf{R}$ by

$$\underline{\underline{\Lambda}}^{-1}(\lambda) = \Omega^{-1}(\lambda)\mathbf{R}^{-1}(\lambda),$$

then F_{ν_k} is explicitly given by

$$F_{\nu_k}(\mu, s) = D(\nu_k, \mu)\Sigma^{-1}\mathbf{A}(\mu)\mathbf{R}^{-1}(\nu_k)\mathbf{B}(s)\Sigma^{-1}D(\nu_k, s)\frac{s}{\Omega'(\nu_k)}. \quad (5.5)$$

The matrix function $F_\nu(\mu, s)$ in (5.4) arises from the singularities of the integrand in (5.3) on the cut $[0, 1]$. To evaluate F_ν, we replace $D(\lambda, \mu)$ and $D(\lambda, s)$ in (5.3) with the expansion (4.1), and then close the contour around the cut $[0, 1]$; the result is the integral term in (5.4) with F_ν given by

$$\begin{aligned} F_\nu(\mu, s) = {} & \delta(\nu - \mu)\Sigma^{-1}\mathbf{A}(\mu)\cdot\tfrac{1}{2}\{[\underline{\underline{\Lambda}}^{-1}(\mu)]^+ \\ & + [\underline{\underline{\Lambda}}^{-1}(\mu)]^-\}\cdot\mathbf{B}(s)\Sigma^{-1}D(\mu, s)s \\ & + \delta(\nu - s)D(s, \mu)\Sigma^{-1}\mathbf{A}(\mu)\cdot\tfrac{1}{2}\{[\Lambda^{-1}(s)]^+ \\ & + [\underline{\underline{\Lambda}}^{-1}(s)]^-\}\cdot\mathbf{B}(s)\Sigma^{-1}s + D(\nu, \mu)\Sigma^{-1}\mathbf{A}(\mu)\cdot\frac{1}{2\pi i} \\ & \times\{[\underline{\underline{\Lambda}}^{-1}(\nu)]^- - [\underline{\underline{\Lambda}}^{-1}(\nu)]^+\}\cdot\mathbf{B}(s)\Sigma^{-1}D(\nu, s)s. \end{aligned} \quad (5.6)$$

The first two terms in this equation contain delta functions, while the third term contains two principal value functions. Thus the integral in (5.4) is a singular integral; $F(x, \mu; s)$ itself however is smooth in all of its variables, as can be seen from the form (5.3).

The function F in (5.2) is thus explicitly given by Eqs. (5.4)—(5.6) for the case of simple eigenvalues ν_n not lying in the cut $[0, 1]$. If the eigenvalues are not simple or lie in the cut $[0, 1]$, then Eq. (5.3) still defines F; however, greater care must be taken in calculating the multiple residues or the contributions due to poles on the cut $[0, 1]$. We shall not consider these complications here.

Finally, if we introduce (4.10), (4.11) into (5.2), we obtain as the final form for the solution of the albedo problem,

$$\psi(x, \mu) = \exp[-\Sigma d(x, \mu)]\psi_0(\mu) + \int_0^1 H(x, \mu; s)\psi_0(s)\, ds, \quad (5.7)$$

where

$$H(x, \mu; s) = F(x, \mu; s) + \int_{-1}^0 F(x, \mu; \sigma)\mathcal{E}(\sigma, s)\, d\sigma. \quad (5.8)$$

The first "singular" term on the right side of (5.7) describes the uncollided neutrons. Therefore, H in the remaining integral term is a regular Green's function for neutrons having undergone one or more collisions.

Let us now indicate how one can prove that the solution $\psi(x, \mu) \in X$. (We already know $\psi \in X_1$.) The key idea is to verify that $H(x, \mu; s) = sG(x, \mu; s)$, where G is continuous in all of its variables. (We shall explicitly verify this for a simple case below.) Then ψ, as defined by (5.7), is in X for ψ_0 satisfying

$$\sum_{i=1}^N \int_0^1 |\mu\psi_{0,i}(\mu)|\, d\mu < \infty,$$

as required in Sec. I.

Now we shall show that the above solution reduces, for one-speed, isotropic scattering problems, to the form derived in Ref. 5. For such problems we may take $\Sigma = 1$, $C(\mu, s) = c/2$, $0 < c < 1$, and $\mathrm{A} = c$, $\mathrm{B} = 1$. Hence, by Eq. (4.11),

$$\mathcal{E}(\mu, s) = \frac{c}{2}\frac{1}{Y(-\mu)}\frac{1}{X(s)}\frac{s}{s - \mu}.$$

Also, by (5.3),

$$F(x, \mu; s) = \frac{1}{2\pi i}\int_{\Gamma^+}\exp(-x/\lambda)\frac{1}{\lambda - \mu}\frac{c}{2\Lambda(\lambda)}\frac{s}{\lambda - s}\, d\lambda,$$

and hence by (5.8),

$$H(x, \mu; s) = \frac{1}{2\pi i}\int_{\Gamma^+}\exp(-x/\lambda)\frac{1}{\lambda - \mu}\frac{cs}{2\Lambda(\lambda)} \times\left(\frac{1}{\lambda - s} + \frac{c}{2X(s)}\int_{-1}^0 \frac{\sigma}{\lambda - \sigma}\frac{1}{X(-\sigma)}\frac{1}{s - \sigma}\, d\sigma\right) ds. \quad (5.9)$$

Except for some simple changes in notation, Eq. (5.9) is identical to Eq. (4.12) of Ref. 5; in this reference, it is shown that H may be rewritten as

$$H(x, \mu; s) = \frac{cs}{2X(s)}\frac{1}{2\pi i}\int_{\Gamma^+}\frac{\exp(-x/\lambda)}{(\lambda - \mu)(\lambda - s)}\frac{1}{X(-\lambda)}\, d\lambda. \quad (5.10)$$

From this simple form, it is obvious without further analysis that ψ, defined by Eqs. (5.7), (5.10) is in X for ψ_0 satisfying

$$\int_0^1 |\mu\psi_0(\mu)|\, d\mu < \infty.$$

There is a detailed discussion in Ref. 5 concerning the connection between the solution (5.7), (5.10), and the singular eigenfunction solution. The results, briefly, are as follows. If one closes the contour Γ^+ in Eq. (5.10) around the singular points $[0, 1] \cup \{\nu_0\}$ of $X(-\lambda)$ and then introduces the resulting expression for H into Eq. (5.7), then one can interchange certain integrations (using the Poincaré—Bertrand formula) to obtain exactly the singular eigenfunction solution. To do this, however, one must require ψ_0 to be Hölder continuous. Thus while the singular eigenfunction solution is more elegant in appearance, the above solution (5.7), (5.10) is mathematically more regular; the boundary data does not occur within any singular integrals, and hence it need not be so smooth.

APPENDIX A: FACTORIZATION OF Λ

Recently, Mullikin[7] has proved the following result: consider the integral equation

$$(\mathbf{I} - \mathcal{K})\cdot\rho(x) = \mathbf{q}(x), \quad 0 < x < \infty, \quad (\mathrm{A}1)$$

where $\mathcal{K}$ is an integral operator with a difference kernel

$$\mathcal{K}\cdot\rho(x) = \int_0^\infty \mathbf{k}(x - y)\cdot\rho(y)\, dy, \quad (\mathrm{A}2)$$

and where ρ and q are defined in a suitable Banach space X. Let $\mathbf{K}(z)$ be the Fourier transform of $\mathbf{k}$:

$$\mathbf{K}(z)=\int_{-\infty}^{+\infty}\exp(ixz)\mathbf{\kappa}(x)\,dx. \quad (A3)$$

Then if the spectral radius of $\mathcal{K}$ is less than one (or more generally if 1 is not in the spectrum of $\mathcal{K}$), $\mathbf{I}-\mathbf{K}(z)$ may be factored on the real axis as follows:

$$(\mathbf{I}-\mathbf{K}(z)):\mathbf{H}_r(z):\mathbf{H}_l(-z)=\mathbf{I}, \quad \mathrm{Im}z=0. \quad (A4)$$

Here $\mathbf{H}$ and $\mathbf{H}_r$ are analytic in z for $\mathrm{Im}z\geq 0$, continuous in z for $\mathrm{Im}z\geq 0$, and nonsingular (i.e., invertible for $\mathrm{Im}z\geq 0$).

In addition, Mullikin has shown that matrix functions which satisfy the nonlinear equations

$$\mathbf{H}_r^{-1}(z)=\mathbf{I}+\frac{1}{2\pi i}\int_{-\infty}^{\infty}\mathbf{H}_l(t):\mathbf{K}(-t)\frac{1}{t+z}dt, \quad \mathrm{Im}z>0, \quad (A5)$$

and

$$\mathbf{H}^{-1}(z)=\mathbf{I}+\frac{1}{2\pi i}\int_{-\infty}^{+\infty}\mathbf{K}(t):\mathbf{H}_r(t)\frac{1}{t+z}dt, \quad \mathrm{Im}z>0, \quad (A6)$$

also satisfy Eq. (A4).

To use these results, we write Eq. (2.1) in the form

$$\left(\mu\frac{\partial}{\partial x}+\Sigma\right)\psi(x,\mu)=\mathbf{A}(\mu)\cdot\rho(x)+q(x,\mu), \quad (A7)$$

with

$$\rho(x)=\int_{-1}^{1}\mathrm{B}(s)\psi(x,s)\,ds. \quad (A8)$$

To obtain an equation for ρ, we invert the operator on the left side of (A7) and use (2.2) and (2.3). We obtain an equation for ψ explicitly in terms of ρ. We then multiply this equation on the left by $\mathbf{B}(\mu)$ and integrate over μ, obtaining Eq. (A1) with $\mathbf{\kappa}$ defined by

$$\mathbf{\kappa}(x)=\begin{cases}\displaystyle\int_0^1\mathrm{B}(\mu)\frac{\exp(-\Sigma x/\mu)}{\mu}\mathrm{A}(\mu)\,d\mu, & x>0,\\ \displaystyle\int_0^1\mathrm{B}(-\mu)\frac{\exp(\Sigma x/\mu)}{\mu}\mathbf{A}(\mu)\,d\mu, & x<0.\end{cases} \quad (A9)$$

Introducing (A9) into (A3) and inverting the x and μ integrations, we find (after some routine algebra) that

$$\mathbf{I}-\mathbf{K}(z)=\underline{\underline{\Lambda}}(i/z), \quad (A10)$$

where $\underline{\underline{\Lambda}}$, we recall, is given by (3.13).

Now we combine $\underline{\underline{\Lambda}}(z)=\underline{\underline{\Lambda}}(-z)$ with (A10) and (A4) to get

$$\underline{\underline{\Lambda}}(z)=\mathbf{H}_l^{-1}(i/z):\mathbf{H}_r^{-1}(-i/z), \quad \mathrm{Re}z=0. \quad (A11)$$

If we define

$$\mathbf{X}(z)=\mathbf{H}_l^{-1}(i/z), \quad \mathrm{Re}z>0, \quad (A12)$$

and

$$\mathbf{Y}(z)=\mathbf{H}_r^{-1}(i/z), \quad \mathrm{Re}z>0, \quad (A13)$$

then, by the properties of $\mathbf{H}(z)$ and $\mathbf{H}_r(z)$, $\mathbf{X}(z)$, $\mathbf{Y}(z)$ are analytic and nonsingular for $\mathrm{Re}z>0$; also, by Eqs. (A11)–(A13),

$$\underline{\underline{\Lambda}}(z)=\mathbf{X}(z):\mathbf{Y}(-z), \quad \mathrm{Re}z=0. \quad (A14)$$

[To compare with Refs. 1 and 2, make the transformation $X(z)\rightleftharpoons Y(-z)$.]

Equations (A5) and (A6) can now be rewritten in terms of $\mathbf{X}$, $\mathbf{Y}$, and $\underline{\underline{\Lambda}}$ using (A10), (A12), and (A13); by suitably deforming the contour of integration (for details, see Ref. 1) we obtain

$$\mathbf{X}(z)=\mathbf{I}-z\int_0^1\frac{1}{2\pi it}[\underline{\underline{\Lambda}}^+(t)-\underline{\underline{\Lambda}}^-(t)]:\mathbf{Y}^{-1}(t)\frac{dt}{z+t}, \quad \mathrm{Re}z>0, \quad (A15)$$

and

$$\mathbf{Y}(z)=\mathbf{I}-z\int_0^1\mathbf{X}^{-1}(t):[\underline{\underline{\Lambda}}^+(t)-\underline{\underline{\Lambda}}^-(t)]\times\frac{1}{2\pi it}\frac{dt}{z+t}, \quad \mathrm{Re}z>0. \quad (A16)$$

To proceed further with this system of nonlinear integral equations, the following equation, derived from (3.13), may be useful:

$$\frac{1}{2\pi it}[\underline{\underline{\Lambda}}^+(t)-\underline{\underline{\Lambda}}^-(t)]=\sum_{j=1}^{N}\mathrm{B}(t\sigma_j)I_j\mathbf{A}(t\sigma_j), \quad 0\leq t\leq 1. \quad (A17)$$

The numerical solution of (A15) and (A16) then provides the solution of the transport equation.[12] In (A17) we adopt the convention $\mathrm{B}(\mu)=\mathbf{A}(\mu)=0$ for $|\mu|>1$. The remarks of the following two paragraphs can be derived easily from results proved in Ref. 1.

Any solution pair $\mathbf{X}(z)$, $\mathbf{Y}(z)$ of (A15) and (A16) can trivially be extended analytically to the entire complex plane except for the cut $[-1,0]$; as an easy consequence, Eq. (A14) holds in the entire complex plane except on the cut $[-1,1]$.

Thus, any solution pair $\mathbf{X}(z)$, $\mathbf{Y}(z)$ of (A15), (A16) satisfies (A14) in the cut plane and $\mathbf{X}(z)$, $\mathbf{Y}(z)$ are analytic for $\mathrm{Re}z>0$. However, in Sec. IV we also require that $\mathbf{X}(z)$ and $\mathbf{Y}(z)$ be invertible for $\mathrm{Re}z>0$ or, equivalently, that $|X(-\nu_n)|=|Y(-\nu_n)|=0$ (the ν_n are defined in Sec. IV, and these zeros of $\mathbf{X}$ and $\mathbf{Y}$ must be of the proper order). These extra constraints will uniquely determine the solution pair of (A15), (A16).

To summarize, if 1 is not in the spectrum of $\mathcal{K}$, then $\underline{\underline{\Lambda}}(z)$ is factorable by Eq. (4.3); $\mathbf{X}(z)$ and $\mathbf{Y}(z)$ are analytic in z except on the cut $[-1,0]$ and satisfy the nonlinear equations (A15) and (A16) and the constraints that $\mathbf{X}(z)$, $\mathbf{Y}(z)$ be invertible for $\mathrm{Re}z>0$.

A simple mathematical criterion that the spectral radius of $\mathcal{K}$ be less than 1 is of course that $\|\mathcal{K}\|<1$. From Eqs. (A2) and (A9), we obtain

$$\|\mathcal{K}\|\leq\sup_{0<y<\infty}\int_{x=0}^{\infty}\|\mathbf{\kappa}(x-y)\|\,dx\leq 2\|\mathrm{A}\|\|\mathrm{B}\|.$$

In this inequality, we treat $\mathrm{B}(\mu)$ as a mapping from $N\times 1$ vectors $\psi(\mu)$ into $NM\times 1$ vectors under the sup norm, and $\mathrm{A}(\mu)$ as a mapping from $N\times NM$ vectors $\psi(\mu)$ into $N\times 1$ vectors under the sup norm.

Thus, if

$$\|\mathrm{A}\|\|\mathrm{B}\|<\tfrac{1}{2},$$

then the half-space $x > 0$ is subcritical, and the above factorization of $\underline{\underline{\Lambda}}$ exists.

APPENDIX B: THE DOMAIN OF THE TRANSPORT OPERATOR

Here we describe the domain of the transport operator, i.e., the operator acting on ψ in Eq. (2.1). This provides a description of the regularity properties possessed by L_1 solutions of transport problems.

If we denote the streaming operator $\mu\,\partial/\partial x$ by T, then by (2.6) we may rewrite (2.1) as

$$(T+L)\psi = q.$$

The domain of $T+L$, is the subspace of X such that $(T+L)h \in X$ for each $h \in D(T+L)$. However, $L: X \to X$ is a bounded operator, and so

$$D(T+L) = D(T). \tag{B1}$$

The operator T, which is just a first-order ordinary differential operator, has the following domain: $\psi \in D(T)$ if and only if

(a) $\psi(x,\mu)$ is continuous in x for almost every μ,

(b) $\psi(x,\mu)$ is differentiable in x for almost every x and μ,

(c) $\psi \in X$ and $\mu\,\partial\psi/\partial x \in X$.

[Condition (c) is obvious. Conditions (a) and (b) follow from $D(T) = D(\lambda I - T) = R((\lambda I - T)^{-1})$ for $\lambda \notin \sigma(T)$, constructing $(\lambda I - T)^{-1}$ for $\lambda \notin \sigma(T)$, and applying elementary results from any real variables text, e.g., Ref. 13.]

Thus the set of functions satisfying (a), (b), and (c) is the domain of T; by (B1), this is also the domain of the original transport operator $T+L$.

*Research supported in part by the Air Force Office of Scientific Research under Grants No. AFOSR-71-2013 and AFOSR-72-2187.

†Present address: Department of Mathematics, University of Delaware, Newark, Delaware 19711.

‡John Simon Guggenheim Memorial Foundation Fellow on leave from Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

[1] R. L. Bowden, S. Sancaktar, and P.F. Zweifel, J. Math. Phys. **17**, 76, 81 (1976). A summary of previous efforts and a representative bibliography are included in these references. We adopt the notation of these references, or see Ref. 7.

[2] S. Sancaktar, Transport Theory Stat. Phys. **4**, 71 (1975).

[3] E.W. Larsen and G.J. Habetler, Comm. Pure Appl. Math. **26**, 525 (1973); E.W. Larsen, Comm. Pure Appl. Math. **27**, 523 (1974).

[4] E.W. Larsen, S. Sancaktar, and P.F. Zweifel, J. Math. Phys. **16**, 1117 (1975).

[5] E.W. Larsen, Comm. Pure Appl. Math. **28**, 729 (1975).

[6] P. Silvennoinen and P. F. Zweifel, J. Math. Phys. **13**, 1114 (1972).

[7] T.W. Mullikin, Transport Theory Stat. Phys. **4**, 215 (1973).

[8] I. Kuscer and I. Vidav, J. Math. Anal. Appl. **25**, 80 (1969).

[9] A. E. Taylor, *Introduction to Functional Analysis* (Wiley, New York, 1967).

[10] I. C. Gohberg and M.G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators* (Amer. Math. Soc., Providence, R.I., 1969), p. 21.

[11] P. Silvennoinen and P. F. Zweifel, Nucl. Sci. Eng. **42**, 103 (1970).

[12] It has been shown that a unique iterative solution to Eqs. (A15) and (A16) exists and that it possesses the required analyticity properties (R. L. Bowden, R. Menikoff, and P. F. Zweifel, to be published).

[13] H. L. Royden, *Real Analysis* (Macmillan, New York, 1968), p. 269, pp. 101—07.

A simple proof of the angular momentum Helmholtz theorem and the relation of the theorem to the decomposition of solenoidal vectors into poloidal and toroidal components

H. E. Moses

Air Force Geophysics Laboratory, Hanscom Air Force Base, Massachusetts 01731
(Received 18 March 1976)

Vector spherical harmonics are used in a simple proof of the angular-momentum Helmholtz theorem. The decomposition of vectors defined on a sphere into two components which this theorem gives is carried out explicitly. Furthermore, the potentials which occur in the theorem are given explicitly in terms of the original vector. The decomposition of solenoidal vectors into poloidal and toroidal components is also carried out explicitly. It is shown how these components are related to the components given by the angular-momentum Helmholtz theorem.

1. THE ANGULAR-MOMENTUM HELMHOLTZ THEOREM

Let the vector operator $\mathbf{L}=\{L_1, L_2, L_3\}$ be defined as a differential operator on the surface of a sphere of radius r by

$$L_1=\left(\sin\phi\frac{\partial}{\partial\theta}+\cot\theta\cos\phi\frac{\partial}{\partial\phi}\right),$$
$$L_2=\left(-\cos\phi\frac{\partial}{\partial\theta}+\cot\theta\sin\phi\frac{\partial}{\partial\phi}\right), \qquad (1)$$
$$L_3=\frac{\partial}{\partial\phi},$$

where θ,ϕ are the usual angles in polar coordinates defined by

$$x_1=r\sin\theta\cos\phi,\quad x_2=r\sin\theta\sin\phi,\quad x_3=r\cos\theta. \qquad (2)$$

Let $\mathbf{f}(\theta,\phi)$ be a vector defined on the surface of the sphere. Then the angular-momentum Helmholtz theorem says that $\mathbf{f}$ can be decomposed into two components

$$\mathbf{f}(\theta,\phi)=\mathbf{f}_1(\theta,\phi)+\mathbf{f}_2(\theta,\phi), \qquad (3)$$

such that

$$\mathbf{L}\times\mathbf{f}_1=-\mathbf{f}_1,\quad \mathbf{L}\cdot\mathbf{f}_2=0. \qquad (4)$$

Furthermore, "potentials" $V(\theta,\phi)$, $\mathbf{A}(\theta,\phi)$ exist such that

$$\mathbf{f}_1(\theta,\phi)=\mathbf{L}V(\theta,\phi),$$
$$\mathbf{f}_2(\theta,\phi)=\mathbf{L}\times\mathbf{A}(\theta,\phi)+\mathbf{A}(\theta,\phi). \qquad (5)$$

The quantities $\mathbf{f},\mathbf{f}_i,V,\mathbf{A}$ are complex in general but may also be real.

The decomposition given by Eqs. (3)–(5) is called the angular-momentum Helmholtz theorem and was given, in part, in Ref. 1 and entirely in Ref. 2. Another proof was given in Ref. 3. Generalizations in several directions are discussed in Refs. 4–6.

The operators $\mathbf{M}=-i\mathbf{L}$ are the components of the orbital angular momentum in quantum mechanics, $\mathbf{M}=-i\mathbf{x}\times\nabla$. This fact accounts for the name of the theorem. It should be noted that the components of $\mathbf{M}$ are also the infinitesimal generators of the rotation group in a scalar representation. This observation has been used in the generalization of Refs. 4–6.

In the present section we shall show how $\mathbf{f}_1$ and $\mathbf{f}_2$ are obtained from $\mathbf{f}$. We shall also obtain explicitly the potentials V and $\mathbf{A}$ from $\mathbf{f}$. In carrying out this program we shall provide another proof of the angular-momentum Helmholtz theorem.

We make use of the vector spherical harmonics $\mathbf{Y}_{JlM}(\theta,\phi)$ which are introduced in Ref. 7. We shall, however, use the notation and properties of the vector spherical harmonics as given in Ref. 8.

Any vector $\mathbf{f}(\theta,\phi)$ defined on a sphere of arbitrary radius can be expanded in terms of vector spherical harmonics,

$$\mathbf{f}(\theta,\phi)=\sum_{JlM} f_{JlM}\mathbf{Y}_{JlM}(\theta,\phi), \qquad (6a)$$

where the coefficients of the expansion f_{JlM} are given by

$$f_{JlM}=\int_0^{2\pi}\int_0^{\pi}\mathbf{Y}^*_{JlM}(\theta,\phi)\cdot\mathbf{f}(\theta,\phi)\sin\theta\, d\theta\, d\phi. \qquad (6b)$$

Let us define $\mathbf{f}_1$ and $\mathbf{f}_2$ by

$$\mathbf{f}_1(\theta,\phi)=\sum_{J,M} f_{JJM}\mathbf{Y}_{JJM}(\theta,\phi),$$
$$\mathbf{f}_2(\theta,\phi)=\sum_{J,M}[f_{J,J+1,M}\mathbf{Y}_{J,J+1,M}(\theta,\phi)+f_{J,J-1,M}\mathbf{Y}_{J,J-1,M}(\theta,\phi)]. \qquad (7a)$$

Clearly,

$$\mathbf{f}(\theta,\phi)=\mathbf{f}_1(\theta,\phi)+\mathbf{f}_2(\theta,\phi). \qquad (7b)$$

We maintain that the decomposition (7b) with $\mathbf{f}_1$ and $\mathbf{f}_2$ given by (7a) is the angular-momentum Helmholtz decomposition, and that $\mathbf{f}_1$ and $\mathbf{f}_2$ satisfy Eq. (4).

To prove this assertion we use the following properties of the vector spherical harmonics:

$$\mathbf{L}\times\mathbf{Y}_{JlM}=\tfrac{1}{2}[J(J+1)-l(l+1)-2]\mathbf{Y}_{JlM},$$
$$\mathbf{L}\cdot\mathbf{Y}_{JlM}=i[l(l+1)]^{1/2}\delta_{Jl}Y_{lM}, \qquad (8)$$

where $Y_{lM}(\theta,\phi)$ are the usual surface harmonics in the notation of Ref. 8, for example. We believe that Eq. (8), which can be proved in various ways from the definition of the vector spherical harmonics, represent new results. That $\mathbf{f}_1$ and $\mathbf{f}_2$ as defined by Eq. (7a) satisfies Eq. (4) now follows from Eq. (8). Also on using Eq. (8) and

$$\mathbf{L}Y_{JM}(\theta,\phi)=i[J(J+1)]^{1/2}\mathbf{Y}_{JJM}(\theta,\phi), \tag{9}$$

which is proved in Ref. 7, we obtain the potentials

$$V(\theta,\phi)=-i\sum_{J=1}^{\infty}\sum_{M=-J}^{J}[J(J+1)]^{-1/2}f_{JJM}Y_{JM}(\theta,\phi)+C,$$

$$\mathbf{A}(\theta,\phi)=\sum_{J,M}[-(J+1)^{-1}f_{J,J+1,M}\mathbf{Y}_{J,J+1,M}(\theta,\phi)+J^{-1}f_{J,J-1,M}\mathbf{Y}_{J,J-1,M}(\theta,\phi)]+\mathbf{L}W(\theta,\phi), \tag{10}$$

where $W(\theta,\phi)$ is an arbitrary function of its arguments and is a kind of "gauge" and C is an arbitrary constant.

Before we leave the subject of the angular-momentum Helmholtz theorem we note the following orthogonality theorem:

$$\int_0^{2\pi}\int_0^{\pi}\mathbf{f}_1^*\cdot\mathbf{f}_2\sin\theta\,d\theta\phi=0. \tag{11}$$

2. DECOMPOSITION OF SOLENOIDAL VECTORS INTO ITS TOROIDAL AND POLOIDAL COMPONENTS

Let us now consider a spherical shell. Let $\mathbf{f}(r,\theta,\phi)$ be a complex vector defined in this shell. (As special cases, the shell may be the entire space, or the space within a sphere or the space external to one.) Let us assume that f is solenoidal, i.e.,

$$\nabla\cdot\mathbf{f}=0. \tag{12}$$

The following theorem is proved in Refs. 9 and 10. The vector f can always be written as the sum of two vectors

$$\mathbf{f}(r,\theta,\phi)=\mathbf{h}_1(r,\theta,\phi)+\mathbf{h}_2(r,\theta,\phi), \tag{13}$$

such that

$$\mathbf{h}_1(r,\theta,\phi)=\mathbf{L}C(r,\theta,\phi),\quad \mathbf{h}_2(r,\theta,\phi)=\nabla\times[\mathbf{L}D(r,\theta,\phi)], \tag{14}$$

where C and D are scalar functions of their arguments. The vector $\mathbf{h}_1$ is called the toroidal part of h, while $\mathbf{h}_2$ is its poloidal part.

Since $\mathbf{f}(r,\theta,\phi)$ is defined on a sphere for every value of r, we can apply the angular-momentum Helmholtz theorem to the vector. We introduce the variable r into $\mathbf{f}_1(r,\theta,\phi)$, $\mathbf{f}_2(r,\theta,\phi)$, $f_{J1M}(r)$, $V(r,\theta,\phi)$, and $\mathbf{A}(r,\theta,\phi)$. We shall show that if $\mathbf{f}(r,\theta,\phi)$ is solenoidal, then there is a close connection between the angular-momentum Helmholtz theorem and the decomposition into toroidal and poloidal components. The relation is the following:

$$\mathbf{h}_1(r,\theta,\phi)=\mathbf{f}_1(r,\theta,\phi),\quad \mathbf{h}_2(r,\theta,\phi)=\mathbf{f}_2(r,\theta,\phi),$$
$$C(r,\theta,\phi)=V(r,\theta,\phi)+K(r), \tag{15}$$

where $K(r)$ is an arbitrary function of r. In proving the relations (15) we shall essentially be proving the solenoidal decomposition theorem, in which the scalar potential $D(r,\theta,\phi)$ will be contructed explicitly.

We first note that $\mathbf{f}_1(r,\theta,\phi)$ is solenoidal. This fact follows from the expansion which is the first of Eq. (7a) and from Eq. (5.9.22) of Ref. 8. Thus a necessary and sufficient condition that $\mathbf{f}$ be solenoidal is that $\mathbf{f}_2$ be so. From Eq. (7a) and from Eqs. (5.9.21) and (5.9.23) of Ref. 8, a necessary and sufficient condition for $\nabla\cdot\mathbf{f}_2=0$ is

$$[J+1]^{1/2}\left[\frac{d}{dr}+\frac{J+2}{r}\right]f_{J,J+1,M}=[J]^{1/2}\left[\frac{d}{dr}-\frac{J-1}{r}\right]f_{J,J-1,M}. \tag{16}$$

Let us define $k_{JM}(r)$ as being a solution of the differential equation

$$\left[\frac{J}{(2J+1)}\right]^{1/2}\left[\frac{d}{dr}-\frac{J}{r}\right]k_{JM}=f_{J,J+1,M}. \tag{17}$$

The solutions are not unique. If k_{JM} is one solution, the general solution is

$$k_{JM}^{(1)}(r)=k_{JM}(r)+D_{JM}r^J, \tag{18}$$

where D_{JM} is an arbitrary constant.

Let us write

$$f_{J,J-1,M}(r)=\left[\frac{(J+1)}{(2J+1)}\right]^{1/2}\left[\frac{d}{dr}+\frac{J+1}{r}\right]k_{JM}(r)+C_{JM}(r). \tag{19}$$

On using Eqs. (17) and (19) in Eq. (16), we obtain a simple differential equation for the function $C_{JM}(r)$,

$$\left[\frac{d}{dr}-\frac{J-1}{r}\right]C_{JM}(r)=0, \tag{20}$$

for which the solution is

$$C_{JM}(r)=C_{JM}r^{J-1}, \tag{21}$$

where the constants C_{JM} are arbitrary.

Instead of using $k_{JM}(r)$ in Eqs. (17) and (19), we may use the function $k_{JM}^{(1)}(r)$ of Eq. (18). It is easily seen that the constants D_{JM} may be so chosen as to cancel the constants C_{JM}. We thus have the following theorem:

The vector $\mathbf{f}_2(r,\theta,\phi)$ is solenoidal if and only if functions $k_{JM}(r)$ can be found such that

$$f_{J,J+1,M}(r)=\left[\frac{J}{(2J+1)}\right]^{1/2}\left[\frac{d}{dr}-\frac{J}{r}\right]k_{JM}(r),$$
$$f_{J,J-1,M}(r)=\left[\frac{(J+1)}{(2J+1)}\right]^{1/2}\left[\frac{d}{dr}+\frac{J+1}{r}\right]k_{JM}(r). \tag{22}$$

Let us now use Eq. (22) in Eq. (7a) to obtain

$$\mathbf{f}_2=\sum_{J,M}\left\{\left[\frac{(J+1)}{(2J+1}\right]^{1/2}\mathbf{Y}_{J,J+1,M}\left[\frac{d}{dr}-\frac{J}{r}\right]k_{JM}+\left[\frac{(J+1)}{(2J+1)}\right]^{1/2}\mathbf{Y}_{J,J-1,M}\left[\frac{d}{dr}+\frac{J+1}{r}\right]k_{JM}\right. . \tag{23}$$

Hence, on using Eqs. (5.9.19) of Ref. 8,

$$\mathbf{f}_2=-i\nabla\times\mathbf{B},\quad \mathbf{B}\equiv\mathbf{B}(r,\theta,\phi)=\sum_{J=1}^{\infty}\sum_{M=-J}^{J}k_{JM}(r)\mathbf{Y}_{JJM}(\theta,\phi). \tag{24}$$

The first of Eq. (24) is, of course, part of the usual Helmholtz theorem which states that a solenoidal vector can be expressed as the curl of a vector potential. However, the particular form of the vector potential given by the second of Eq. (24) leads to

$$\mathbf{B}(r:\theta,\phi)=i\mathbf{L}D(r,\theta,\phi),$$
$$D(r,\theta,\phi)=-\sum_{J=1}^{\infty}\sum_{M=-J}^{J}[J(J+1)]^{-1/2}Y_{JM}(\theta,\phi)k_{JM}(r) \tag{25}$$

when one uses Eq. (9).

The use of Eq. (25) in Eq. (24) leads to the identification of $\mathbf{f}_2$ with $\mathbf{h}_2$. The identification of $\mathbf{f}_1$ with $\mathbf{h}_1$ is trivial.

Because of this identification when f is solenoidal, it seems reasonable to call $\mathbf{f}_1$ and $\mathbf{f}_2$ of the angular-momentum Helmholtz theorem the toroidal and poloidal components (in a more general sense) even when f is *not* solenoidal.

[1]J. S. Lomont and H. E. Moses, Nuovo Cimento **16**, 96 (1960).
[2]J. S. Lomont and H. E. Moses, Commun. Pure Appl. Math. **14**, 69 (1961).
[3]J. B. Keller, Commun. Pure Appl. Math. **14**, 77 (1961).
[4]J. T. Cannon and T. F. Jordan, J. Math. Phys. **5**, 299 (1964).
[5]H. Leutwyler, Nuovo Cimento **37**, 543, 556 (1965).
[6]H. Stetkaer, Trans. Am. Math. Soc. **214**, 349 (1975).
[7]J. M. Blatt and V. F. Weisskopf, *Theoretical Nuclear Physics* (Wiley, New York, 1952).
[8]A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U. P., Princeton, New Jersey, 1960).
[9]W. M. Elasser, Phys. Rev. **69**, 106 (1946).
[10]S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability* (Oxford U. P., Oxford, 1961).

Complete extension of the symmetry axis of the Tomimatsu–Sato solution of the Einstein equations

C. Reina

Istituto di Scienze Fisiche dell'Università, Milano, Italy
(Received 9 January 1976; revised manuscript received 12 April 1976)

The symmetry axis of the simplest Tomimatsu–Sato field is considered. Since this manifold is not geodesically complete for every value of the parameters occurring in the metric, a complete extension is given, and it is shown that its causal structure is very similar to that of the symmetry axis of the Kerr field.

1. INTRODUCTION

Ten years after the discovery by Kerr[1] of the first axisymmetric rotating solution of Einstein's equations, new rotating fields were found by Tomimatsu and Sato.[2,3] The main difference between these two classes of solutions is that the quadrupole moment of the T–S fields is larger than that of the Kerr field, and that the former solutions exhibit a number of ring singularities among which the outermost is unshielded by an event horizon.

The structure of these manifolds has been investigated by several authors, who have considered the geodesic problem[3,4] and, in the case of the simplest T–S metric, the behavior of the metric near the poles ($x=1$, $y=\pm 1$ in prolate spheroidal coordinates).[5,6] In particular, Ernst[5] introduced a new representation of this T–S metric, showing that the full four-dimensional geodesic problem can be completely solved in the neighborhood of the poles.

In this paper the bidimensional metric on the axis of the simplest T–S field is studied, using extensively the method adopted by Carter[7] in the case of the axis of the Kerr solution. Although this problem is rather more restricted than the maximal extension of the full four-dimensional metric, it is nevertheless significant to have found a complete extension of the bidimensional metric which is exact and C_0 on its domain.

In spite of the differences between the Kerr and T–S solutions, it is found that they have a very similar causal structure when restricted to the axis.

2. T-S FIELD IN QUASISPHEROIDAL COORDINATES

The axisymmetric line element in canonical coordinates reads

$$ds^2 = f^{-1}[e^{2\gamma}(d\rho^2 + dz^2) + \rho^2 d\varphi^2] - f(dt - \omega\, d\varphi)^2, \tag{1}$$

where f, ω, γ are functions of ρ and z only. In prolate spheroidal coordinates (x, y) defined by the mapping

$$\begin{cases} \rho = k(x^2-1)^{1/2}(1-y^2)^{1/2}, \\ z = kxy, \end{cases} \tag{2}$$

the line element (1) takes the form

$$ds^2 = k^2 f^{-1}\left[e^{2\gamma}(x^2-y^2)\left(\frac{dx^2}{x^2-1} + \frac{dy^2}{1-y^2}\right) + (x^2-1)(1-y^2)\,d\varphi^2\right] - f(dt - \omega d\varphi)^2. \tag{3}$$

For the T–S fields the metric functions f, γ, ω are expressed in terms of three polynomials A, B, C of x, y in the following way[3]:

$$f = \frac{A}{B}, \quad \omega = \frac{2mq}{A}(1-y^2)C, \quad e^{2\gamma} = \frac{A}{p^{2\delta}(x^2-y^2)^{\delta^2}} \tag{4}$$

where m is a parameter describing the mass of the source of the field, p and q are real constants subjected to the condition $p^2+q^2=1$, and δ is an integer parameter taking the values 2, 3, 4. The explicit expressions for A, B, C depend on the chosen value of δ. The simplest T–S metric was obtained[2] for $\delta=2$, yielding

$$\begin{aligned} A &= p^4(x^2-1)^4 + q^4(1-y^2)^4 - 2p^2q^2(x^2-1)(1-y^2) \\ &\quad \times [2(x^2-1)^2 + 2(1-y^2)^2 + 3(x^2-1)(1-y^2)], \\ B &= [p^2(x^2+1)(x^2-1) - q^2(y^2+1)(1-y^2) + 2px(x^2-1)]^2 \\ &\quad + 4q^2y^2[px(x^2-1) + (px+1)(1-y^2)]^2, \\ C &= p^3x(x^2-1)[2(x^2+1)(x^2-1) + (x^2+3)(1-y^2)] \\ &\quad - p^2(x^2-1)[4x^2(x^2-1) + (3x^2-1)(1-y^2)] \\ &\quad + q^2(px+1)(1-y^2)^3. \end{aligned} \tag{5}$$

For the explicit form of the polynomials A, B, C for $\delta=3$ the work by T–S is cited.[3]

In the following it will be useful to work in quasi-spheroidal coordinates (r, θ) defined by

$$\frac{1}{\delta}px + 1 = \frac{r}{m}, \quad y = \cos\theta. \tag{6}$$

The line element (3) becomes

$$\begin{aligned} ds^2 &= \frac{\tilde{B}}{[(r-m)^2 - H^2\cos^2\theta]^{\delta^2-1}}\left(\frac{dr^2}{r^2 - 2mr + \alpha^2} + d\theta^2\right) \\ &\quad + \frac{B}{A}(r^2 - 2mr + \alpha^2)\sin^2\theta\, d\varphi^2 \\ &\quad - \frac{A}{B}\left(dt - \frac{2mq}{A}\sin^2\theta \cdot C\, d\varphi\right)^2, \end{aligned} \tag{7}$$

where the arbitrariness of the scale k is used in order to put $k = H = pm/\delta$, and where

$$\alpha = m^2 - H^2, \quad \tilde{B} = (H^{2\delta^2}/p^{2\delta})B. \tag{8}$$

On the symmetry axis $\sin\theta = 0$ (i.e., $y^2=1$) this metric reduces to the form

$$ds^2_{\text{ax}} = \frac{\rho^4}{\Delta^4}dr^2 - \frac{\Delta^4}{\rho^4}dt^2, \tag{9}$$

where

$$\rho^4 = (r-m)^2(r^2+m^2) + (m^2-\alpha^2)(r^2-\alpha^2), \tag{10}$$

$$\Delta^4 = (r^2 - 2mr + \alpha^2)^2. \tag{11}$$

Note that for $m \neq \alpha$, $\rho^4 > 0$ for all r, while for $m = \alpha$, $\rho^4 = 0$ for $r = m$. Since Δ^4 has no real zeros for $m^2 < \alpha^2$, in this case the manifold $-\infty < r < +\infty$, $-\infty < t < +\infty$, with the metric (9) is complete.

3. GEODESIC COMPLETENESS

In order to see the necessity of an extension for $m^2 \geq \alpha^2$, we introduce new coordinates t', r' defined by

$$dt' = dt' + \frac{\Delta^4 - \rho^4}{\Delta^4}\, dr', \quad dr = dr'. \tag{12}$$

The metric becomes

$$ds^2_{\text{ax}} = (1+\tau)\, dr'^2 + 2\tau\, dr'\, dt' - (1-\tau)\, dt'^2, \tag{13}$$

where

$$\tau = \frac{2mr(r-m)^2 + (m^2-\alpha^2)[(r-m)^2 - 2(mr-\alpha^2)]}{(r^2+m^2)(r-m)^2 + (m^2-\alpha^2)(r^2-\alpha^2)}. \tag{13a}$$

Upon introducing a null coordinate u such that

$$t' = u - r', \quad r = r', \tag{14}$$

as was done by Finkelstein for the Schwarzschild manifold, the metric (13) becomes

$$ds^2_{\text{ax}} = -(1-\tau)\, du^2 + 2\, du\, dr. \tag{15}$$

Geodesic trajectories can be derived from the Lagrangian

$$\mathcal{L} = \tfrac{1}{2}[2\dot{u}\dot{r} - (1-\tau)\dot{u}^2], \tag{16}$$

where the dot indicates the derivative with respect to an affine parameter, λ say. The Euler–Lagrange equation obtained varying the action with respect to u is immediately integrated, giving

$$-(1-\tau)\dot{u} + \dot{r} = -E,$$

where E is a constant. This equation together with the normalization condition $\mathcal{L} = \epsilon$ ($\epsilon = 0, \pm 1$ for null, spacelike and timelike geodesics, respectively) yields the two equations

$$\dot{u} = \frac{E \pm \sqrt{E^2 + \epsilon(1-\tau)}}{(1-\tau)},$$

$$\dot{r} = \pm\sqrt{E^2 + \epsilon(1-\tau)}.$$

For $m^2 < \alpha^2$, the expression $1 - \tau = \rho^4/\Delta^4$ has no real zeros and $\dot{u}$, $\dot{r}$ are bounded functions of r. This implies that each geodesic $u(\lambda)$, $r(\lambda)$ can be continued to arbitrary values of the affine parameter λ. Therefore, the manifold $-\infty < r < +\infty$, $-\infty < t < +\infty$, with the metric (15) is geodesically complete in this case.

For $m^2 > \alpha^2$, u diverges at $r = r_\pm$, and the manifold is incomplete. This can be shown explicitly for null geodesics ($\epsilon = 0$). Redefining λ so that $E = 1$, one has the following equations for "ingoing" and "outgoing" goedesics:

$$u = C_1, \quad r = -\lambda, \quad u = C_2 + F(r), \quad r = \lambda,$$

where C_1, C_2 are constants, and

$$F(r) = \int \frac{2\,dr}{1-\tau} = 2r + D_1 \ln|r - r_+| - D_2 \ln|r - r_-| - D_3 \frac{r}{(r-r_+)(r-r_-)}, \tag{17}$$

with

$$D_1 = 2\left(\frac{2m^2-\alpha^2}{(m^2-\alpha^2)^{1/2}} + m\right),$$

$$D_2 = 2\left(\frac{2m^2-\alpha^2}{(m^2-\alpha^2)^{1/2}} - m\right),$$

$$D_3 = 2(m^2-\alpha^2).$$

Since $F(r)$ diverges for $r = r_\pm$, outgoing geodesics cannot penetrate the surface $r = r_+$.

For $m^2 = \alpha^2$, it has been shown[8] that all T—S spaces are equivalent to extreme Kerr ($m^2 = a^2$), so the complete extension has been given already[7] for this case.

4. COMPLETE EXTENSION

Introduce now a second null coordinate w, defined by

$$F(r) = u + w. \tag{18}$$

Since $F(r)$ is monotonic in the regions,

I: $r > r_+$,

II: $r_+ > r > r_-$,

III: $r_- > r$,

one must specify to which region one is referring, in order that the mapping (18) be well defined. With the coordinates u, w the metric assumes the canonical double null form

$$ds^2_{\text{ax}} = (1-\tau)\, dw\, du, \tag{19}$$

where again the factor $1-\tau$ is degenerate at $r = r_\pm$.

Following Carter, one can introduce the manifold $\mathcal{M}^*$ spanned by coordinates ψ, ξ, ranging from $-\infty$ to $+\infty$. Let r_n, a_m be the lines

$$\begin{aligned} &r_n) \quad \psi = -\xi + \pi/2 + n\pi, \\ &a_m) \quad \psi = \xi + \pi/2 + m\pi \\ &\qquad (m, n = 0, \pm 1, \pm 2, \cdots), \end{aligned} \tag{20}$$

and let Q_{nm} be the intersections of the two strips bounded by the lines a_m, a_{m-1} and r_n, r_{n-1}, respectively. The ψ, ξ coordinates are defined by the relations

$$\begin{aligned} u &= \tan\tfrac{1}{2}(\psi + \xi), \\ w &= \cot\tfrac{1}{2}(\psi - \xi). \end{aligned} \tag{21}$$

The squares Q_{kk} are images of the region II, $Q_{h,h-1}$ are images of the region III if h is odd and are images of the region I if h is even, and finally the squares $Q_{j-1,j}$ are images of the region I if j is odd and of the region III if j is even.

The metric becomes

$$ds^2_{\text{ax}} = \Omega^2\, d\psi\, d\xi \tag{22}$$

where

$$\Omega^2 = (1-\tau)\, \tfrac{1}{4}\sec^2\left(\frac{\psi+\xi}{2}\right) \csc^2\left(\frac{\psi-\xi}{2}\right).$$

It can be easily shown that this conformal factor is continuous and positive definite on the manifold $\mathcal{M}^*$.

The conformal diagram for the axis of the T—S solution considered here is identical to that for the Kerr axis.[4] Therefore, the two axis have the same causal structure, i. e., they are conformally related.

ACKNOWLEDGMENTS

I am grateful to the referee of this paper for constructive suggestions.

[1]P. Kerr, Phys. Rev. Lett. **11**, 238 (1963).
[2]A. Tomimatsu and H. Sato, Phys. Rev. Lett. **23**, 1344 (1972).
[3]A. Tomimatsu and H. Sato, Progr. Theor. Phys. **50**, 95 (1973).
[4]G.W. Gibbons and R.A.Russell-Clark, Phys. Rev. Lett. **30**, 398 (1973).
[5]F.J. Ernst, J. Math. Phys. (to be published).
[6]J.E. Economou and F.J. Ernst, J. Math. Phys. **17**, 52 (1976).
[7]B. Carter, Phys. Rev. **141**, 1242 (1966).
[8]W. Kinnersley and E.F. Kelley, J. Math. Phys. **15**, 2121 (1974).

Gauge theories for space–time symmetries. I*

J. P. Harnad†

Department of Physics, Carleton University, Ottawa, Ontario, Canada

R. B. Pettitt

Department of Mathematics, University of Ottawa, Ottawa, Ontario, Canada

(Received 19 December 1975)

The general formulation of gauge invariant field theories based upon space–time symmetries is developed and given its geometrical interpretation. The consequences of gauge invariance, in the form of identities and conservation laws, are derived and the field equations are obtained from a class of gauge invariant Lagrangians. This is the first paper of a series, the subsequent work treating specific cases, in particular, conformal invariance.

1. INTRODUCTION

An upsurge of interest has occurred in recent years with regard to gauge field theories; notably, in the search for a unifying framework for the fundamental interactions of elementary particles. The concept of gauge transformations as related to matter fields (classical or quantum) may be traced to the work of Weyl.[1] The term was introduced in his formulation of a unified field theory of electromagnetism and gravitation (1919) and the concept was carried into the domain of quantum mechanics in his study of the electromagnetic interactions of Dirac electrons (1929). The gravitational interaction of the Dirac electron as a classical field was itself treated by Weyl (1950) as a gauge field theory for the Lorentz group acting upon locally Lorentzian frames (vierbeins). Non-Abelian gauge theories associated with internal symmetry groups were treated by Yang and Mills[2] (1954) whose work together with that of Weyl was generalized by Utiyama[3] (1956) to include arbitrary internal symmetries and the gravitational interaction for matter fields of any spin. Sciama[4] (1961) provided further insight into the vierbein formalism within the Palatini approach in which vierbeins and gauge fields are subjected to independent Euler–Lagrange variations in obtaining the field equations from a Lagrangian. When given a space–time geometrical interpretation, the emerging structure is seen to be a slight generalization of the Riemannian geometry of general relativity to a geometry in which the affine connection has an asymmetric part (torsion) which is related in the dynamical scheme to the intrinsic spin current of matter. This type of theory was originally proposed by Cartan[5] (1924) and its ramifications have been recently subjected to considerable study.[6,7] In particular, the work of Trautman[8] has set the theory within a clear geometrical setting, utilizing the coordinate free notations of modern differential geometry.

A formulation of this gauge theory based upon the *inhomogeneous* Lorentz group was obtained by Kibble[9] (1961). His approach more closely resembles the theories of the Yang–Mills type in that no *a priori* geometrical objects (like the vierbein fields) need to be introduced. Instead, one begins in a flat space–time with Lorentz invariant interaction of matter fields and introduces gauge fields associated with the full inhomogeneous Lorentz group, giving them only an *a posteriori* interpretation in geometrical terms. The gauge fields in this context can be understood as a Cartan connection for the principal bundle with the inhomogeneous Lorentz group as structure group with the canonical choice[10] for the part of the connection lying in the translation subalgebra. This choice implies that the vierbein fields themselves (or rather their duals) become the gauge fields associated with translations. The corresponding "curvature" is the torsion. As emphasized by Sciama,[11] any attempt to extend the geometrical structure of space–time so as to include additional gauge fields would necessarily involve an extension of the class of infinitesimal holonomy groups. Within the context of linear frames, this may only involve linear or affine structure groups, possibly complex ones. In particular, if scale transformations are included, the emerging theory is of the Weyl (1919) type.[12] Other alternatives investigated by Sciama[13] were the group of complex unitary transformations and the symplectic group. Numerous attempts to formulate a unified gauge theory based upon combined space–time and internal symmetries have also been made.[14,15] Of particular interest within the context of gauge theories of the Yang–Mills type are the unified theories of weak and electromagnetic (and possibly strong) interactions of Weinberg[16] and Salam.[17]

One fault of all these approaches [as was pointed out in particular by Sciama[18] (1961) and Weinberg[19] (1967)] is that the underlying symmetry groups are not simple and therefore independent invariants may be formed from subclasses of gauge fields. This means that the underlying gauge theory is not a genuinely unified one, even if, being based upon space–time symmetries, it is a "geometrized" one. If one wishes to modify the structure of the Weyl theory by imbedding the group of Lorentz plus scale-transformations (the "Weyl group") in one which is simple, the most natural choice is the conformal group. This is a particularly reasonable extension since, as shown by Mack and Salam[20] (1968), any Lagrangian which is invariant under the Weyl group (in Minkowski space) is automatically invariant under the conformal group, provided no derivative couplings are involved. However, such extensions involve new complications since the transformations realized in a four-dimensional manifold are no longer linear, and therefore the methods of Weyl, Sciama, Utiyama, and

Kibble may not be applied without suitable generalizations. In particular, if we wish the infinitesimal holonomy group to include all possible conformal transformations, we cannot utilize merely linear, orthonormal frames (vierbeins) but must turn to second order frames in order to define completely the notion of infinitesimal parallel transfer.

The present paper is the first of a series applying the gauge method to a wide class of space–time (and/or internal) symmetry transformations, including all special cases previously treated in the literature. We first establish the general formalism for gauge theories based upon space–time transformation groups, the identities and conservation laws following from the requirements of invariance, and the field equations obtained from a class of invariant gauge-field Lagrangians. In Part A the development parallels and generalizes the work of Sciama and Kibble, the notation being explicitly coordinate dependent, and the geometrical interpretation deliberately avoided. In Part B the notions of gauge groups and fields are related to the differential geometric concepts of connections in principal G-bundles. All the fields and geometrical entities are expressed first in the coordinate free notation of differential geometry and are then related to the notations of the gauge approach. The gauge fields, vierbeins, and matter fields together with their transformation properties are given their respective geometrical interpretations. The field equations and identities are then expressed in the more compact notation of the calculus of exterior differential forms.

In the subsequent paper we shall introduce the second order frame structure, which is the suitable generalization of vierbeins needed to study the gauge conformal group. The physical interpretation of the geometrical ideas involved will be examined in greater detail, as will the identities and field equations associated with conformally gauge invariant theories.

PART A

2. REPRESENTATIONS, SYMMETRIES AND CONSERVATION LAWS

Consider a group G whose elements can be realized (at least locally) as transformations acting on the space–time manifold and characterized by a set of n parameters $\{\epsilon^a\}_{a=1,\ldots,n}$ such that the infinitesimal coordinate transformations may be written

$$x^\mu \xrightarrow{g} x'^\mu \simeq x^\mu + \xi^\mu, \tag{2.1}$$

$$\xi^\mu = \epsilon^a \zeta^\mu_a. \tag{2.2}$$

Here ζ^μ_a are n linearly independent vector functions of x. As usual, Eq. (2.1) may be viewed either as a passive (coordinate) transformation assigning new coordinate labels to the same point or as an active (point) transformation mapping the point with coordinates x^μ into the one with coordinates x'^μ. We shall, in what follows, generally adopt the passive view. Equation (2.1) may also be taken to represent any arbitrary point or coordinate transformation independently of the identification (2.2) as an action of the group G.

Let us furthermore consider a subgroup $G_0 \subset G$ of dimension $m \leq n$. The group G_0 will be assumed to have a representation acting upon the vector space V in which all the fields $\{\psi_A\}_{A=1,\ldots,r}$ take their values (the subscript A distinguishing the fields will be suppressed with a summation convention understood whenever bilinear combinations occur). The group G_0 is interpreted as the internal part of G which acts not only upon the space–time coordinates, but also upon the field components (e.g., internal spin). The representations of G_0 acting on V will be denoted by $\{\rho(g) : g \in G_0,\ \rho(g) : V \to V\}$.

The group G_0 may be a proper subgroup of G (e.g., the homogeneous Lorentz subgroup of the Poincaré group) or it may simply be G itself. In the event that G can be decomposed into a semi-direct product of G_0 with some other subgroup

$$G = Z \times_\tau G_0, \tag{2.3}$$

where $Z \cap G_0 = \{e\}$ and $GZG^{-1} = Z$, the representation $\rho(G_0)$ may also be considered as a representation of G itself, with the elements of the subgroup Z being simply mapped into the identity. In this trivial sense, the group G_0 may be replaced by the whole group G considered as acting upon the field components. Another decomposition which can sometimes be made is the identification of G_0 as the group generated by all those infinitesimal transformations $\mathring{\zeta}^\mu_a$, for which

$$\mathring{\zeta}^\mu_a (x = 0) = 0. \tag{2.4}$$

This defines G_0 as the isotropy subgroup at $x = 0$. As may be seen by making a Taylor expansion of ξ^μ about this point, such a G_0 has dimension $(n-4)$ or greater, since there can only be 4 independent vectors η^μ in the expansion

$$\xi^\mu = \eta^\mu + \epsilon^a \mathring{\zeta}^\mu_a \tag{2.5}$$

and these may be identified with the parameters of a four- (or fewer) dimensional Abelian subgroup of G (coordinate translations), the remaining parameters labeling the infinitesimal transformations of G_0.

The field representations will be assumed to transform under the combined actions of G upon coordinates and G_0 upon the field components as follows for $g \in G$:

$$\psi(x) \xrightarrow{g} \rho(g_0(g,x))\, \psi(g^{-1}x). \tag{2.6}$$

where

$$g_0(g,x) \in G_0 \tag{2.7}$$

and

$$g_0(g'g,x) = g_0(g',x)\, g_0(g, g'^{-1}x). \tag{2.8}$$

We should like to emphasize that the transformation property (2.6), which defines a representation of G upon *fields* given a representation $\rho(G_0)$ of the subgroup G_0 within the vector space V, could have been obtained in various ways. In particular, the method of induced representations[21] gives rise to a transformation property of type (2.6), as does the method of nonlinear realizations,[22] and, as we shall see in the sequel, such a relation arises in a very natural way within the formalism of higher order frames.

In certain simple cases, the dependence upon x in $g_0(g,x)$ is absent, in which case Eq. (2.8) shows that this defines a group homomorphism $G \to G_0$. In particular, this is obviously the case when $G=G_0$, where the representation (2.6) reduces simply to a direct product of the representation $\rho(G_0)$ with the quasiregular representation of G. If G has the semidirect product structure (2.3), this may also be done [extending the representation $\rho(G_0)$ to $\rho(G)$ in the trivial manner discussed above]. If G happens to be a simple group (e.g., the conformal group), no such homomorphism to a group of smaller dimension can exist, and the x dependence in $g_0(g,x)$ must necessarily remain.

At the level of infinitesimal transformations, Eq. (2.6) reduces to

$$\delta_0\psi \equiv \psi'(x)-\psi(x) \simeq \epsilon^a \hat{T}_a \psi, \tag{2.9}$$

where the $\hat{T}_a$ are the infinitesimal generators of the representation (2.6) of G.

Thus the $\hat{T}_a$ form a basis for the corresponding representation of the algebra $\mathcal{G}$, and we have

$$\hat{T}_a = \mathring{\beta}_a^b \mathring{T}_b - \xi_a^\mu \partial_\mu \tag{2.10}$$

with Lie brackets

$$[\hat{T}_a, \hat{T}_b] = f_{ab}^c \hat{T}_c, \tag{2.11}$$

where f_{ab}^c are the structure constants of $\mathcal{G}$ within this basis, $\beta_a^b(x)$ are functions of the coordinates x determined by the infinitesimal form of $\rho(g_0(g,x))$, and $\{\mathring{T}_a\}_{a=1,\dots,m}$ designate the basis elements of the representation $\rho(\mathcal{G}_0)$, where $\mathcal{G}_0$ is the Lie algebra of G_0. We have chosen to define the basis for $\mathcal{G}$ by extending the basis for $\mathcal{G}_0$ in order that corresponding components may be labeled with the same indices, with the understanding that a degree sign ° over any summed quantity indicates a restriction of the range of summation to the subspace spanned by $\{\mathring{T}_a\}$. Thus we may write the commutation relations within this subrepresentation as

$$[\mathring{T}_a, \mathring{T}_b] = \mathring{f}_{ab}^c \mathring{T}_c. \tag{2.12}$$

The linear differential operators $-\xi_a^\mu \partial_\mu$ generate the quasiregular representation of G and hence satisfy

$$[\xi_a^\mu \partial_\mu, \xi_b^\nu \partial_\nu] = -f_{ab}^c \xi_c^\mu \partial_\mu. \tag{2.13}$$

Combining Eqs. (2.10)–(2.13) gives us the following differential equation that must be satisfied by the β_a^b's in order that $\hat{T}_a$ really generate a representation:

$$\mathring{\beta}_a^c \mathring{\beta}_b^d \mathring{f}_{cd}^e - \xi_{[a}^\nu \mathring{\beta}_{b],\nu}^e = \mathring{f}_{ab}^c \mathring{\beta}_c^e, \tag{2.14}$$

where the symbol [ ] denotes antisymmetrization with regard to the adjacent, included indices. This relation is just the infinitesimal form of Eq. (2.8). [We note in passing that the form of the β_a^b's is determined by the group and does not depend on the representation $\rho(G)$ involved. Therefore a solution of (2.14) allows one to immediately give the representation of $\mathcal{G}$ induced by *any* given representation of $\mathcal{G}_0$.]

Under the changes of coordinates (2.1) induced by $g\in G$, the infinitesimal change in the field ψ, considered at the same point (with new coordinates $x'=gx$) is given by

$$\delta\psi \equiv \psi'(x')-\psi(x) \simeq \mathring{\chi}^a \mathring{T}_a \psi, \tag{2.15}$$

where

$$\mathring{\chi}^a \equiv \epsilon^b \mathring{\beta}_b^a(x). \tag{2.16}$$

The derivative of the field transforms as follows:

$$\delta(\partial_\mu\psi) = \partial_\mu(\delta\psi) - \partial_\nu\psi\partial_\mu\xi^\nu = \partial_\mu \mathring{\chi}^a \mathring{T}_a\psi + \mathring{\chi}^a \mathring{T}_a \partial_\mu\psi - \partial_\nu\psi\partial_\mu\xi^\nu. \tag{2.17}$$

We note that the first term in (2.17) vanishes for the case when β_b^a *and* ϵ^a are constants, independent of x. If β_b^a is x dependent, $\partial_\mu\psi$ does not have a linear homogeneous transformation property like that of ψ, even under the rigid action of the group (constant ϵ^a). In this case, it is not possible to form Lagrangians out of ψ and $\partial_\mu\psi$ alone which are *manifestly* invariant under the actions of the group. If we are considering *nonrigid* actions of the group, that is, if the ϵ^a are allowed to vary from point to point, then $\partial_\mu\psi$ does not enjoy linear homogeneous transformation properties in any event, and we are led to the construction of "covariant" derivatives which do. This we defer until the next section, and merely assert here that, even for x-dependent β_a^b, it is possible to form Lagrangians which are invariant (though not necessarily *manifestly* invariant) under the rigid action of the group.[22]

Let us suppose, then, that there is a Lagrangian density $L(\psi, \partial_\mu\psi)$ depending on the fields and their first derivatives only, such that the corresponding action integral gives rise, through the variational principle, to field equations which are form invariant under transformations of the fields and coordinates [Eqs. (2.15), (2.1)] generated by G. A sufficient condition for this to hold is that the corresponding change in the Lagrangian satisfy

$$\delta L + \epsilon^a \partial_\mu \xi_a^\mu L = 0. \tag{2.18}$$

[The second term in (2.18) is present due to the Jacobian factor involved in the transformation of the volume element.] This gives rise, in the usual way, to the identities

$$L_{[\psi]}\delta_0\psi + \partial_\mu\left[\frac{\partial L}{\partial\partial_\mu\psi}\,\delta\psi + T_\nu^\mu \xi^\nu\right] = 0, \tag{2.19}$$

where

$$L_{[\psi]} \equiv \frac{\partial L}{\partial\psi} - \partial_\mu\left[\frac{\partial L}{\partial\partial_\mu\psi}\right] \tag{2.20}$$

is the Eulerian derivative of L with respect to ψ and

$$T_\nu^\mu = -\delta_\nu^\mu L + \frac{\partial L}{\partial\partial_\mu\psi}\partial_\nu\psi \tag{2.21}$$

is the canonical energy–momentum tensor. Equation (2.19) is valid for any variation $\delta\psi$ in the fields and ξ^μ in the coordinates under which L transforms as a scalar density. Substituting (2.2) and (2.15) with arbitrary constant ϵ^a's and assuming the Euler–Lagrange equations to hold for all ψ's,

$$L_{[\psi]} = 0,$$

we obtain the conservation laws

$$\partial_\mu J_a^\mu = 0, \tag{2.22}$$

where

$$J_a^\mu \equiv \mathring{\beta}_a^b \mathring{S}_b^\mu + T_\nu^\mu \xi_a^\nu, \tag{2.23}$$

$$S_a^\mu \equiv \frac{\partial L}{\partial\partial_\mu\psi}\,\mathring{T}_a\psi. \tag{2.24}$$

Equation (2.23) is interpreted as defining the conserved current as a combination of an "intrinsic" part S_a^μ and an orbital part $T_\nu^\mu \zeta_a^\nu$. The fact that the resulting total current is not simply the sum of S_a^μ plus the orbital part is related, as we shall see in the sequel, to the nonlinearity of the group action upon space—time coordinates.

A slight generalization of (2.18) which also implies invariance of the field equations is

$$\delta L + \epsilon^a \partial_\mu \zeta_a^\mu L = \partial_\mu (\epsilon^a F_a^\mu), \tag{2.25}$$

where F_a^μ is some set of vector fields whose Eulerian variation vanishes outside the same domain as that of ψ. In this case, the conservation laws hold for the modified currents:

$$J'^\mu_a \equiv J_a^\mu - F_a^\mu. \tag{2.26}$$

3. GAUGE INVARIANCE

We now wish to investigate the extension of the invariance properties considered above to include arbitrary, nonrigid group transformations and coordinate changes. The procedure for forming invariant Lagrangians for this wider class of "gauge" transformations is well known for the case of internal symmetries and for linear space—time symmetries such as Lorentz invariance.[23] One starts with a Lagrangian invariant under the rigid transformations and replaces the terms involving derivatives of the fields by suitably defined "covariant derivatives" through the introduction of minimally coupled gauge fields. If the Jacobian determinant of the coordinate transformation is not unity, a further slight modification must be made to ensure that the resulting Lagrangian transforms as a scalar density under the wider class of transformations. The procedure used for the case of linear (e.g., Lorentz) transformations may be extended to our more general class of groups provided the following condition is met. When identifying the group element $g_0(g,x)$ acting upon the "internal" space V of field components, it is essential that the *linear* part (if any) of this transformation be simply given by the Jacobian matrix of the transformation (2.2). That is, we must have (for infinitesimal transformations with constant ϵ^a)

$$\chi_j^i = \epsilon^a \beta_a^b = \epsilon^a \zeta_{a,j}^i, \tag{3.1}$$

where χ_j^i is the infinitesimal parameter corresponding to a linear one-parameter subgroup of G_0 labelled by the index $b \equiv \binom{i}{j}$. The corresponding identification of β_a^b as the derivative of ζ_a^i arises in a natural way within the context of frames of higher order, as we shall see in the sequel. Assuming (3.1) to hold, we then may rewrite the transformation property of a Lagrangian which is an invariant density under the action of the group upon the fields and coordinates as [cf. Eq. (2.18)]

$$\delta L + \chi_i^i L = 0. \tag{3.2}$$

It is now straightforward to apply the minimal coupling procedure, provided the β_a^b of the previous section are x independent. If not, as we have seen above, the ordinary derivatives do not themselves have linear homogeneous transformation properties even under the rigid actions of the group. In a sense, as we may see from Eqs. (2.15)—(2.17), the rigid action of the group upon space—time coordinates gives rise to a nonrigid action upon the fields. However, it is possible to apply a similar procedure starting with a Lagrangian that has already been defined in a manifestly invariant manner with regard to rigid group transformations. The procedure for forming such Lagrangians with or without the introduction of subsidiary (Goldstone) fields has been studied extensively[24] through the methods of nonlinear group realizations. The essential result for our purposes is that it is possible to modify the derivatives $\partial_\mu \psi$ suitably so as to define covariant derivatives $d_\mu \psi$ for *rigid* actions of the group, with the linear homogeneous transformation property

$$\begin{aligned}\delta(d_\mu \psi) &= \epsilon^a \beta_a^b \mathring{T}_b \psi - \epsilon^a \zeta_{a,\mu}^\nu d_\nu \psi \\ &= \chi^a \mathring{T}_a \psi - \chi_\mu^\nu d_\nu \psi.\end{aligned} \tag{3.3}$$

We may then use an invariant Lagrangian $L(\psi, d_\mu \psi)$ satisfying condition (3.2) and substitute for $d_\mu \psi$ a new covariant derivative $\psi_{;i}$ such that under any *nonrigid* action of the group the same transformation property is enjoyed by $\psi_{;i}$:

$$\delta \psi_{;i} = \chi^a \mathring{T}_a \psi - \chi_i^j \psi_{;j}. \tag{3.4}$$

(The use of Roman rather than Greek indices will become clarified later on.) An invariant Lagrangian which is a scalar under coordinate changes may then be formed in a straightforward way.

We first note that since the ϵ^a's are now to have arbitrary space—time dependence we may just as well choose as our independent functions the set $\{\chi^a, \xi^\mu\}$ thus freeing the action of the group G_0 upon the fields from coordinate transformations. (We shall henceforth also drop the $\circ$ indicating the limitation of the range of summation for group indices to the one parameter subgroups in G_0 and always take this to be the case.) Thus Eq. (2.16) is to represent an arbitrary, space—time dependent group action and (2.1) an arbitrary coordinate change. Of course, we may choose the coordinate change to be generated by the group G as in (2.2) and the corresponding action upon the fields to be of the form (2.15), but we need not do so.

Now, in view of the transformation property (2.17) for $\partial_\mu \psi$, we are led to define a covariant derivative

$$\nabla_\mu \psi = \partial_\mu \psi + \omega_\mu^a \mathring{T}_a \psi, \tag{3.5}$$

which transforms as

$$\delta(\nabla_\mu \psi) = \chi^a \mathring{T}_a \nabla_\mu \psi - \nabla_\nu \psi \partial_\mu \xi^\nu \tag{3.6}$$

provided the gauge fields ω_μ^a transform as follows:

$$\delta \omega_\mu^a = -\partial_\mu \chi^a + \mathring{f}_{bc}^a \chi^b \mathring{\omega}_\mu^c - \omega_\nu^a \partial_\mu \xi^\nu. \tag{3.7}$$

Thus the $\{\omega_\mu^a\}_{a=1,\ldots,n-m}$ are components of coordinate covariant vectors and have a homogeneous part transforming under the adjoint representation of G_0 plus an additional inhomogeneous term characteristic of gauge fields. The next step is to proceed as in the linear case and introduce a further set of four covariant vector fields b_μ^i $(i = 0,1,2,3)$ together with their inverses h_i^μ:

$$h_i^\mu b_\nu^i = \delta_\nu^\mu, \quad h_i^\mu b_\mu^j = \delta_i^j. \tag{3.8}$$

We require these fields to satisfy the following transformation properties:

$$\delta h^\mu_i = \partial_\nu \xi^\mu h^\nu_i - \chi^j_i h^\mu_j, \quad \delta b^i_\mu = -\partial_\mu \xi^\nu b^i_\nu + \chi^i_j b^j_\mu. \tag{3.9}$$

The covariant derivative $\nabla_\mu \psi$ is then referred to the vectors h^μ_i as basis by defining

$$\psi_{;i} \equiv h^\mu_i \nabla_\mu \psi, \tag{3.10}$$

which satisfies the transformation property (3.4). If $d_\mu\psi$ is then replaced in $L(\psi, d_\mu\psi)$ by $\psi_{;i}$, the new Lagrangian will transform as in (3.2) even under nonrigid actions of the group and coordinate changes.

To obtain a Lagrangian which is a group invariant and a coordinate scalar density, we now define

$$\mathcal{L} \equiv bL, \tag{3.11}$$

where

$$b \equiv \det\{b^i_\mu\}. \tag{3.12}$$

It follows from (3.9) that b transforms as

$$\delta b = -\partial_\mu \xi^\mu b + \chi^i_i b, \tag{3.13}$$

and hence

$$\delta \mathcal{L} + \partial_\mu \xi^\mu \mathcal{L} = 0. \tag{3.14}$$

Finally, we should like to emphasize that the minimal coupling procedure outlined here is only applicable if the relation (3.1) is valid. In general, however, we need not necessarily start with a Lagrangian which satisfies (2.18) for rigid actions of the group. We may simply use the fields ψ and their covariant derivatives $\psi_{;i}$ to form invariant "densities" (under the action of G_0) satisfying (3.2) and then multiply by the quantity b as in (3.11) to obtain a group invariant and coordinate scalar density.

4. IDENTITIES AND CONSERVATION LAWS

We now consider a Lagrangian $\mathcal{L}(\psi, \partial_\mu\psi, h^\mu_i, \omega^a_\mu, x)$ depending on the matter and gauge fields, the first derivatives of the matter fields, and possibly the coordinates, and investigate the consequences following from the invariance condition (3.14). The variation $\delta\mathcal{L}$ may be written as

$$\delta\mathcal{L} = \frac{\partial\mathcal{L}}{\partial Q_A}\,\partial Q_A + \frac{\partial\mathcal{L}}{\partial Q_{A,\mu}}\,\delta Q_{A,\mu} + \frac{\partial\mathcal{L}}{\partial x^\mu}\,\xi^\mu, \tag{4.1}$$

where $\{Q_A\}$ denotes all the fields, matter and gauge, upon which $\mathcal{L}$ depends. Here we have

$$\delta Q_A = Q'_A(x') - Q_A(x), \tag{4.2}$$

$$\delta Q_{A,x} = Q'_{A,\mu'}(x') - Q_{A,\mu}(x). \tag{4.3}$$

Equation (4.1) is equivalent to

$$\partial_\mu\left[\left(\delta^\mu_\nu \mathcal{L} - \frac{\partial\mathcal{L}}{\partial\partial_\mu Q_A}\,\partial_\nu Q_A\right)\xi^\nu + \frac{\partial\mathcal{L}}{\partial\partial_\mu Q_A}\,\delta Q_A\right] + \mathcal{L}_{[Q_A]}\delta_0 Q_A = 0, \tag{4.4}$$

where, as before,

$$\mathcal{L}_{[Q_A]} = \frac{\partial\mathcal{L}}{\partial Q_A} - \partial_\mu \frac{\partial\mathcal{L}}{\partial\partial_\mu Q_A} \tag{4.5}$$

is the Eulerian derivative and

$$\delta_0 Q_A \equiv Q'_A(x) - Q_A(x) = \delta Q_A - \partial_\mu Q_A \xi^\mu \tag{4.6}$$

is the substantial variation of Q_A. Since the Lagrangian depends on derivatives only of the matter fields and the Eulerian derivatives with respect to these vanish by the variational principle, we have

$$\partial_\mu\left[\frac{\partial\mathcal{L}}{\partial\partial_\mu\psi}\,\delta\psi - \mathcal{T}^\mu_\nu \xi^\nu\right] + \mathcal{L}_{[h^\mu_i]}\delta_0 h^\mu_i + \mathcal{L}_{[\omega^a_\mu]}\delta_0\omega^a_\mu = 0. \tag{4.7}$$

Here $\mathcal{T}^\mu_\nu$ is the canonical energy—momentum tensor—density, for the Lagrangian $\mathcal{L}$. Now substituting the expressions (2.15), (2.17), (3.7), (3.8), and (4.6) into (4.7) and noting that the terms χ^a, $\partial_\mu\chi^a$, ξ^ν, $\partial_\mu\xi^\nu$ are all independent, we may equate each of their coefficients separately to zero, so obtaining a number of identities. From coordinate invariance (terms in ξ^ν, $\partial_\mu\xi^\nu$) we obtain the following two relations:

$$\mathcal{T}^\mu_\nu = t^\mu_\nu - \mathcal{S}^\mu_a \omega^a_\nu, \tag{4.8}$$

$$\partial_\mu \mathcal{T}^\mu_\nu + t^i_\mu h^\mu_{i,\nu} + \mathcal{S}^\mu_a \omega^a_{\mu,\nu} = 0, \tag{4.9}$$

where

$$t^i_\mu \equiv \frac{\partial\mathcal{L}}{\partial h^\mu_i} \tag{4.10}$$

$$t^\mu_\nu \equiv h^\mu_i t^i_\nu \tag{4.11}$$

and

$$\mathcal{S}^\mu_a \equiv \frac{\partial\mathcal{L}}{\partial\omega^a_\mu}. \tag{4.12}$$

We note that because of the transformation properties of $\mathcal{L}$ and ω^a_μ, the quantity $\mathcal{S}^\mu_a$ defined in (4.12) is a coordinate vector density transforming under the group action according to the co-adjoint representation of the group:

$$\delta\mathcal{S}^\mu_a = \partial_\nu\xi^\mu \mathcal{S}^\nu_a - \partial_\nu\xi^\nu \mathcal{S}^\mu_a - f^c_{ab}\chi^b \mathcal{S}^\mu_c. \tag{4.13}$$

The quantity t^μ_ν, on the other hand, is a coordinate tensor density and a group invariant:

$$\delta t^\mu_\nu = \partial_\sigma\xi^\mu t^\sigma_\nu - \partial_\nu\xi^\sigma t^\mu_\sigma - \partial_\sigma\xi^\sigma t^\mu_\nu. \tag{4.14}$$

Taking the divergence of $\mathcal{T}^\mu_\nu$ in (4.8) and subtracting from (4.9), we obtain the relation

$$\partial_\mu t^\mu_\nu - t^i_\mu h^\mu_{i,\nu} = D_\mu \mathcal{S}^\mu_a \omega^a_\nu + \mathcal{S}^\mu_a R^a_{\mu\nu}, \tag{4.15}$$

where

$$D_\mu \mathcal{S}^\mu_a \equiv \partial_\mu \mathcal{S}^\mu_a - f^c_{ab}\omega^b_\mu \mathcal{S}^\mu_c \tag{4.16}$$

and

$$R^a_{\mu\nu} \equiv \partial_\mu\omega^a_\nu - \partial_\nu\omega^a_\mu + f^a_{bc}\,\omega^b_\mu\omega^c_\nu. \tag{4.17}$$

In (4.17) we have introduced the "gauge curvature" $R^a_{\mu\nu}$ associated with the gauge fields ω^a_μ. It follows from the transformation property (3.7) for the gauge fields that $R^a_{\mu\nu}$ has a homogeneous linear transformation property corresponding to the adjoint representation of the group, while it is a covariant, antisymmetric tensor under coordinate transformation:

$$\delta R^a_{\mu\nu} = -\partial_\mu\xi^\alpha R^a_{\alpha\nu} - \partial_\nu\xi^\alpha R^a_{\mu\alpha} + f^a_{bc}\chi^b R^c_{\mu\nu}. \tag{4.18}$$

This result makes it possible to form manifestly invariant Lagrangians for the gauge fields alone. Relation (4.16), furthermore, may be understood as defining the gauge-covariant divergence of the current $\mathcal{S}^\mu_a$, in view of its transformation property (4.13). Since S_a^μ is a coordinate vector density, this is a coordinate covariant relation. As we shall see shortly, relation (4.15) may also be cast in a manifestly covariant form.

We now turn to the identities following from the group invariance of the Lagrangian [the terms in χ^a and $\partial_\mu \chi^a$ in relation (4.7)]. These relations give us firstly an identification of S_a^μ as the intrinsic current of matter associated with the one-parameter subgroup with generator $\mathring{T}_a$:

$$S_a^\mu = \frac{\partial L}{\partial \partial_\mu \psi} \mathring{T}_a \psi. \tag{4.19}$$

Using this relation, the remaining identities may be expressed in the following simple form:

$$D_\mu S_a^\mu = 0 \quad (\text{for } \mathring{T}_a \text{ not in the linear subalgebra}), \tag{4.20}$$

$$D_\mu S_j^{i\mu} = t_\mu^i h_j^\mu. \tag{4.21}$$

Here, we have separated the results for the linear subgroup of G_0 from the rest and defined

$$S_j^{i\mu} \equiv \frac{\partial L}{\partial \partial_\mu \psi} \sigma_j^i \psi = \frac{\partial L}{\partial \omega_{i\mu}^j}, \tag{4.22}$$

which is the current associated with the one-parameter linear subgroup of G_0 whose infinitesimal generator is σ_j^i for the given representation of ψ. Relation (4.20) thus tells us that the gauge covariant divergence of the currents associated with the nonlinear parts of the group G_0 vanish. Substituting the remaining, nonvanishing divergence given by (4.21), into (4.15), we obtain the following, important relation:

$$\partial_\mu t_\nu^\mu + \Gamma_{\mu\nu}^\alpha t_\alpha^\mu = R_{\mu\nu}^a S_a^\mu, \tag{4.23}$$

where

$$\Gamma_{\mu\nu}^\alpha \equiv -(h_{i,\nu}^\alpha - \omega_{i\nu}^j h_j^\alpha) b_\mu^i$$
$$= (b_{\mu,\nu}^i + \omega_{j\nu}^i b_\mu^j) h_i^\alpha. \tag{4.24}$$

Equation (4.23) has very direct significance for the underlying dynamics of gauge field theories, and also for the geometrical interpretation of the results. First, we note that, in view of the transformation properties of the gauge fields $\omega_{j\mu}^i, b_\mu^i$ [Eqs. (3.7), (3.9)], the quantity defined by Eq. (4.24) is a group invariant which transforms as follows under coordinate changes:

$$\delta\Gamma_{\mu\nu}^\alpha = -\partial_\mu \xi^\sigma \Gamma_{\sigma\nu}^\alpha - \partial_\nu \xi^\sigma \Gamma_{\mu\sigma}^\alpha + \partial_\sigma \xi^\alpha \Gamma_{\mu\nu}^\sigma - \partial_\mu \partial_\nu \xi^\alpha. \tag{4.25}$$

We recognize this as the transformation property of an affine connection. Furthermore, if we define the second order symmetric tensor

$$\mathcal{G}_{\mu\nu} \equiv b_\mu^i b_\nu^j \eta_{ij} \tag{4.26}$$

(where η_{ij} is the Minkowski metric), together with its inverse

$$\mathcal{G}^{\mu\nu} \equiv h_i^\nu h_j^\nu \eta^{ij}, \tag{4.27}$$

the group transformation property of this entity is given by

$$\delta \mathcal{G}_{\mu\nu} = (\chi_k^i b_\mu^k b_\nu^j + \chi_k^j b_\mu^i b_\nu^k) \eta_{ij}. \tag{4.28}$$

Now let us separate the infinitesimal parameters of the group into an antisymmetric part, trace, and traceless symmetric part:

$$\chi_j^i = \chi_{aj}^i + \chi_{sj}^i + \chi \delta_j^i, \tag{4.29}$$

where

$$\chi_{ak}^i \equiv \tfrac{1}{2}(\chi^{ij} - \chi^{ji}) \eta_{jk}, \tag{4.30}$$

$$\chi_{sk}^i \equiv \tfrac{1}{2}(\chi^{ij} + \chi^{ji}) \eta_{jk} - \chi \delta_k^i, \tag{4.31}$$

$$\chi \equiv \tfrac{1}{4}\chi_i^i \tag{4.32}$$

(raising and lowering of Roman indices done here with the Minkowski metric). Then (4.28) may be written as

$$\delta \mathcal{G}_{\mu\nu} = 2\chi \mathcal{G}_{\mu\nu} + 2\chi_{sij} b_\mu^i b_\nu^j. \tag{4.33}$$

Thus, the antisymmetric part of χ_j^i, corresponding to orthogonal (Lorentz) transformations, does not alter $\mathcal{G}_{\mu\nu}$ at all while the diagonal part, corresponding to scale transformation, multiplies it by a factor $(1+2\chi)$. If there is a traceless symmetric part in the infinitesimal linear transformations of G_0, the change in the quantity $\mathcal{G}_{\mu\nu}$ cannot be expressed in terms of its original value alone. Within the geometrical interpretation of the theory, we shall be led to regarding $\mathcal{G}_{\mu\nu}$ as a generalization of the metric of Riemannian geometry. With this identification, we may obtain an interpretation of the tensor t_ν^μ. Noting that if the Lagrangian L depends upon the fields h_i^μ (or b_μ^i) only through the tensor $\mathcal{G}_{\mu\nu}$, then we have the following relation:

$$t_\nu^\mu = \frac{1}{2} \frac{\partial L}{\partial \mathcal{G}_{\mu\sigma}} \mathcal{G}_{\nu\sigma}. \tag{4.34}$$

The right-hand side of (4.34) is, within Riemannian geometry, the symmetrical energy—momentum tensor. If we retain this interpretation in the more general case, the relation (4.23) has a clear physical interpretation. The left-hand side, when integrated over a space—time tube traced out by a particle of infinitesimal spatial extent gives the covariant derivative (with connection $\Gamma_{\mu\nu}^\alpha$) of the unit tangent vector along the path.[25] If the right-hand side of (4.23) vanished, this would give rise to the equation for geodesic motion. The nonvanishing term $R_{\mu\nu}^a S_a^\mu$ must hence be interpreted as the noninertial force density giving the deviation from geodesic paths. With suitable interpretation of the current and the gauge curvature $R_{\mu\nu}^a$, this may be viewed as a generalization of the Lorentz force of electrodynamics. A force of this sort has been shown to arise for matter with intrinsic spin within the Cartan—Sciama—Kibble theory, and also within the conventional general-relativistic approach to spinning matter.[26]

5. FIELD EQUATIONS IN A LAGRANGIAN MODEL

The next step in formulating a gauge field theory is to choose a Lagrangian for the gauge fields alone which has the necessary invariance property under gauge and coordinate transformations. The sum of matter Lagrangian with minimal coupling plus gauge fields Lagrangian may then be subjected to the Euler—Lagrange variational procedure to yield the field equations. To illustrate this method, let us consider gauge Lagrangians of the following form:

$$L_G = -(N/4)\, b R_{\mu\nu}^a R_{\sigma\tau}^b \,\mathcal{G}^{\mu\sigma} \mathcal{G}^{\nu\tau} g_{ab}. \tag{5.1}$$

Here g_{ab} is the group metric defined by

$$g_{ab} = f_{ad}^c f_{bc}^d, \tag{5.2}$$

and $N/4$ is a normalization constant. Note that this Lagrangian is an invariant only for groups whose linear parts consist at most of orthogonal (Lorentz) and scale transformations, in view of the transformation property (4.33) for the metric $g_{\mu\nu}$. The action integral which is required to be stationary for arbitrary variation of matter fields ψ, gauge fields ω^a_μ, and frame fields h^μ_i which vanish, together with their first derivatives, outside a finite space—time domain, is then

$$S=\int[\mathcal{L}+\mathcal{L}_G]d^4x. \tag{5.3}$$

The resulting field equations are:

$$ND_\mu[bR^{\mu\nu}_a]=\mathcal{S}^\nu_a, \tag{5.4}$$

$$Nb[R^a_{\nu\sigma}R^{\mu\sigma}_a-\tfrac{1}{4}\delta^\mu_\nu R^a_{\sigma\tau}R^{\sigma\tau}_a]=t^\mu_\nu, \tag{5.5}$$

$$\mathcal{L}_{[\psi]}=0. \tag{5.6}$$

The raising and lowering of space—time and group indices is accomplished with the respective coordinate and group metric. For fields ψ which remain unchanged under actions of the translation subgroup ($\mathring{T}_i=0$), the corresponding currents in (5.4) vanish, giving us the four vector equations

$$D_\mu\{g_{ia}bR^{a\mu\nu}\}=0, \tag{5.7}$$

where the subscript i refers to the one-parameter subgroup of translations in the x^i direction. From Eq. (5.4), we have the following conservation law:

$$\partial_\mu[\mathcal{S}^\mu_a+f^c_{ab}\omega^b_\nu R^{\nu\mu}_c]=0, \tag{5.8}$$

which leads us to making the identification of a current $\mathcal{J}^\mu_a$ carried by the gauge fields:

$$\mathcal{J}^\mu_a\equiv f^c_{ab}\omega^b_\nu R^{\nu\mu}_c. \tag{5.9}$$

Equation (5.8) is then a true conservation law for the total current, but is not manifestly group covariant, as compared with the covariant conservation law (4.20).

PART B

6. FIBRE BUNDLES

In Part B we will first summarize certain basic notions of differential geometry, introducing in particular the covariant derivative in terms of a connection form in a principal fibre bundle[27]; we then interpret from this geometrical viewpoint the gauge fields defined in Part A. Similar geometrical presentations of gauge field theories have been given by other authors.[28]

Consider a four-dimensional (smooth) manifold M. The set of all (smooth) vector fields on M is denoted by $\mathcal{X}(M)$. Consider also a principal bundle $P=P(M,G)$ over M with structure group G. The bundle P is a local direct product of M with G; we have the projection $\pi:P\to M$ and for each $q\in M$ the fibre $\pi^{-1}(q)$ is a copy of G attached to M at the point q. The group G thus acts on the fibres, and $R_g(f)\equiv f\cdot g$ denotes the right action of $g\in G$ on $f\in P$. For instance when P is the linear frame bundle $L(M)$, the fibre $\pi^{-1}(q)$ consists of all sets of basis vectors for the tangent space M_q at q and the structure group $G=\mathrm{GL}(4)$ which permutes these bases among themselves.

One is interested in physical fields on which the structure group G acts; for this one needs vector bundles associated with $P(M,G)$. Thus, consider a vector space V and a representation $\rho:G\to\mathrm{GL}(V)$. On the direct product $P\times V$ we call (f,v) and $(f',v')\in P\times V$ equivalent if

$$(f',v')=(fg,\rho(g^{-1})v)\ \text{ for some } g\in G, \tag{6.1}$$

and for this equivalence relation we denote by $[f,v]$ the equivalence class containing (f,v). Then the set $E=E(M,G,\rho)$ of such equivalence classes endowed with a differentiable manifold structure (naturally related to that on P) is called the vector bundle with fibre V associated with $P(M,G)$ via the representation ρ. A ρ-field over M is a cross section $\psi:M\to E$ (i.e., a mapping satisfying $\pi_E\circ\psi=\mathrm{id}_M$, where $\pi_E:E\to M$ is the projection map of the bundle E onto M). Such a ρ-field ψ can be identified with a V-valued function $\tilde\psi:P\to V$ which is ρ-invariant in the sense that

$$\tilde\psi(f\cdot g)=\rho(g^{-1})\tilde\psi(f)\ \text{ for all } f\in P,\ g\in G. \tag{6.2}$$

Conversely any such function gives rise to a ρ-field ψ. Explicitly $\tilde\psi$ is characterized by ψ through the following relationship:

$$\psi(q)=[f,\tilde\psi(f)]\ \text{ for all } q\in M,\ f\in\pi^{-1}(q). \tag{6.3}$$

Denote by $\mathcal{G}$ the Lie algebra of G. Then the action of G on P induces a homomorphism $A\to A^*$ from $\mathcal{G}$ into $\mathcal{X}(P)$; explicitly for $A\in\mathcal{G}$ and $f\in P$, $(A^*)_f$ is the tangent vector at $t=0$ to the curve $R_{\exp tA}(f)$ obtained by right action on f by the one-parameter subgroup $\exp tA$. We call $(A^*)_f$ a vertical vector at f: It is tangent to the fibre $\pi^{-1}(q)$ where $q=\pi(f)$. These vertical vectors form a $(\dim G)$-dimensional subspace of the tangent space P_f to $P(M,G)$ at the point f.

7. CONNECTIONS AND COVARIANT DIFFERENTIATION

A connection in $P(M,G)$ specifies in the tangent spaces P_f a smooth G-invariant distribution $\mathcal{H}$ of horizontal subspaces $\mathcal{H}_f$ which are complementary to the space of vertical vectors. Thus a tangent vector $\tilde X_f\in P_f$ can be written as

$$\tilde X_f=\bar X_f+\mathrm{vert}(\tilde X_f), \tag{7.1}$$

where $\bar X_f$ and $\mathrm{vert}(\tilde X_f)$ are its horizontal and vertical parts respectively. Now

$$\mathrm{vert}(\tilde X_f)=(A^*)_f \tag{7.2}$$

for a unique $A\in\mathcal{G}$; the $\mathcal{G}$-valued 1-form ω defined by

$$\omega(\tilde X_f)=A \tag{7.3}$$

has the properties

$$\omega(A^*)=A \tag{7.4a}$$

and

$$R^*_g\omega=(\mathrm{ad}g^{-1})\cdot\omega\ \text{ for } g\in G. \tag{7.4b}$$

[Here $R^*_g\omega$ is the pullback of ω by the mapping R_g.] This connection form ω completely characterizes the connection. Relative to a basis $\{t_a\}$ for $\mathcal{G}$ we express

$$\omega=\omega^a t_a, \tag{7.5}$$

where the ω^a are 1-forms on P ($a=1,2,\ldots,\dim G$).

Consider a given connection, characterized by its connection form ω. Consider also a tangent vector field

$X \in \mathcal{X}(M)$, i.e., a field X of directions on the manifold M. Now every such $X \in \mathcal{X}(M)$ has a unique horizontal lift $\bar{X} \in \mathcal{X}(P)$; i.e., there is a uniquely determined tangent vector field $\bar{X} \in \mathcal{X}(P)$ such that, for all $f \in P$,

$$\pi_*(\bar{X}_f) = X_{\pi(f)} \quad \text{and} \quad \bar{X}_f \in \mathcal{H}_f. \tag{7.6}$$

Now $\bar{X}$ is invariant by G; that is,

$$(R_g)_*(\bar{X}_f) = \bar{X}_{R_g(f)} \tag{7.7}$$

for all $g \in G$, $f \in P$. Consequently for the ρ-invariant V-valued function $\tilde{\psi}$ corresponding to a given ρ-field ψ, the new ρ-invariant V-valued function $\bar{X}\tilde{\psi}$ (which we also denote by $\nabla_X \tilde{\psi}$) defines a corresponding ρ-field $\nabla_X \psi$: $M \to E$ which we call the covariant derivative of ψ (along the flow lines of the vector field X). Considering all the various representations ρ_α of G and the associated vector bundles E_α, one has many types of fields and the operator ∇_X acts on these fields as a covariant derivative: covariant in the sense that if ψ_α is a ρ_α-field then $\nabla_X \psi_\alpha$ is a ρ_α-field also, and derivative in the sense that ∇_X obeys a Leibniz rule when applied to tensor products of fields.

For a representation $\rho: G \to \mathrm{GL}(V)$, consider a V-valued k-form φ on P for which

$$R_g^* \varphi = \rho(g^{-1})\, \varphi \tag{7.8}$$

for all $g \in G$. The exterior covariant derivative $D\varphi$ is the $(k+1)$-form defined by

$$D\varphi(\tilde{X}_0, \tilde{X}_1, \ldots, \tilde{X}_k) = d\varphi(\bar{X}_0, \bar{X}_1, \ldots, \bar{X}_k), \tag{7.9}$$

where $d\varphi$ is the usual exterior derivative of φ. It follows that $D\varphi$ projects naturally to define a $(k+1)$-form on M. We note that

$$\begin{aligned} D\varphi &= d\varphi + (\rho \circ \omega) \wedge \varphi \\ &= d\varphi + \omega^a \wedge (\rho(t_a)\, \varphi). \end{aligned} \tag{7.10}$$

In particular for the connection 1-form ω, the related $\mathcal{G}$-valued 2-form $\Omega = D\omega$ is the curvature form of the connection, and we have

$$\Omega(\tilde{X}, \tilde{Y}) = d\omega(\tilde{X}, \tilde{Y}) + \tfrac{1}{2}[\omega(\tilde{X}), \omega(\tilde{Y})], \tag{7.11}$$

where $\tilde{X}$ and $\tilde{Y} \in \mathcal{X}(P)$, and the symbol [,] denotes the Lie bracket operation in $\mathcal{G}$. This 2-form Ω defines a $\mathcal{G}$-valued 2-form R as a tensor field on M; relative to a coordinate system $\{x^\mu\}$ on U_A, we have

$$R = R_{\mu\nu}\, dx^\mu \wedge dx^\nu \quad \text{with} \quad R_{\mu\nu} = 2\Omega_A(\bar{\partial}_\mu, \bar{\partial}_\nu), \tag{7.12}$$

where the vector fields $\{\partial_\mu\}$ are the partial derivative fields on U_A (i.e., $\partial_\mu = \partial/\partial x^\mu$). This 2-form R is referred to as the $\mathcal{G}$-valued curvature tensor on M associated with ω.

Consider a local section σ_A of M; i.e., a mapping $\sigma_A: U_A \to P$ of some open set U_A in M into P such that $\pi \circ \sigma_A = \mathrm{id}_U$. For example, when P is the bundle $L(M)$ of linear frames, then σ_A is a smooth choice of frame for all points in U_A; that is, such a section σ_A corresponds to a local "moving frame" of Cartan. Returning to the general case, we use σ_A to refer the fields, connection form and curvature to the manifold. Thus we introduce the following objects defined on $U_A \subset M$:

$$\psi_A \equiv \tilde{\psi} \circ \sigma_A, \tag{7.13a}$$

$$\omega_A \equiv \sigma_A^* \omega, \tag{7.13b}$$

$$\Omega_A \equiv \sigma_A^* \Omega \equiv d\omega_A + \tfrac{1}{2}[\omega_A, \omega_A]. \tag{7.13c}$$

Similarly for the covariant derivative $\nabla_X \tilde{\psi}$ we consider $(\nabla_X \tilde{\psi})_A = (\nabla_X \tilde{\psi}) \circ \sigma_A$, and we have in fact

$$(\nabla_X \tilde{\psi})_A = X\psi_A + \rho(\omega_A(X))\, \psi_A. \tag{7.14}$$

For a second local section $\sigma_B: U_B \to P$ we have the relation on $U_A \cap U_B$ that

$$\sigma_B = \sigma_A \cdot g_{AB} \equiv R_{g_{AB}} \circ \sigma_A, \tag{7.15}$$

where $g_{AB}: U_A \cap U_B \to G$ is an appropriate function from the open set $U_A \cap U_B$ in M into the group G; g_{AB} is sometimes called a transition function for the bundle structure on $P(M, G)$. We have the following relations showing how the new choice of section affects the fields, connection form and curvature:

$$\psi_B = \rho(g_{AB}^{-1})\, \psi_A \tag{7.16a}$$

$$\omega_B = \mathrm{ad}(g_{AB}^{-1})\omega_A + g_{AB}^{-1}\, dg_{AB} \tag{7.16b}$$

$$\Omega_B = \mathrm{ad}(g_{AB}^{-1})\, \Omega_A, \tag{7.16c}$$

where $g_{AB}^{-1}\, dg_{AB}$ is a commonly used shorthand for the following. Consider a vector $X \in M_q$ for some $q \in U_A \cap U_B$; then $(g_{AB})_*(X)$ is a vector tangent to the group G at $g_{AB}(q)$ and applying the differential $(L_{g_{AB}(q)^{-1}})_*$ of left translation by the inverse of $g_{AB}(q) \in G$, we obtain the tangent vector $(L_{g_{AB}(q)^{-1}})_* \circ (g_{AB})_*(X)$ at the group identity; such a vector may be identified with an element of the Lie algebra $\mathcal{G}$ and that element of $\mathcal{G}$ we denote by $g_{AB}^{-1}\, dg_{AB}(X)$. Thus for each $X \in M_q$ with $q \in U_A \cap U_B$:

$$\omega_B(X) = \mathrm{ad}(g_{AB}(q))^{-1}\omega_A(X) + \{L_{g_{AB}(q)^{-1}}\}_* \circ \{g_{AB}(q)\}_*(X), \tag{7.17}$$

which is clearly more neatly written as:

$$\omega_B(X) = \mathrm{ad}(g_{AB}^{-1}) \circ \omega_A(X) + g_{AB}^{-1} \circ dg_{AB}(X). \tag{7.18}$$

We recognize (7.18) as the transformation property of gauge fields of the type introduced in Part A. In the next section we develop in more detail the geometrical interpretation of the gauge theory.

8. GEOMETRY OF GAUGE THEORIES

In Part A we introduced on a four-dimensional manifold M certain gauge fields related to invariance of a Lagrangian defined on M under transformations associated with a symmetry group G. We will show here that one can think of these gauge fields ω_μ^a as the components of a connection form ω on a principal bundle $P(M, G)$, and that the covariant derivative $\psi_{;i}$ of (3.4) agrees with that given by (7.14) from the bundle viewpoint.

A. Gauge fields and the curvature

The nonrigid gauge variations treated in Sec. 3 of Part A correspond to considering a 1-parameter family of sections $\sigma_{B(t)}$ for values of t say in an open interval $I_\epsilon = (-\epsilon, \epsilon)$ about 0 and with $\sigma_B(t)$ related to the given section σ_A by the transition functions $g_{AB(t)}(q) = \exp(-t\chi_q)$, where χ is $\mathcal{G}$-valued function on U_A. The function ψ_χ and the forms ω_χ and Ω_χ are defined on $U_A \times I_\epsilon$ by

$$\psi_\chi(q,t)=\psi_{B(t)}(q), \tag{8.1a}$$

$$\omega_\chi(q,t)=\omega_{B(t)}(q), \tag{8.1b}$$

$$\Omega_\chi(q,t)=\Omega_{B(t)}(q). \tag{8.1c}$$

By using Eqs. (7.16), one computes the substantial variation induced by χ:

$$\delta_\chi(\psi_\chi)=\rho(\chi)\psi_A, \tag{8.2a}$$

$$\delta_\chi(\omega_\chi(X))=[\chi,\omega_A(X)]-X\chi, \tag{8.2b}$$

$$\delta_\chi(\Omega_\chi)=[\chi,\Omega_A]. \tag{8.2c}$$

In terms of the basis $\{t_a\}$ for $\mathcal{G}$ we can express χ, ω_A, and Ω_A as follows:

$$\chi=\chi^a t_a, \tag{8.3a}$$

$$\omega_A=\omega_A^a t_a, \tag{8.3b}$$

$$\Omega_A=\Omega_A^a t_a. \tag{8.3c}$$

Moreover, the structure constants of $\mathcal{G}$ are defined by $[t_a,t_b]=f_{ab}^c t_c$ and we define the operators $T_a=\rho(t_a)$ for $a=1,2,\ldots,\dim\mathcal{G}$. Then Eqs. (8.2) read thus:

$$\delta_\chi\psi_\chi=\chi^a T_a\psi_A, \tag{8.4a}$$

$$\delta_\chi\omega_\chi^a(X)=f_{bc}^a\chi^b\omega_A^c(X)-X\chi^a, \tag{8.4b}$$

$$\delta_\chi\Omega_\chi=f_{bc}^a\chi^b\Omega_A^c. \tag{8.4c}$$

In this form we can compare with equations of Part A. Equation (8.4a) corresponds to (2.15). Relative to a given coordinate system $\{x^\mu\}$ on U_A, we define

$$\omega_{A\mu}^a\equiv\omega_A^a(\partial_\mu). \tag{8.5}$$

Then, for $X=\partial_\mu$, Eq. (8.4b) corresponds to the condition (3.7) imposed on the gauge fields, and (8.4c) corresponds to (4.18); the additional terms in (3.7) and (4.18) arise from including coordinate transformations as well as group transformations. [In connection with (4.18), note that the curvature 2-form $R=R^a t_a=R_{\mu\nu}^a t_a\,dx^\mu\wedge dx^\nu$ is determined relative to a coordinate system on U_A by the components $R_{\mu\nu}^a=2\Omega_A^a(\partial_\mu,\partial_\nu)$ which by (7.13c) agrees with the gauge curvature introduced in (4.17).]

The gauge fields h_i^μ [cf. (3.8)] define a set of four local vector fields $\{h_i=h_i^\mu\partial_\mu\}_{i=0,1,2,3}$, which are linearly independent and which may be equivalently thought of as a local section over U_A of a bundle of linear frames whose structure group is the linear part of the group G under consideration.

B. Covariant differentiation of matter fields

Let us now relate the covariant derivative of a ρ-field given in Eq. (7.14) with the covariant derivative of Part A. Considering U_A as a coordinate neighborhood with coordinate system $\{x^\mu\}$ and related tangent vectors $\{\partial_\mu\}$ we obtain from (7.14) with $X=\partial_\mu$:

$$\begin{aligned}(\nabla_{\partial_\mu}\tilde\psi)_A&=\partial_\mu(\psi_A)+\rho(\omega_A(\partial_\mu))\psi_A\\&=\partial_\mu(\psi_A)+\omega_{A\mu}^a(T_a\psi_A),\end{aligned} \tag{8.6}$$

which corresponds precisely to the covariant derivative of a field introduced from the gauge point of view in (3.4) of Part A.

C. Currents, conservation laws, and field equations

Before treating the currents and the field equations of Secs. 4 and 5 we introduce some further notions concerned with various ρ-fields on M.

The defining representation of GL(4) on $V=R^4$ gives the usual tensor representations $\rho(r,s)$ on $V^r\otimes(V^*)^s$, where V^r (resp. $(V^*)^s$) denotes the r-fold (resp. s-fold) tensor product of V (resp. its dual V^*). A $\rho(r,s)$-field on M is a tensor field of type (r,s); the antisymmetric tensors of type $(0,s)$ are the differential forms on M, for which the wedge product $\wedge$ is defined. We also have locally defined densities of weight k, i.e., local sections of a real line bundle over M constructed via the kth power of the determinant representation det of GL(4). The tensor product of such a local density with a tensor field, say of type (r,s), gives a tensor density of weight k; in other words, a local section of the bundle $E(M,\mathrm{GL}(4),\rho)$ where ρ is the representation $\det^k\otimes\rho(r,s)$ on $R^+\otimes V^r\otimes(V^*)^s$. Thus Eq. (4.26) gives the components of a tensor density $\mathfrak{g}$ of weight 2 corresponding to a local section of the bundle with $\rho=\det^2\otimes\rho(0,2)$.

Consider a Lorentz metric g on M. We extend g to a metric $\bar g$ on the Grassmann algebra of differential forms by defining for each $q\in M$:

$$\bar g_q(\alpha,\beta)=\det(g_q(\alpha_i,\beta_i)), \tag{8.7}$$

where α and β are decomposable k-forms at q given by $\alpha=\alpha_1\wedge\alpha_2\wedge\cdots\wedge\alpha_k$ and $\beta=\beta_1\wedge\beta_2\wedge\cdots\wedge\beta_k$ [and by extending (8.7) bilinearly for indecomposable forms]; moreover, k-forms and k'-forms are orthogonal if $k\neq k'$. In terms of $\bar g$ we define a metric duality between k-forms [tensors of type $(0,k)$] and antisymmetric tensor fields of type $(k,0)$; the dual ψ^+ of a k-form ψ is characterized by

$$\bar g(\varphi,\psi)=\psi^+(\varphi)\quad\text{for all } k\text{-forms } \varphi \tag{8.8}$$

[where the term $\psi^+(\varphi)$ denotes the complete contraction of ψ^+ with φ].

We assume now that M is an oriented manifold. Then M admits globally defined densities. Moreover, for any given metric g we have the related global volume form τ_g [a tensor of type $(0,4)$]. Relative to $\bar g$ and τ_g one can introduce[29] the Hodge star operator $*$ which maps a k-form φ to a $(4\text{-}k)$-form $*\varphi$ characterized by

$$\bar g(*\varphi,\lambda)\tau_g=\varphi\wedge\lambda\quad\text{for each } (4\text{-}k)\text{-form } \lambda. \tag{8.9}$$

Explicitly for a k-form expressed in terms of a moving coframe $\{e^i\}_{i=0,1,2,3}$ by

$$\varphi=\sum_R\varphi_{r_1\cdots r_k}e^{r_1}\wedge\cdots\wedge e^{r_k}, \tag{8.10}$$

where the sum is over all sets of ordered indices $R=\{r_1<r_2<\cdots<r_k\}$, we have

$$\begin{aligned}*\varphi=\sum_R&\varphi_{r_1\cdots r_k}\epsilon(r_1,\ldots,r_k,r_1',\ldots,r_{4-k}')\\&\times(-1)^{\#(R')}e^{r_1'}\wedge\cdots\wedge e^{r_{4-k}'},\end{aligned} \tag{8.11}$$

where $R'=\{r_1'<r_2'\cdots<r_{4-k}'\}$ is the ordered set of indices complementary to R (i.e., $R\cup R'=\{0,1,2,3\}$), $\epsilon(\alpha,\beta,\gamma,\delta)=\mathrm{sgn}\binom{0\ 1\ 2\ 3}{\alpha\ \beta\ \gamma\ \delta}$, and

$$\#(R') = \begin{cases} 0 & \text{if } R' \text{ does not contain the index } 0 \\ 1 & \text{if } R' \text{ does contain the index } 0 \end{cases}. \quad (8.12)$$

In Sec. 4 we showed how variations of the invariant Lagrangian give rise to certain currents related to the gauge fields. Consider the (locally defined) gauge fields $\{h_i\}$ and $\{\omega^a\}$. The fields $\{h_i\}$ yield a density of weight 1:

$$b = \det^{-1}(h_i^\mu), \quad (8.13)$$

and, as mentioned above, determine locally a $\det^2 \otimes \rho(0,2)$ tensor density $\mathcal{g}$ of weight 2; this local metric $\mathcal{g}$ defines locally a Hodge star and metric duality.

The vectors h_i and the 1-forms $b^i = b^i_\mu dx^\mu$ of (3.8) are related by the duality thus:

$$(b^i)^* = \eta^{ik} h_k, \quad (8.14)$$

and we have the local volume form $\beta = b^0 \wedge b^1 \wedge b^2 \wedge b^3$. The Lagrangian $\mathcal{L} = bL$ is a density of weight 1 and the integral of the associated 4-form $L\beta$ gives the action. To each gauge field ω^a, we have the coupled current 3-form

$$S_a = \frac{\delta(L\beta)}{\delta\omega^a}, \quad (8.15)$$

which corresponds to a current vector density $({}^*S_a)^*$ via the Hodge stars and metric duals. More explicitly if

$$S_a = (S_a)_{\alpha\beta\gamma} dx^\alpha \wedge dx^\beta \wedge dx^\gamma \quad (8.16)$$

expresses the 3-form relative to coordinates then

$$({}^*S_a)^* = (S_a)_{\alpha\beta\gamma} \epsilon^{\alpha\beta\gamma\mu} \partial_\mu, \quad (8.17)$$

whence we see that the currents S_a^μ of (4.12) are the components of the vector density $({}^*S_a)^*$. For the gauge fields $\{h_i\}$ a variation δh_i induces a change $\delta(L\beta)$ in the action integrand $L\beta$ given by

$$\delta(L\beta) = -{}^*(t^i(\delta h_i)), \quad (8.18)$$

where t^i is a 1-form whose coordinate components are given by (4.10).

To treat an invariant Lagrangian for the gauge fields as in Sec. 5, we suppose now that the Lie algebra $\mathcal{G}$ is semisimple. Then the Killing–Cartan form g of (5.2) is a nondegenerate quadratic form on $\mathcal{G}$, and defines an isomorphism $\mathcal{G} \to \mathcal{G}^*$ of $\mathcal{G}$ with its dual space $\mathcal{G}^*$. Moreover, if the group G has conformal linear part relative to the Minkowski metric on R^4, we introduce as in (5.1) an invariant Lagrangian density $\mathcal{L}_G = bL_G$, where the associated action integrand $L_G \cdot \beta$ is the 4-form

$$L_G \cdot \beta = (N/2)\{R^* \wedge ({}^*R)\}, \quad (8.19)$$

where R^* is the $\mathcal{G}^*$-valued form dual to the $\mathcal{G}$-valued curvature form R and ${}^*R$ is the Hodge star of R (with respect to the chosen orientation on M and the local metric $\mathcal{g}$). Using

$$R = R^a t_a = (d\omega_A^a + \tfrac{1}{2}[\omega_A \wedge \omega_A]^a)\, t_a, \quad (8.20)$$

we obtain in invariant form the field equation (5.4) obtained by variation of the gauge fields ω^a:

$$S = (2N)\{D({}^*R^*)\}. \quad (8.21)$$

Similarly the field equation (5.5) obtained by variation of the vierbein field, h_i, can be expressed as:

$$t^i(X) = \mathcal{L}_G \cdot b^i(X) + (Nb/4) \cdot \bar{g}(X \lrcorner R^*, (b^i)^* \lrcorner R), \quad (8.22)$$

where for fixed $X \in \mathcal{X}(M)$ we define the interior product[30] $X \lrcorner R^*$ as the 1-form satisfying:

$$(X \lrcorner R^*)(Y) = 2 \cdot R^*(X, Y) \quad \text{for all } Y \in \mathcal{X}(M). \quad (8.23)$$

Also, in this context the divergence relation (4.20) may be written as

$$D(S_a) = 0 \quad (8.24)$$

for t_a in the nonlinear part of the algebra $\mathcal{G}$, and for the linear part (4.21) may be written as

$$D(S_j^i) = {}^*(t^i(h_j)). \quad (8.25)$$

Finally, defining a tensor density t of type (1,1) and weight 1 by $t = h_i \circ t^i$ and by considering, for each $X \in \mathcal{X}(M)$, $\nabla(t)X$ as a linear transformation mapping $Z \in \mathcal{X}(M)$ into $\nabla_Z(t) X \in \mathcal{X}(M)$, we can express the relation (4.23) (connected with geodesic deviation) as follows:

$$\mathrm{tr}[\nabla(t) X] = R^*(({}^*S)^*, X). \quad (8.26)$$

ACKNOWLEDGMENTS

The authors would like to acknowledge helpful discussions with M. Perroud regarding certain aspects of this work. One of the authors (J.P.H.) would like to thank the Physics Department of Carleton University for their hospitality during the course of this work.

*J.P.H. and R.B.P. acknowledge partial support from the National Research Council of Canada under operating grants A1574 and A9248, respectively.

†Present address (J.P.H.); Centre de Recherches Mathématiques, Université de Montréal, Montréal, Québec, Canada.

[1]H. Weyl, Sitzungsber. Preuss. Akad. Wiss., 464 (1918); Ann. Phys. (Leipzig) **59**, 101 (1919); *Raum, Zeit, Materie* (Springer, Berlin, 1921) [*Space, Time, Matter* (Dover, New York, 1961)]; Z. Phys. **56**, 330 (1929); Phys. Rev. **77**, 699 (1950)
[2]C.N. Yang and R.L. Mills, Phys. Rev. **96**, 191 (1954).
[3]R. Utiyama, Phys. Rev. **101**, 1597 (1956).
[4]D.W. Sciama, *Festschrift for Infeld* (Pergamon, New York, 1962).
[5]E. Cartan, C.R. Acad. Sci. (Paris) **174**, 593 (1922); Ann. Ecole Norm. **40**, 325 (1923); **41**, 1 (1925).
[6]A. Trautman, Bull. Akad. Polon. Sci. **20**, 184, 503, 895 (1972); *ibid.* **21**, 345 (1973); Rep. Math. Phys. **1**, 29 (1970).
[7]F.W. Hehl, General Relativity Gravitation **4**, 333 (1973); **5**, 491 (1974); F.W. Heyl and P. von der Heyde, Ann. Inst. Henri Poincaré **19**, 179 (1973).
[8]A. Trautman, see Ref. 6.
[9]T.W.B. Kibble, J. Math. Phys. **2**, 212 (1961).
[10]S. Kobayashi, *Transformation Groups in Differential Geometry* (Springer-Verlag, Berlin, 1972).
[11]D.W. Sciama, J. Math. Phys. **2**, 472 (1961). Also see Ref. 4.
[12]A. Bregman, Progr. Theor. Phys. **49**, 667 (1973); J.M. Charap and W. Tait, Proc. R. Soc. Lond. A **349**, 249 (1974); P.G.O. Freund, Ann. Phys. (N.Y.) **84**, 440 (1974). See also Weyl's 1919 paper of Ref. 1, and Ref. 4.
[13]D.W. Sciama, see Ref. 11.
[14]C.J. Isham, A. Salam, and J. Strathdee, Lett. Nuovo

Cimento A **5**, 969 (1972); Phys. Rev. D **8**, 2600 (1973).
[15]P.A.M. Dirac, Proc. R. Soc. Lond. A **333**, 403 (1973).
[16]S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967).
[17]A. Salam, in *Elementary Particle Theory Nobel Symposium* 1968, edited by Nils Svartholm (Wiley, New York, 1968).
[18]D.W. Sciama, see Ref. 4.
[19]S. Weinberg, see Ref. 16.
[20]G. Mack and A. Salam, Ann. Phys. (N.Y.) **53**, 174 (1969).
[21]G. Mack and A. Salam, see Ref. 20.
[22]C.J. Isham, A. Salam, and J. Strathdee, see Ref. 14.
[23]T.W.B. Kibble, see Ref. 9.
[24]Van der Merwe, Nuovo Cimento Lett. **12**, 214 (1975); D.V. Volkov, Sov. J. Particles & Nuclei **4**, 1 (1973).
[25]A. Schild, in *Vol.* 8 *of Lectures in Applied Mathematics*, edited by J. Ehlers (Am. Math. Soc., Providence, R.I., 1967).
[26]M. Mathisson, Acta. Phys. Polon. **6**, 163 (1937); A. Papapetrou, Proc. R. Soc. Lond. A **209**, 249 (1951); F.A.E. Pirani, talk at Brandeis Summer Institute (1964); see also Ref. 25.
[27]The results stated in Part B are treated in full detail in the following books: S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963); S. Sternberg, *Lectures on Differential Geometry* (Prentice-Hall, Englewood Cliffs, New Jersey, 1964). Also see S. Kobayashi, Ref. 10.
[28]Y.M. Cho, J. Math. Phys. **16**, 2029 (1975); R. Herman, *Geometry, Physics, and Systems* (Marcel Dekker, New York, 1973); also see A. Trautman, Ref. 6.
[29]H. Flanders, *Differential Forms* (Academic, New York, 1963); also see S. Sternberg, Ref. 27.
[30]See Chapter 1 of S. Kobayashi and K. Nomizu, Ref. 27.

Diffraction by a half-plane perpendicular to the distinguished axis of a gyrotropic medium

R. A. Hurd and S. Przeździecki*

Division of Electrical Engineering, National Research Council of Canada, Ottawa, Ontario, K1A 0R8, Canada
(Received 1 April 1976)

The Wiener–Hopf–Hilbert method is used to obtain an exact solution to the problem of diffraction by a perfectly conducting half-plane in a gyrotropic medium, when the distinguished axis of the medium is perpendicular to the half-plane. The incident field is a plane wave whose direction of propagation is perpendicular to the edge of the half-plane. The problem has not previously been solved exactly. The answer is given in terms of Fourier transforms of the field components; these turn out to be simple algebraic functions. But the field quantities themselves are not, in general, expressible in terms of known functions. A few special cases are investigated and possible generalizations of the problem are mentioned.

1. INTRODUCTION

In this paper we give an exact, closed-form solution to a previously unsolved diffraction problem—that of a plane wave falling on a perfectly conducting half-plane embedded in an unbounded gyrotropic medium, whose distinguished axis is perpendicular to the half-plane. The direction of incidence is assumed to be perpendicular to the edge of the half-plane. The problem is a two-mode one; that is, both ordinary and extraordinary waves can propagate, and coupling between them occurs at the edge. This "diffraction" coupling manifests itself as a pair of simultaneous Wiener–Hopf equations, whose unknowns are closely related to the Fourier transforms of the field quantities. The difficulty of treating such equations is well-known, and this particular set had previously been thought insoluble.[1,2] Recently, however, a new method of treatment was devised,[3] in which Wiener–Hopf problems are converted to Hilbert problems. The latter always seem simpler than the former, and are exactly solvable in the present case. The solution, moreover, is elementary, in that the transforms comprise just algebraic functions.

Before proceeding, it is worthwhile to give a short survey of existing exact solutions for half-planes in anisotropic media. When the distinguished axis of a gyrotropic medium is parallel to the edge of the half-plane, there are solutions by Seshadri and Rajagopal,[4] Jull,[5] and De Santis.[6] This seems to be the only gyrotropic problem yet solved. For a uniaxial medium, Felsen[7] and Rulf[8] obtained solutions when the distinguished axis was perpendicular to the edge. Przeździecki[9] took the axis perpendicular to the plate, but allowed both magnetic and electric anisotropy and general skew incidence. When the axis is in the plane of the plate but otherwise arbitrarily directed, there is a solution by Williams.[10] Rosenbaum[11] considered the case of the axis arbitrarily directed in a plane perpendicular to the plate. Finally, a catalog of problems solvable by simple transformations of the Wiener–Hopf equations was prepared by Hurd.[2]

2. PROPERTIES OF THE GYROTROPIC MEDIUM

This section describes the gyrotropic medium, examines the propagation of plane waves in it, and lists relevant analytic properties of the propagation constants and other quantities.

A. Tensor permittivity

We introduce a rectangular coordinate system $\{x, y, z\}$ and suppose that the distinguished axis of the (lossless) gyrotropic medium lies in the z direction. The permittivity is then given by

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon & -i\epsilon_g & 0 \\ i\epsilon_g & \epsilon & 0 \\ 0 & 0 & \epsilon_a \end{pmatrix}, \tag{2.1}$$

with ϵ, ϵ_g, and ϵ_a real.

It is known that wave phenomena in such a medium depend critically on the relative sizes and on the algebraic signs of the elements of $\boldsymbol{\epsilon}$.[12,13] In this paper we suppose that

$$\epsilon \geqslant \epsilon_a > 0, \tag{2.2a}$$

$$\epsilon - \epsilon_a > |\epsilon_g|. \tag{2.2b}$$

These assumptions[14] ensure that propagation is of a fairly normal type, that is, no backward wave can occur; they also make the relative positions of some poles and branch points definite. It is not anticipated that relaxing (2.2) will invalidate the method of analysis.

B. Plane wave propagation

In the ensuing analysis we shall use the method of plane wave spectra. Accordingly, we first study the propagation of plane waves in the medium. Consider a plane wave whose components are

$$\mathbf{E} = \mathbf{E}_0 \exp[i(\alpha x + \gamma z) - i\omega t], \tag{2.3a}$$

$$\mathbf{H} = \mathbf{H}_0 \exp[i(\alpha x + \gamma z) - i\omega t]. \tag{2.3b}$$

With the time dependence $\exp(-i\omega t)$ suppressed henceforth, Maxwell's curl equations

$$\nabla \times \mathbf{E} = i\omega\mu_0 \mathbf{H}, \tag{2.4a}$$

$$\nabla \times \mathbf{H} = -i\omega\boldsymbol{\epsilon}\mathbf{E}, \tag{2.4b}$$

become

$$\mathbf{K}\mathbf{E}_0 = \omega\mu_0\mathbf{H}_0, \tag{2.5a}$$

$$\mathbf{K}\mathbf{H}_0 = -\omega\boldsymbol{\epsilon}\mathbf{E}_0, \tag{2.5b}$$

where

$$\mathbf{K}=\begin{pmatrix}0 & -\gamma & 0\\ \gamma & 0 & -\alpha\\ 0 & \alpha & 0\end{pmatrix}.$$

We can eliminate $\mathbf{H}_0$ from (2.5) and obtain

$$(\mathbf{KK}+\omega^2\mu_0\boldsymbol{\epsilon})\mathbf{E}_0=0. \tag{2.6}$$

Written out in full, (2.6) is

$$\begin{pmatrix}k^2-\gamma^2 & -ik_g^2 & \alpha\gamma\\ ik_g^2 & k^2-\alpha^2-\gamma^2 & 0\\ \alpha\gamma & 0 & k_a^2-\alpha^2\end{pmatrix}\begin{pmatrix}E_{0x}\\ E_{0y}\\ E_{0z}\end{pmatrix}=0, \tag{2.7}$$

where $k^2=\omega^2\mu_0\epsilon$, $k_g^2=\omega^2\mu_0\epsilon_g$, and $k_a^2=\omega^2\mu_0\epsilon_a$. For nonzero solutions, the determinant of (2.7) must vanish. That is,

$$k_a^2\gamma^4+[\alpha^2(k^2+k_a^2)-2k^2k_a^2]\gamma^2 + (k_a^2-\alpha^2)[k^4-k_g^4-k^2\alpha^2]=0. \tag{2.8}$$

Equation (2.8) determines four values of γ as functions of α. We denote them by $\pm\gamma_1$ and $\pm\gamma_2$. By virtue of the branches to be chosen, γ_1 and γ_2 give waves which propagate or are attenuated in the positive z direction. These are the extraordinary and ordinary waves mentioned before. The solution of a general diffraction problem must normally contain superpositions of waves of both types.

Some relations involving γ_1 and γ_2 are

$$\gamma_j=[\tfrac{1}{2}D(\alpha)+\tfrac{1}{2}(-)^j\Delta(\alpha)]^{1/2},\quad j=1,2, \tag{2.9}$$

with

$$D(\alpha)=(1+k^2k_a^{-2})[2k^2k_a^2(k^2+k_a^2)^{-1}-\alpha^2], \tag{2.10a}$$

$$\Delta(\alpha)=(k^2k_a^{-2}-1)[(\alpha_1^2-\alpha^2)(\alpha_2^2-\alpha^2)]^{1/2}. \tag{2.10b}$$

The constants α_1 and α_2 in (2.10b) are given by

$$\alpha_j^2=2k_g^2k_a^2(k^2-k_a^2)^{-2}\{k_g^2-i(-)^j[(k^2-k_a^2)^2-k_g^4]^{1/2}\},\quad j=1,2. \tag{2.11}$$

The following relations are easily proved:

$$\gamma_1^2+\gamma_2^2=D(\alpha), \tag{2.12a}$$

$$\gamma_2^2-\gamma_1^2=\Delta(\alpha), \tag{2.12b}$$

$$\gamma_1^2\gamma_2^2=k^2k_2^{-2}(k_1^2-\alpha^2)(k_2^2-\alpha^2), \tag{2.12c}$$

where

$$k_1^2=k^2-k_g^4k^{-2}, \tag{2.13a}$$

$$k_2^2=k_a^2. \tag{2.13b}$$

C. Analytic properties

In this section we set down various properties of the quantities defined in Sec. 2, Part B. These results will be used in succeeding sections, and are derived or justified in Appendix A.

(a) k_1 and k_2 are real and satisfy $k_1>k_2>0$. For a slightly lossy medium they are displaced into the first quadrant.

(b) α_j lies in the jth quadrant of the complex α plane and approaches the real axis only when $k_g\to 0$. Also, it is *assumed* that $\mathrm{Re}(\alpha_1)<k_2$.[15]

(c) γ_1 has branch points at $\pm k_1$, $\pm\alpha_1$, and $\pm\alpha_2$, while γ_2 has branch points at $\pm k_2$, $\pm\alpha_1$, and $\pm\alpha_2$. The positions of these points and the associated branch cuts are shown in Fig. 1.

(d) γ_j behaves as $(k_j^2-\alpha^2)^{1/2}F_j(\alpha)$, where $F_j(\alpha)$ is analytic on and near the branch cut contours from $\pm k_j$. In particular, γ_j changes sign as its branch cut is crossed. γ_j has a positive imaginary part if medium losses are present, when α is real.

(e) $\gamma_1+\gamma_2$ and $\gamma_1\gamma_2$ do not have branch points at $\pm\alpha_1$ and $\pm\alpha_2$ while $\gamma_1+\gamma_2$ has no zeros in the finite complex α plane.

(f) $k^2-\alpha^2-\gamma_1^2$ has simple zeros at $\alpha=\pm k_2$; $k^2-\alpha^2-\gamma_2^2$ has no zeros.

3. FORMULATION OF THE WIENER-HOPF PROBLEM

A perfectly conducting half-plane occupies the region $x\geq 0$, $z=0$, and a plane wave is incident upon it. The E_z component of this wave is given by

$$E_z^{(i)}=A_0\exp[i(\alpha_0x+\gamma_0z)], \tag{3.1}$$

where γ_0 is of either γ_1 or γ_2 type. The surrounding medium is gyrotropic, with characteristics given by (2.1) and (2.2). The problem is to find a scattered electromagnetic field **E**, **H** which

(i) obeys Maxwell's equations;

(ii) satisfies $E_x+E_x^{(i)}=0$, $E_y+E_y^{(i)}=0$ for $x\geq 0$, $z=0$;

(iii) satisfies an edge condition of the form $E_x=O(r^{-\nu_1})$ and $E_y+E_y^{(i)}=O(r^{\nu_2})$ as r, the distance from the edge, tends to zero. Here ν_1 and ν_2 satisfy $0<\nu_{1,2}<1$;

(iv) decays as $\exp(-ar)$, $a>0$, as $r\to\infty$ when the medium is lossy.

It is a consequence of the symmetry of the problem that E_z, H_x, H_y are odd functions of z, whilst H_z, E_x, E_y are even. It suffices, then, to consider only the region $z\geq 0$. To complete the boundary conditions for this region we add:

(v) $H_x=H_y=0$ for $x\leq 0$, $z=0$,

which follow from the continuity and asymmetry of H_x and H_y across $z=0$ for $x\leq 0$.

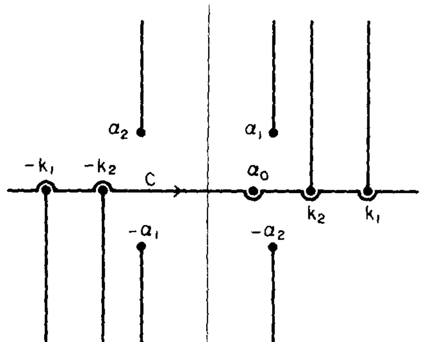


FIG. 1. Positions of the branch points $\pm k_1$, $\pm k_2$, $\pm\alpha_1$, $\pm\alpha_2$ and the pole α_0 in the complex α plane. Also shown are the integration contour C and the branch cuts.

Let us assume the following plane wave representation for E_z:

$$E_z = \int_C [A(\alpha)\exp(i\gamma_1 z) + B(\alpha)\exp(i\gamma_2 z)]\exp(i\alpha x)\,d\alpha, \tag{3.2}$$

where $A(\alpha)$ and $B(\alpha)$ are undetermined functions. The contour C follows the real axis except for indentations above $-k_1$ and $-k_2$ and below α_0, k_1, and k_2. See Fig. 1.

Using (2.7) and setting

$$f(\alpha,\gamma) = (\alpha^2 - k_2^2)(\alpha\gamma)^{-1}, \tag{3.3a}$$

$$g(\alpha,\gamma) = -ik_g^2 f(\alpha,\gamma)(k^2 - \alpha^2 - \gamma^2)^{-1}, \tag{3.3b}$$

we derive

$$E_x = \int_C [f(\alpha,\gamma_1)A(\alpha)\exp(i\gamma_1 z) + f(\alpha,\gamma_2)B(\alpha)\exp(i\gamma_2 z)]\exp(i\alpha x)\,d\alpha, \tag{3.4a}$$

$$E_y = \int_C [g(\alpha,\gamma_1)A(\alpha)\exp(i\gamma_1 z) + g(\alpha,\gamma_2)B(\alpha)\exp(i\gamma_2 z)]\exp(i\alpha x)\,d\alpha, \tag{3.4b}$$

and from (2.5a),

$$H_x = \frac{-1}{\omega\mu_0}\int_C [\gamma_1 g(\alpha,\gamma_1)A(\alpha)\exp(i\gamma_1 z) + \gamma_2 g(\alpha,\gamma_2)B(\alpha)\exp(i\gamma_2 z)]\exp(i\alpha x)\,d\alpha, \tag{3.5a}$$

$$H_y = \frac{1}{\omega\mu_0}\int_C \{[\gamma_1 f(\alpha,\gamma_1) - \alpha]A(\alpha)\exp(i\gamma_1 z) + [\gamma_2 f(\alpha,\gamma_2) - \alpha]B(\alpha)\exp(i\gamma_2 z)\}\exp(i\alpha x)\,d\alpha, \tag{3.5b}$$

$$H_z = \frac{1}{\omega\mu_0}\int_C [\alpha g(\alpha,\gamma_1)A(\alpha)\exp(i\gamma_1 z) + \alpha g(\alpha,\gamma_2)B(\alpha)\exp(i\gamma_2 z)]\exp(i\alpha x)\,d\alpha. \tag{3.5c}$$

Clearly, condition (i) is satisfied if the integrals converge; this will be verified *a posteriori*. Further, from property (d), γ_1 and γ_2 always have at least a limiting positive imaginary part, so that (iv) is also satisfied.

We denote the region of the complex α plane lying above the contour C by u, and the part below by l. To satisfy condition (ii) it is sufficient to take

$$\begin{pmatrix} f(\alpha,\gamma_1) & f(\alpha,\gamma_2) \\ g(\alpha,\gamma_1) & g(\alpha,\gamma_2)\end{pmatrix}\begin{pmatrix} A(\alpha) \\ B(\alpha)\end{pmatrix} = \begin{pmatrix} U_1(\alpha) \\ U_2(\alpha)\end{pmatrix} - \frac{A_0}{2\pi i(\alpha-\alpha_0)}\begin{pmatrix} f(\alpha_0,\gamma_0) \\ g(\alpha_0,\gamma_0)\end{pmatrix} = \begin{pmatrix} V_1(\alpha) \\ V_2(\alpha)\end{pmatrix}, \text{ say,} \tag{3.6}$$

where $U_1(\alpha)$ and $U_2(\alpha)$ are analytic functions of α in $u + C$. In like manner, condition (v) can be satisfied if

$$\begin{pmatrix} -\gamma_1 g(\alpha,\gamma_1) & -\gamma_2 g(\alpha,\gamma_2) \\ \gamma_1 f(\alpha,\gamma_1) - \alpha & \gamma_2 f(\alpha,\gamma_2) - \alpha\end{pmatrix}\begin{pmatrix} A(\alpha) \\ B(\alpha)\end{pmatrix} = \begin{pmatrix} L_1(\alpha) \\ L_2(\alpha)\end{pmatrix} \tag{3.7}$$

with $L_1(\alpha)$ and $L_2(\alpha)$ analytic in $l + C$.

A pair of simultaneous Wiener–Hopf equations may now be obtained by eliminating $A(\alpha)$ and $B(\alpha)$ from (3.6) and (3.7),

$$\frac{2}{\gamma_1+\gamma_2}\begin{pmatrix} -ik_g^2 & \alpha^2 - k^2 - \gamma_1\gamma_2 \\ k^2 + \gamma_1\gamma_2 k_2^2(k_2^2 - \alpha^2)^{-1} & ik_g^2\end{pmatrix}\begin{pmatrix} V_1(\alpha) \\ V_2(\alpha)\end{pmatrix} = \begin{pmatrix} L_1(\alpha) \\ L_2(\alpha)\end{pmatrix}, \tag{3.8}$$

or in compact form

$$\mathbf{G}(\alpha)\mathbf{V}(\alpha) = \mathbf{L}(\alpha), \quad \text{on } C. \tag{3.9}$$

Equation (3.9) is to be solved subject to the conditions

$$V_1(\alpha) = O(\alpha^{\nu_1 - 1}), \tag{3.10a}$$

$$V_2(\alpha) = O(\alpha^{-\nu_2 - 1}), \tag{3.10b}$$

as $|\alpha| \to \infty$ in $u + C$. This is a consequence of condition (iii).

It does not seem possible to solve (3.9) using ordinary Wiener–Hopf techniques.

4. THE WIENER-HOPF-HILBERT METHOD

The Wiener–Hopf–Hilbert (or WHH) method was recently introduced as a means of simplifying (and sometimes solving) equations such as (3.9). In Ref. 3, the method is treated rather cursorily. In the present work we try to establish it more rigorously; also the method of attack is altered somewhat. We begin with a short introduction.

A. Brief outline

Consider the homogeneous version of (3.9), which we write as $\mathbf{G}\Psi = \Phi$, with Ψ and Φ analytic in $u + C$ and $l + C$, respectively. (Occasionally we drop the notational dependence on α of various quantities. No confusion should arise.) This equation is first converted to a system of Hilbert problems $\Psi_+ = \mathbf{H}\Psi_-$ on the branch cuts Γ_1 and Γ_2 in l (Fig. 2). These Hilbert problems can be solved exactly and yield an infinite set of functions $\{\Psi^{(j)}\}$ which are analytic in $u + C$. We calculate the function $\mathbf{G}\Psi^{(j)}$ and require it to be analytic in $l + C$. Only a subset of $\{\Psi^{(j)}\}$ will have this property. This subset is further reduced by imposing the conditions at infinity (3.10). Finally the remaining $\{\Psi^{(j)}\}$ are assembled in a standard way to solve the *inhomogeneous* equation (3.9) and hence, the entire problem.

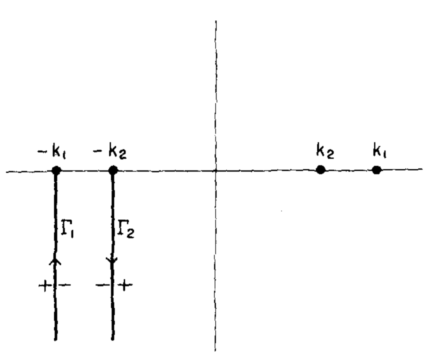


FIG. 2. The branch cuts Γ_1 and Γ_2 in the lower half α plane. The (+) and (-) signs identify the sides of the contours and the arrows identify directions of integration.

B. Derivation of the Hilbert equations

We write the homogeneous version of (3.9) as

$$\mathbf{G}(\alpha)\mathbf{\Psi}(\alpha)=\mathbf{\Phi}(\alpha) \quad \text{on } C, \tag{4.1}$$

with $\mathbf{\Psi}(\alpha)$ analytic in $u+C$ and $\mathbf{\Phi}(\alpha)$ analytic in $l+C$. As indicated, our aim is to convert (4.1) to a set of Hilbert equations on the contours Γ_1 and Γ_2. To do this, the functions of (4.1) must be capable of analytic extension into $l-\Gamma_1-\Gamma_2$. This is evidently true of $\mathbf{\Phi}(\alpha)$, since $\mathbf{\Phi}(\alpha)$ is analytic in l; it is equally true of $\mathbf{G}(\alpha)$ since its only singularities in l are the branch points $-k_1$ and $-k_2$. (Note that by Appendix A, Part 5, $\gamma_1+\gamma_2$ and $\gamma_1\gamma_2$ do not have branch points at $\pm\alpha_1$ and $\pm\alpha_2$, and $\gamma_1+\gamma_2$ has no zeros anywhere.) It is not quite obvious that $\mathbf{\Psi}(\alpha)$ can also be continued, but this can be seen as follows. We calculate $\mathbf{G}^{-1}(\alpha)$,

$$\mathbf{G}^{-1}(\alpha)=\frac{k_2^2-\alpha^2}{2k_2^2\gamma_1\gamma_2(\gamma_1+\gamma_2)} \times\begin{pmatrix} -ik_g^2 & k^2-\alpha^2+\gamma_1\gamma_2 \\ -k^2+k_2^2\gamma_1\gamma_2(\alpha^2-k_2^2)^{-1} & -ik_g^2 \end{pmatrix}. \tag{4.2}$$

Now the only singularities of $\mathbf{G}^{-1}(\alpha)$ in l are branch points at $-k_1$ and $-k_2$. From (4.1) we obtain

$$\mathbf{\Psi}(\alpha)=\mathbf{G}^{-1}(\alpha)\mathbf{L}(\alpha) \quad \text{on } C. \tag{4.3}$$

The right-hand side of (4.3) is analytic in $l-\Gamma_1-\Gamma_2$, hence $\mathbf{\Psi}(\alpha)$ can be analytically extended into this region and Eq. (4.1) must hold there.

Let us denote the two sides of Γ_j by (+) and (−), and assign senses of direction, as shown in Fig. 2. Let the limiting values of the functions on the (+) and (−) sides be identified by (+) and (−) subscripts. Then, from Appendix A, Part 4,

$$\gamma_{1_-}=-\gamma_{1_+} \quad \text{on } \Gamma_1, \tag{4.4a}$$

$$\gamma_{2_-}=-\gamma_{2_+} \quad \text{on } \Gamma_2, \tag{4.4b}$$

and since $\mathbf{\Phi}(\alpha)$ is analytic in $l+C$,

$$\mathbf{\Phi}_-(\alpha)=\mathbf{\Phi}_+(\alpha) \quad \text{on } \Gamma_1 \text{ and } \Gamma_2. \tag{4.5}$$

From (4.1) and (4.5) it follows that

$$\mathbf{G}_-(\alpha)\mathbf{\Psi}_-(\alpha)=\mathbf{G}_+(\alpha)\mathbf{\Psi}_+(\alpha) \tag{4.6}$$

or

$$\mathbf{\Psi}_+(\alpha)=\mathbf{H}(\alpha)\mathbf{\Psi}_-(\alpha), \tag{4.7}$$

where with the (+) subscript on γ_{j+} omitted, we have

$$\mathbf{H}(\alpha)=\pm\frac{1}{\gamma_2^2-\gamma_1^2}\begin{pmatrix} \alpha^2(1-k^2/k_2^2) & -2ik_g^2(k_2^2-\alpha^2)/k_2^2 \\ 2ik_g^2 & -\alpha^2(1-k^2/k_2^2) \end{pmatrix}. \tag{4.8}$$

The (+) sign applies to Γ_1, the (−) sign to Γ_2. Equation (4.7) is the desired Hilbert problem.

C. Solution of the Hilbert problem

We introduce a new unknown vector $\mathbf{\Psi}'(\alpha)$ via the transformation

$$\mathbf{\Psi}(\alpha)=\mathbf{T}\mathbf{\Psi}'(\alpha) \tag{4.9}$$

with

$$\mathbf{T}=\begin{pmatrix} \alpha^2(k_2^2-k^2) & 1 \\ 2ik_g^2k_2^2 & 0 \end{pmatrix}. \tag{4.10}$$

Since $\det(\mathbf{T})\neq 0$, $\mathbf{\Psi}'(\alpha)$ will have the same region of analyticity as $\mathbf{\Psi}(\alpha)$, except at infinity. (We assume $k_g \neq 0$; this restriction can be relaxed later.) In terms of $\mathbf{\Psi}'(\alpha)$, (4.7) becomes

$$\mathbf{\Psi}'_+(\alpha)=\mathbf{T}^{-1}\mathbf{H}(\alpha)\mathbf{T}\mathbf{\Psi}'_-(\alpha). \tag{4.11}$$

The matrix of this system is

$$\mathbf{T}^{-1}\mathbf{H}(\alpha)\mathbf{T}=\pm\begin{pmatrix} 0 & [k_2^2(\gamma_2^2-\gamma_1^2)]^{-1} \\ k_2^2(\gamma_2^2-\gamma_1^2) & 0 \end{pmatrix}. \tag{4.12}$$

In component form, (4.11) becomes

$$\Psi'_{1_+}(\alpha)=\pm\Psi'_{2_-}(\alpha)k_2^{-2}(\gamma_2^2-\gamma_1^2)^{-1}, \tag{4.13a}$$

$$\Psi'_{2_+}(\alpha)=\pm k_2^2(\gamma_2^2-\gamma_1^2)\Psi'_{1_-}(\alpha), \tag{4.13b}$$

with the sign convention as before.

Multiplication in (4.13) gives

$$\Psi'_{1_+}(\alpha)\Psi'_{2_+}(\alpha)=\Psi'_{1_-}(\alpha)\Psi'_{2_-}(\alpha) \tag{4.14}$$

which is satisfied by any rational fraction with poles only at $-k_1$ and $-k_2$. It suffices to take

$$\Psi'_1(\alpha)\Psi'_2(\alpha)=(k_1+\alpha)^m(k_2+\alpha)^n, \tag{4.15}$$

where m and n are unspecified integers.

Division in (4.13) yields

$$[\Psi'_1(\alpha)/\Psi'_2(\alpha)]_+\cdot[\Psi'_1(\alpha)/\Psi'_2(\alpha)]_-=k_2^{-4}(\gamma_2^2-\gamma_1^2)^{-2}. \tag{4.16}$$

On taking logs,

$$[\log\Psi'_1(\alpha)/\Psi'_2(\alpha)]_+ +[\log\Psi'_1(\alpha)/\Psi'_2(\alpha)]_- =-\log[k_2^4(\gamma_2^2-\gamma_1^2)^2]. \tag{4.17}$$

This is a standard Hilbert problem, and as shown in Appendix B, Part 1, has the solution

$$\Psi'_1(\alpha)/\Psi'_2(\alpha)=\rho^{-2}(\alpha), \tag{4.18}$$

where

$$\rho(\alpha)=[(k^2-k_2^2)(k_2-k_1)^{-2}s(\alpha,\alpha_1)s(\alpha,-\alpha_1) \times s(\alpha,\alpha_2)s(\alpha,-\alpha_2)]^{1/2}, \tag{4.19}$$

and in which

$$s(\alpha,\zeta)=\eta_1(\alpha)\eta_2(\zeta)+\eta_2(\alpha)\eta_1(\zeta), \tag{4.20}$$

with $\eta_j(\alpha)=(k_j+\alpha)^{1/2}$.

The combination of (4.15) and (4.18) yields

$$\Psi'_1(\alpha)=\pm[\eta_1(\alpha)]^m[\eta_2(\alpha)]^n\rho^{-1}(\alpha), \tag{4.21a}$$

$$\Psi'_2(\alpha)=\pm[\eta_1(\alpha)]^m[\eta_2(\alpha)]^n\rho(\alpha). \tag{4.21b}$$

The signs in (4.21) and m,n are not entirely arbitrary; restrictions are found if (4.21) is substituted in (4.13). We obtain

$$\Psi'_1(\alpha)=[\eta_1(\alpha)]^m[\eta_2(\alpha)]^n\rho^{-1}(\alpha), \tag{4.22a}$$

$$\Psi'_2(\alpha)=(-)^n[\eta_1(\alpha)]^m[\eta_2(\alpha)]^n\rho(\alpha), \quad m+n=\text{odd}. \tag{4.22b}$$

Then, using (4.10), the solutions to the basic Hilbert problem (4.7) are

$$\Psi_1(\alpha)=[\eta_1(\alpha)]^m[\eta_2(\alpha)]^n[\alpha^2(k_2^2-k^2)\rho^{-1}(\alpha)+(-)^n\rho(\alpha)], \tag{4.23a}$$

$$\Psi_2(\alpha)=2ik_g^2k_2^2[\eta_1(\alpha)]^m[\eta_2(\alpha)]^n\rho^{-1}(\alpha), \quad m+n=\text{odd}. \tag{4.23b}$$

D. Lower bounds on m and n

Thus far, the only restrictions on m and n are that they be integers with $m+n$ odd. We now find lower bounds for them. As forecast in Sec. 4, Part A, this is done by requiring analyticity of the functions $\mathbf{G}(\alpha)\mathbf{T}\mathbf{\Psi}'(\alpha)$ in $l+C$.

Theorem: The function $\mathbf{\Phi}(\alpha)=\mathbf{G}(\alpha)\mathbf{T}\mathbf{\Psi}'(\alpha)$ has no branch points in $l+C$.

Proof: Clearly the only possible branch points of $\mathbf{\Phi}(\alpha)$ lie at $-k_1$ and $-k_2$. Now on either Γ_1 or Γ_2 we have

$$\mathbf{\Phi}_+=\mathbf{G}_+\mathbf{T}\mathbf{\Psi}'_+=\mathbf{G}_+\mathbf{T}\mathbf{T}^{-1}\mathbf{H}\mathbf{T}\mathbf{\Psi}'_-=\mathbf{G}_+\mathbf{H}\mathbf{T}\mathbf{\Psi}'_-=\mathbf{G}_+\mathbf{G}_+^{-1}\mathbf{G}_-\mathbf{T}\mathbf{\Psi}'_-=\mathbf{\Phi}_-.$$

Since there is no change in $\mathbf{\Phi}(\alpha)$ across Γ_j, the theorem is proved. An immediate corollary is that $\mathbf{\Phi}(\alpha)$ possesses a power series expansion containing only *even* powers of $\eta_j(\alpha)$ in the neighborhood of $\alpha=-k_j$.

Note that the theorem does not say that $\mathbf{\Phi}(\alpha)$ is analytic in $l+C$. In particular, $\mathbf{\Phi}(\alpha)$ can have poles at $\alpha=-k_j$. The exclusion of these poles sets lower bounds on m and n. An inspection of $\mathbf{G}(\alpha)$ in (3.8) shows that the leading term of $\Phi_2(\alpha)$ is $O\{[\eta_2(\alpha)]^{-1}\Psi_j'(\alpha)\}$ near $\alpha=-k_2$ unless cancellation occurs. If n is even, this cancellation must occur; a direct calculation shows that the leading term is $O\{[\eta_2(\alpha)]^n\}$, from which we deduce that $n>-2$. When n is odd, the leading term is $O\{[\eta_2(\alpha)]^{n-1}\}$. Thus $n>-1$. Near $\alpha=-k_1$, the leading term is $O[\Psi_j'(\alpha)]$ and we obtain $m>-2$ for m even, and $m>-3$ for m odd. Similar but less restrictive results are obtained if $\Phi_1(\alpha)$ is considered. In summary, the constraints

$$m\geq -1,\quad n\geq 0 \quad \text{with } m+n=\text{odd} \tag{4.24}$$

must be satisfied if $\mathbf{\Phi}(\alpha)$ is to be analytic in $l+C$. Therefore, *every* $\mathbf{\Psi}(\alpha)$ *which satisfies* (4.23) *and* (4.24) *is a solution of the homogeneous Wiener–Hopf equation* (4.1).

Before we can obtain upper bounds on m and n we need the form of the solution to the inhomogeneous equation (3.8).

E. Synthesis of the solution

Suppose there are p solution vectors $\mathbf{\Psi}^{(j)}$, $j=1,2,\ldots,p$ to (4.1), of which at least two, say $\mathbf{\Psi}^{(1)}$ and $\mathbf{\Psi}^{(2)}$, have components satisfying $\Psi_1^{(1)}\Psi_2^{(2)}-\Psi_2^{(1)}\Psi_1^{(2)}\neq 0$. Then a solution of the inhomogeneous equation (3.9) is

$$\mathbf{U}(\alpha)=A_0\,\frac{\mathbf{I}-\mathbf{X}(\alpha)\mathbf{X}^{-1}(\alpha_0)}{2\pi i(\alpha-\alpha_0)}\begin{pmatrix} f(\alpha_0,\gamma_0)\\ g(\alpha_0,\gamma_0)\end{pmatrix}+\sum_{j=1}^{p}\lambda_j\mathbf{\Psi}^{(j)}, \tag{4.25}$$

where $\mathbf{I}$ is the unit matrix, λ_j are arbitrary constants and

$$\mathbf{X}(\alpha)=\begin{pmatrix}\Psi_1^{(1)} & \Psi_1^{(2)}\\ \Psi_2^{(1)} & \Psi_2^{(2)}\end{pmatrix}. \tag{4.26}$$

The proof of the statement is immediate. Since $\det[\mathbf{X}(\alpha)]\neq 0$, $\mathbf{X}^{-1}(\alpha)$ exists and so does $\mathbf{U}(\alpha)$. The result then follows by substitution of (4.25) in (3.9).

F. Upper bounds on m and n

These are obtained through the order relations (3.10),

$$V_1(\alpha)=O(\alpha^{\nu_1-1}),\quad V_2(\alpha)=O(\alpha^{-\nu_2-1}),$$

as $|\alpha|\to\infty$ in $u+C$, $0<\nu_j<1$. From Appendix B, Part 2, we have $\rho(\alpha)=O(\alpha)$; thus from (4.23),

$$\Psi_1(\alpha)=O[\alpha^{(m+n)/2+1}], \tag{4.27a}$$

$$\Psi_2(\alpha)=O[\alpha^{(m+n)/2-1}]. \tag{4.27b}$$

Since only powers of $\alpha^{1/2}$ can occur in the asymptotic expression for $\mathbf{V}(\alpha)$, we must have $\nu_1=\nu_2=\frac{1}{2}$. Hence

$$V_1(\alpha)=O(\alpha^{-1/2}),\qquad V_2(\alpha)=O(\alpha^{-3/2}). \tag{4.28a}$$

We have reached a possibly new result: *For a half-plane which is perpendiular to the axis of a gyrotropic medium, the field singularity at the edge is the same as for an edge in free space.*

According to (4.27), $\Psi_1(\alpha)/\Psi_2(\alpha)=O(\alpha^2)$, so if no cancellation of leading terms takes place, (4.25) would give $V_1(\alpha)/V_2(\alpha)=O(\alpha^2)$, contradicting (4.28). Hence cancellation occurs; moreover it must occur in the component $V_1(\alpha)$. Since the ultimate behavior of $V_1(\alpha)$ is $O(\alpha^{-1/2})$, Eq. (4.25) indicates that the maximum exponent of α permitted in (4.27a) is $\frac{3}{2}$. Therefore

$$m+n\leq 1, \tag{4.29}$$

establishing the upper bound and showing that only three vectors are possible. They are denoted as follows:

$$\mathbf{\Psi}^{(1)}:\ m=-1,\quad n=0,$$
$$\mathbf{\Psi}^{(2)}:\ m=0,\quad n=1,$$
$$\mathbf{\Psi}^{(3)}:\ m=1,\quad n=0,$$

of which $\mathbf{\Psi}^{(3)}(\alpha)$ is just $(k_1+\alpha)\mathbf{\Psi}^{(1)}(\alpha)$.

G. The complete solution

We now complete the solution to (3.9). Since $\mathbf{\Psi}^{(1)}$ and $\mathbf{\Psi}^{(2)}$ are independent we can form $\mathbf{X}(\alpha)$ from them [Eq. (4.26)]. ($\mathbf{\Psi}^{(3)}$ and $\mathbf{\Psi}^{(2)}$ could also have been used, with no change in the final result.) We next construct a new vector

$$\mathbf{Z}=-\frac{A_0}{2\pi i}\,\mathbf{X}^{-1}(\alpha_0)\begin{pmatrix} f(\alpha_0,\gamma_0)\\ g(\alpha_0,\gamma_0)\end{pmatrix} \tag{4.30}$$

which can also be expressed as

$$\mathbf{Z}=-\frac{A_0(\alpha_0^2-k_2^2)}{8\pi i k_2^2\alpha_0\gamma_0\rho(\alpha_0)}\begin{pmatrix} 2k_2^2\eta_1(\alpha_0)+\dfrac{\eta_1(\alpha_0)}{k^2-\alpha_0^2-\gamma_0^2}[\alpha_0^2(k_2^2-k^2)-\rho^2(\alpha_0)]\\ -2k_2^2/\eta_2(\alpha_0)-\dfrac{1}{\eta_2(\alpha_0)(k^2-\alpha_0^2-\gamma_0^2)}[\alpha_0^2(k_2^2-k^2)+\rho^2(\alpha_0)]\end{pmatrix}. \tag{4.31}$$

Now $[\mathbf{X}(\alpha)\mathbf{Z}]_1(\alpha-\alpha_0)^{-1}=O(\alpha^{1/2})$ as $|\alpha|\to\infty$, so that only $\Psi^{(1)}$ can be used as the complementary function. From (4.25),

$$\mathbf{V}(\alpha)=\mathbf{X}(\alpha)\mathbf{Z}(\alpha-\alpha_0)^{-1}+\lambda\Psi^{(1)}, \tag{4.32}$$

with the constant λ to be found. With the help of (4.23), Eq. (4.32) becomes

$$V_1(\alpha)=(\alpha-\alpha_0)^{-1}\{[\eta_1(\alpha)]^{-1}\rho^{-1}(\alpha)Z_1[\alpha^2(k_2^2-k^2)+\rho^2(\alpha)]+\eta_2(\alpha)\rho^{-1}(\alpha)Z_2[\alpha^2(k_2^2-k^2)-\rho^2(\alpha)]\}$$
$$+\lambda[\eta_1(\alpha)]^{-1}\rho^{-1}(\alpha)[\alpha^2(k_2^2-k^2)+\rho^2(\alpha)], \tag{4.33a}$$

$$V_2(\alpha)=2ik_g^2k_2^2(\alpha-\alpha_0)^{-1}\rho^{-1}(\alpha)\{Z_1[\eta_1(\alpha)]^{-1}+Z_2\eta_2(\alpha)\}+2ik_g^2k_2^2\lambda[\eta_1(\alpha)]^{-1}\rho^{-1}(\alpha). \tag{4.33b}$$

Now $\rho(\alpha)\to\rho_\infty\alpha$ as $|\alpha|\to\infty$ (Appendix II, Part 2). Then

$$V_1(\alpha)\to\alpha^{1/2}\rho_\infty^{-1}\{[k_2^2-k^2-\rho_\infty^2]Z_2+[k_2^2-k^2+\rho_\infty^2]\lambda\}+O(\alpha^{-1/2}), \tag{4.34a}$$

$$V_2(\alpha)\to 2ik_g^2k_2^2\alpha^{-3/2}\rho_\infty^{-1}(Z_2+\lambda)+O(\alpha^{-5/2}). \tag{4.34b}$$

The leading term of $V_1(\alpha)$ must vanish. This determines λ,

$$\lambda=(\rho_\infty^2-k_2^2+k^2)(\rho_\infty^2+k_2^2-k^2)^{-1}Z_2. \tag{4.35}$$

It can be shown that $\lambda+Z_2\neq 0$; hence $V_2(\alpha)=O(\alpha^{-3/2})$, as required.

Summarizing, the (unique) solution to the inhomogeneous Wiener—Hopf equation (3.9) is given by (4.33), in which $\mathbf{Z}$ is given by (4.31), and λ by (4.35). If the function $\mathbf{L}(\alpha)$ is required, it can be found from (4.33) and (3.8).

5. THE FIELD QUANTITIES

The amplitudes $A(\alpha)$ and $B(\alpha)$ can be found from (3.6), once $\mathbf{V}(\alpha)$ is known,

$$A(\alpha)=\frac{-ik_g^2\gamma_1\alpha}{k_2^2(\gamma_2^2-\gamma_1^2)(k^2-\alpha^2-\gamma_2^2)}\times[ik_g^2V_1(\alpha)+(k^2-\alpha^2-\gamma_2^2)V_2(\alpha)], \tag{5.1a}$$

$$B(\alpha)=\frac{ik_g^2\gamma_2\alpha}{k_2^2(\gamma_2^2-\gamma_1^2)(k^2-\alpha^2-\gamma_1^2)}\times[ik_g^2V_1(\alpha)+(k^2-\alpha^2-\gamma_1^2)V_2(\alpha)]. \tag{5.1b}$$

Of the eight possible branch points, $+k_2$ is missing from $A(\alpha)$ while $+k_1$ is missing from $B(\alpha)$. The only pole is at $\alpha=\alpha_0$.

The field quantities may now be found by substituting (5.1) in (3.2), (3.4), and (3.5). However, it seems impossible to evaluate the integrals in terms of tabulated functions in the general case. Even a steepest descents approximation, although possible,[1] is very difficult and would be beyond the scope of the present paper. Instead we discuss some general properties and special cases.

A. General properties

We first show that the integrals (3.2), (3.4), and (3.5) converge. For $z>0$ the terms $\exp(i\gamma_jz)$ ensure this, since $\mathrm{Im}(\gamma_j)>0$ as $\alpha\to\pm\infty$ (Appendix A, Part 4). When $z=0$, a more delicate investigation is needed. We already have $V_1(\alpha)=O(\alpha^{-1/2})$ and $V_2(\alpha)=O(\alpha^{-3/2})$. It is easy to check that $\gamma_1\to i|\alpha|$ and $\gamma_2\to ikk_2^{-1}|\alpha|$ as $\alpha\to\pm\infty$. Hence $\gamma_2^2-\gamma_1^2=O(\alpha^2)$, $k^2-\alpha^2-\gamma_1^2=O(1)$, and $k^2-\alpha^2-\gamma_2^2=O(\alpha^2)$. Thus $A(\alpha)=O(\alpha^{-3/2})$ and $B(\alpha)=O(\alpha^{-1/2})$, so all the integrals converge. Incidentally, (3.7) shows that $L_1(\alpha)=O(\alpha^{-1/2})$ and $L_2(\alpha)=O(\alpha^{-3/2})$.

It is interesting that the points $\pm\alpha_1$ and $\pm\alpha_2$ are not branch points of the integrands (3.2), (3.4), and (3.5). This can be seen from the following argument. Clearly, $\mathbf{V}(\alpha)$ does not have branch points at $\pm\alpha_j$. Also, $\gamma_1\leftrightarrow\gamma_2$ when the point α_j is circled. Hence $A(\alpha)\leftrightarrow B(\alpha)$, by (5.1); and $f(\alpha,\gamma_1)\leftrightarrow f(\alpha,\gamma_2)$, $g(\alpha,\gamma_1)\leftrightarrow g(\alpha,\gamma_2)$. Thus the total integrands are unchanged and, as mentioned in Sec. 2, the position of the α_j relative to the k_j is immaterial. Note that these results are not true of the individual ordinary and and extraordinary components.

Lastly, the point $\alpha=\alpha_0$ is easily seen to be a pole only of $A(\alpha)$ for γ_1 type incidence and only of $B(\alpha)$ for γ_2 type incidence.

B. Field components in the plane $z=0$

In this case, it is fairly easy to obtain far field expressions. Consider the E_y component

$$E_y=\int_C V_2(\alpha)\exp(i\alpha x)\,d\alpha. \tag{5.2}$$

For $x>0$, the contour C can be closed by a large semicircle in u. The only contribution is from the pole $\alpha=\alpha_0$, and yields $E_y=-E_y^{(i)}$. When $x<0$, the contour is closed by a large semicircle in l, indented around the branch cuts Γ_1 and Γ_2. The semicircle contribution vanishes and we are left with

$$E_y=\int_{\Gamma_1+\Gamma_2}[V_{2_+}(\alpha)-V_{2_-}(\alpha)]\exp(i\alpha x)\,d\alpha, \tag{5.3}$$

$$V_{2_+}(\alpha)-V_{2_-}(\alpha)=2ik_g^2k_2^2\rho^{-1}(\alpha)\left\{\left[\frac{Z_1}{\alpha-\alpha_0}+\lambda\right]\times\left[1\mp\frac{\rho^2(\alpha)}{k_2^2(\gamma_2^2-\gamma_1^2)}\right]\frac{1}{\eta_1(\alpha)}+\frac{Z_2\eta_2(\alpha)}{\alpha-\alpha_0}\left[1\pm\frac{\rho^2(\alpha)}{k_2^2(\gamma_2^2-\gamma_1^2)}\right]\right\}, \tag{5.4}$$

with the top signs for Γ_1. For large $|x|$, the dominant contributions come from near the points $-k_1$ and $-k_2$ and yield

$$E_y\approx K_1|x|^{-1/2}\exp(-ik_1x)+K_2|x|^{-3/2}\exp(-ik_2x),\quad K_j=\text{const.} \tag{5.5}$$

This result tacitly assumes that only $\eta_j(\alpha)$ in (5.4) is rapidly varying near $\alpha=-k_j$. This is true of the terms $[1\pm\rho^2(\alpha)k_2^2(\gamma_2^2-\gamma_1^2)^{-1}]$ but not easily decided for $\lambda+Z_1(\alpha-\alpha_0)^{-1}$. It is tempting to conjecture that the latter is zero at $\alpha=-k_1$ when the incident field is of

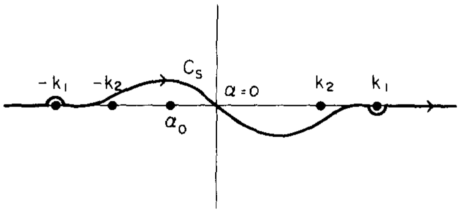


FIG. 3. Steepest descents contour C_s showing capture of the pole $\alpha=\alpha_0$.

the γ_2 type; otherwise we deduce that the scattered γ_2 type field is always dominated by the γ_1 type regardless of the kind of incident wave. Similar results are obtained for the E_x component.

For E_z we have

$$E_z = -k_2^{-2}\int_C \alpha L_2(\alpha)\exp(i\alpha x)\,d\alpha; \tag{5.6}$$

hence $E_z=0$ for $x<0$, $z=0$, as we had assumed earlier.

C. Field components in the plane $x=0$

In this case, a steepest descents evaluation for large z can be carried out without too much trouble. Saddle points are given by the zeros of

$$\gamma_j' = \frac{d\gamma_j}{d\alpha}=0. \tag{5.7}$$

The only relevant solution is $\alpha=0$; all others give exponentially small contributions as $z\to\infty$. We find

$$E_x \approx \left(\frac{-2\pi}{z}\right)^{1/2}\exp(i\pi/4)\times\left[\frac{f(0,\gamma_1)A(0)}{\sqrt{\gamma_1''}}\exp(i\gamma_1 z)+\frac{f(0,\gamma_2)B(0)}{\sqrt{\gamma_2''}}\exp(i\gamma_2 z)\right], \tag{5.8a}$$

$$E_y \approx \left(\frac{-2\pi}{z}\right)^{1/2}\exp(i\pi/4)\times\left[\frac{g(0,\gamma_1)A(0)}{\sqrt{\gamma_1''}}\exp(i\gamma_1 z)+\frac{g(0,\gamma_2)B(0)}{\sqrt{\gamma_2''}}\exp(i\gamma_2 z)\right]. \tag{5.8b}$$

The quantities γ_j and γ_j'' are evaluated at $\alpha=0$. The steepest descents contour is shown in Fig. 3. Note that if $\alpha_0<0$, this pole will be crossed when deforming the contour. This is proper since the field point lies in the geometrical shadow. No significant simplifications of the expressions (5.8) seem possible.

D. Reduction to uniaxial medium

In the limit $\epsilon_g=0$, the medium is uniaxial and the Wiener–Hopf system (3.8) splits into two distinct equations which are solvable by standard techniques. Although there are solutions in the literature,[8] it is easier, for purposes of comparison, to solve the reduced set (3.8) directly.

As $\epsilon_g\to 0$ we see that $\alpha_1,\alpha_2\to 0$, $k_1\to k$, $\gamma_1^2\to k^2-\alpha^2$, and $\gamma_2^2\to(k^2/k_2^2)(k_2^2-\alpha^2)$. Then (3.8) reduces to

$$-2(k^2-\alpha^2)^{1/2}V_2(\alpha)=L_1(\alpha), \tag{5.9a}$$

$$2kk_2V_1(\alpha)(k_2^2-\alpha^2)^{-1/2}=L_2(\alpha). \tag{5.9b}$$

These have the solutions

$$V_2(\alpha)=-A_0 g(\alpha_0,\gamma_{10})\eta_1(\alpha_0)/2\pi i(\alpha-\alpha_0)\eta_1(\alpha) \tag{5.10}$$

when the incident wave is a γ_1 type; and

$$V_1(\alpha)=-A_0 f(\alpha_0,\gamma_{20})\eta_2(\alpha)/2\pi i(\alpha-\alpha_0)\eta_2(\alpha_0) \tag{5.11}$$

for γ_2 type incidence. [We have introduced the notation $\gamma_{j0}=\gamma_j(\alpha_0)$.]

Turning to our solution, as given in (4.33), we first note the following limiting values as $\epsilon_g\to 0$.

$$\rho(\alpha)=(k+k_2)^{1/2}(k-k_2)^{-1/2}\times[2kk_2+\alpha(k+k_2)+2\sqrt{kk_2}\,\eta_1(\alpha)\eta_2(\alpha)], \tag{5.12a}$$

$$\alpha^2\rho^{-1}(\alpha)(k_2^2-k^2)+\rho(\alpha)=-4\sqrt{kk_2}\,[(k+k_2)/(k-k_2)]^{1/2}\eta_1(\alpha)\eta_2(\alpha), \tag{5.12b}$$

$$\alpha^2\rho^{-1}(\alpha)(k_2^2-k^2)-\rho(\alpha)=2[(k+k_2)/(k-k_2)]^{1/2}(2kk_2+k\alpha+k_2\alpha), \tag{5.12c}$$

$$\lambda=\tfrac{1}{2}(k+k_2)Z_2(kk_2)^{-1/2}. \tag{5.12d}$$

Thus

$$V_1(\alpha)\to\eta_2(\alpha)\left(\frac{k+k_2}{k-k_2}\right)^{1/2}\left[\frac{-4\sqrt{kk_2}\,Z_1+2(2kk_2+\alpha k+\alpha k_2)Z_2}{\alpha-\alpha_0}-2(k+k_2)Z_2\right], \tag{5.13a}$$

$$V_2(\alpha)\to 2ik_g^2k_2^2\rho^{-1}(\alpha)\times\left[\frac{Z_1\eta_1(\alpha)^{-1}+Z_2\eta_2(\alpha)}{\alpha-\alpha_0}+\frac{(k+k_2)Z_2}{2\sqrt{kk_2}\,\eta_1(\alpha)}\right]. \tag{5.13b}$$

The quantities Z_1 and Z_2 take rather different forms depending on the type of incident wave. Suppose it is of γ_1 type. Then $k^2-\alpha_0^2-\gamma_{10}^2\to 0$ and

$$Z_1\to\frac{2C_0\eta_2(\alpha_0)}{k^2-\alpha_0^2-\gamma_{10}^2}\left(\frac{k+k_2}{k-k_2}\right)^{1/2}[2kk_2+\alpha_0(k+k_2)], \tag{5.14a}$$

$$Z_2\to\frac{4C_0\eta_2(\alpha_0)\sqrt{kk_2}}{k^2-\alpha_0^2-\gamma_{10}^2}\left(\frac{k+k_2}{k-k_2}\right)^{1/2}, \tag{5.14b}$$

where

$$C_0=-\frac{A_0}{8\pi i}\,\frac{\eta_1(\alpha_0)}{\eta_2(\alpha_0)}\,\frac{(\alpha_0^2-k_2^2)}{k_2^2\alpha_0\gamma_{10}}.$$

Upon substitution in (5.13) we obtain

$$V_1(\alpha)=0, \tag{5.15a}$$

$$V_2(\alpha)=\frac{A_0}{2\pi}\cdot\frac{\eta_1(\alpha_0)}{\eta_1(\alpha)}\cdot\frac{k_g^2}{k^2-\alpha_0^2-\gamma_{10}^2}\cdot\frac{\alpha_0^2-k_2^2}{\alpha_0\gamma_{10}}\,\frac{1}{\alpha-\alpha_0}, \tag{5.15b}$$

which agrees with (5.10). [Note that (5.15b) involves taking the limit $k_g^2(k^2-\alpha_0^2-\gamma_{10}^2)^{-1}$, which turns out to be infinite. This is an expression of the fact that $E_z=0$ for the γ_1 mode in the uniaxial case.]

When the incidence is of the γ_2 type, we find

$$Z_1\to\frac{4C_0(k_2+\alpha_0)\gamma_{10}}{(k^2-\alpha_0^2-\gamma_{20}^2)\gamma_{20}}\sqrt{kk_2}\left(\frac{k+k_2}{k-k_2}\right)^{1/2}\eta_1(\alpha_0), \tag{5.16a}$$

$$Z_2\to\frac{2C_0[2kk_2+\alpha_0(k+k_2)]\gamma_{10}}{(k^2-\alpha_0^2-\gamma_{20}^2)\eta_1(\alpha_0)\gamma_{20}}\left(\frac{k+k_2}{k-k_2}\right)^{1/2}, \tag{5.16b}$$

which yields

$$V_1(\alpha) = -\frac{A_0}{2\pi i}\cdot\frac{\alpha_0^2 - k_2^2}{\alpha_0\gamma_{20}}\cdot\frac{\eta_2(\alpha)}{\eta_2(\alpha_0)}\cdot\frac{1}{\alpha-\alpha_0}, \tag{5.17a}$$

$$V_2(\alpha) = 0, \tag{5.17b}$$

in agreement with (5.11).

6. CONCLUSION

We have solved a new diffraction problem—that of a plane wave incident upon a half-plane embedded in a gyrotropic medium whose distinguished axis is perpendicular to the plate. The method used is the newly invented Wiener—Hopf—Hilbert technique,[3] in which a pair of homogeneous simultaneous Wiener—Hopf equations is reduced to a much simpler pair of Hilbert problems on the branch cuts. A set of characteristic solutions to the latter is obtained, and from this set the entire solution to the problem is obtained. The solution satisfies all the conditions of the problem, and must, by the usual uniqueness condition, be correct.

The solution requires certain inequalities to hold for the components of the permittivity tensor. However, we believe that the problem can also be solved when the inequalities are relaxed, but a complete investigation would involve much more work (e.g., see Refs. 12 and 13) and is beyond the scope of the present paper.

Other generalizations are possible. We can also solve the problem of a similarly oriented half-plane, but with magnetic and electric anisotropy and arbitrary skew incidence. This problem will be treated in a subsequent paper.

ACKNOWLEDGMENT

We wish to thank S. R. Seshadri for his kindness in sending us some unpublished notes on the behavior of the propagation constants.

APPENDIX A: DERIVATION OF SOME USEFUL RELATIONS

1. Relative positions of k_1 and k_2

Equation (2.2a) shows that $k_2(=k_a)$ is real and positive. We have from (2.13a),

$$\begin{aligned} k_1^2 - k_2^2 &= k^2 - k_g^4 k^{-2} - k_a^2 \\ &= k^{-2}[k^2(k^2-k_a^2) - k_g^4] \\ &> k^{-2}[(k^2-k_a^2)^2 - k_g^4] \quad \text{by (2.2a)} \\ &> 0 \quad \text{by (2.2b)}. \end{aligned}$$

This establishes the relative positions of k_1 and k_2 on the real axis; if a small loss is added to the medium it can be shown that they are displaced into the first quadrant. This consideration determines the placement of the contour C.

2. Relative positions of α_1 and α_2

According to (2.11), the real part of α_j^2 ($j=1,2$) is always positive; and by (2.2b), $(k^2-k_a^2)^2 - k_g^4$ is also positive. Hence, α_1^2 lies in the first quadrant and α_2^2 in the fourth. After taking the square root in a consistent way, we deduce that α_j lies in the jth quadrant. It is seen that only in the limit $k_g=0$ does α_j approach the real axis; this limiting value is $\alpha_j=0$. No precise statement can be made about $\text{Re}(\alpha_1)$; it can be greater than k_1 or less than k_2. To make the problem definite, we assume $\text{Re}(\alpha_1) < k_2$ and eventually demonstrate that the relative position is immaterial.

3. The branch points and cuts of γ_1 and γ_2

As is evident from (2.9), (2.10b), and (2.12c), the possible branch points of γ_j lie at the eight points $\pm k_1$, $\pm k_2$, $\pm\alpha_1$, and $\pm\alpha_2$. Consider $D(\alpha)$, defined in (2.10a). For $\alpha=0$ it is positive and changes sign at $\alpha^2 = 2k^2k_a^2(k^2+k_a^2)^{-1}$ $(=k_3^2$, say). Now

$$k_3^2 > 2k^2k_a^2(k^2+k^2)^{-1} = k_a^2 = k_2^2. \tag{A1}$$

Also,

$$\begin{aligned} k_1^2 &= k^2 - k_g^4 k^{-2} \\ &> k^2 - k^{-2}(k^2-k_a^2)^{-2} \quad \text{by (2.2b)} \\ &= k_a^2(2 - k_a^2k^{-2}) \\ &= k_a^2(k^2+k_a^2)^{-1}[2k^2 + k_a^2(1-k_a^2k^{-2})] \\ &> 2k_a^2k^2(k^2+k_a^2)^{-1} = k_3^2. \end{aligned} \tag{A2}$$

Hence, $D(\pm k_2) > 0$ and $D(\pm k_1) < 0$. It will be shown subsequently that $\Delta(\alpha)$ is real and negative for all real α. Since the zeros of $D(\alpha) \pm \Delta(\alpha)$ lie at $\pm k_1$ and $\pm k_2$ by (2.12c) and are simple, we must have

$$D(\pm k_2) + \Delta(\pm k_2) = 0, \tag{A3a}$$

$$D(\pm k_1) - \Delta(\pm k_1) = 0. \tag{A3b}$$

Therefore, $\gamma_j(\pm k_j) = 0$. Quite clearly $\pm\alpha_j$ $(j=1,2)$ are branch points of both γ_1 and γ_2; thus we have the result that γ_j has six branch points: $\pm k_j$, $\pm\alpha_1$, $\pm\alpha_2$.

Figure 1 shows the placement of the branch points and cuts. We have to choose branches such that $\Delta(\alpha)$ is negative for all real α and such that γ_1 and γ_2 are either positive real or positive imaginary for real α. Consider a typical square root $\sigma(\alpha) = (\zeta^2-\alpha^2)^{1/2}$ and choose the branch for which $\sigma(0)=\zeta$. If the real part of ζ is positive, the signs of $\sigma(\alpha)$ as a function of α are as shown in Fig. 4; while if negative, Fig. 5 applies. Since $\Delta(\alpha)$ contains the product of a similar pair of

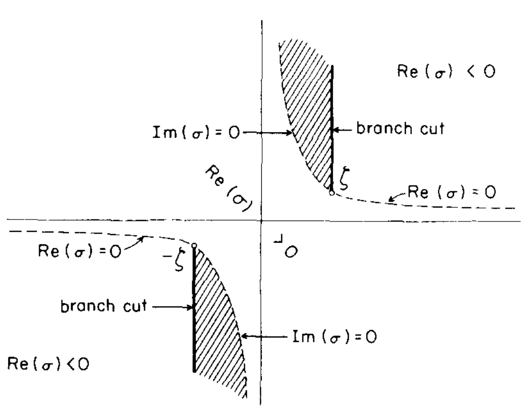


FIG. 4. Branch cuts of a typical square root $\sigma = (\zeta^2-\alpha^2)^{1/2}$ for ζ in the first quadrant. The signs of $\text{Re}(\sigma)$ are shown. In the shaded area $\text{Im}(\sigma) < 0$, elsewhere $\text{Im}(\sigma) \geq 0$.

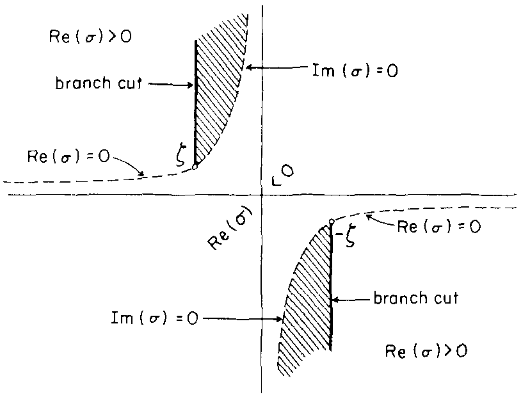


FIG. 5. Branch cuts of a typical square root $\sigma=(\zeta^2-\alpha^2)^{1/2}$ when ζ lies in the second quadrant. The signs of $\mathrm{Re}(\sigma)$ are shown. In the shaded area $\mathrm{Im}(\sigma)<0$, elsewhere $\mathrm{Im}(\sigma)\geq 0$.

roots, we deduce that $\mathrm{Re}[\Delta(\alpha)]<0$ for all real α. There remains the choice of the sign of the root in (2.9); we choose it so that $\gamma_j(0)>0$, $j=1,2$. With this branch, $\mathrm{Im}(\gamma_j)>0$ for $|\alpha|>k_j$, if the branch points are circled as shown in Fig. 1. We do not show it, but if the medium is slightly lossy, $\mathrm{Im}(\gamma_j)>0$ for *all* real α.

4. Behavior of γ_j near the branch points $\pm k_j$

This follows immediately from (2.12c) and the result in Appendix A, Part 3, that the points $\pm k_1$ are not branch points of γ_2, nor $\pm k_2$ of γ_1. Thus γ_j behaves as $(k_j^2-\alpha^2)^{1/2}F_j(\alpha)$, where $F_j(\alpha)$ is analytic on and near the branch cut contours from $\pm k_j$. It is then clear that γ_j changes sign as the branch cut is crossed.

5. Combinations of γ_1 and γ_2

It is quite obvious from their definitions that $\gamma_1+\gamma_2$ and $\gamma_1\gamma_2$ do not have branch points at $\pm\alpha_1$ and $\pm\alpha_2$. Furthermore, $\gamma_1+\gamma_2$ is never zero in the finite α plane, for if it were, we would have $\gamma_1^2=\gamma_2^2$, giving $\Delta(\alpha)=0$, and $\alpha=\pm\alpha_j$, $j=1,2$. Then $\gamma_1=\pm[\frac{1}{2}D(\alpha_j)]^{1/2}$, $\gamma_2=\pm[\frac{1}{2}D(\alpha_j)]^{1/2}$; but γ_1 and γ_2 are both positive at $\alpha=0$, so the same signs must be chosen. Therefore $\gamma_1=\gamma_2$ and $\gamma_1+\gamma_2\neq 0$.

6. The zeros of $k^2-\alpha^2-\gamma_1^2$ and $k^2-\alpha^2-\gamma_2^2$

Consider the product $(k^2-\alpha^2-\gamma_1^2)(k^2-\alpha^2-\gamma_2^2)$. On multiplying and using (2.12) we obtain

$$(k^2-\alpha^2-\gamma_1^2)(k^2-\alpha^2-\gamma_2^2)=k_g^4k_2^{-2}(\alpha^2-k_2^2). \tag{A4}$$

Now $\gamma_2(\pm k_2)=0$, hence $k^2-\alpha^2-\gamma_2^2$ is not zero at $\alpha=\pm k_2$; so $k^2-\alpha^2-\gamma_1^2$ has simple zeros at $\alpha=\pm k_2$, while $k^2-\alpha^2-\gamma_2^2$ has none.

APPENDIX B: DERIVATION AND PROPERTIES OF $\rho(\alpha)$

1. Derivation

The Hilbert problem (4.17) can be written in the form

$$\phi_+(\alpha)+\phi_-(\alpha)=g(\alpha), \tag{B1}$$

where $\phi(\alpha)=\log[\Psi_1'(\alpha)/\Psi_2'(\alpha)]$ and $g(\alpha)=-\log[k_2^4(\gamma_2^2-\gamma_1^2)^2]$. Introduce

$$\psi(\alpha)=[\eta_1(\alpha)\eta_2(\alpha)]^{-1}\phi(\alpha), \tag{B2}$$

and substitute in (B1),

$$\psi_+(\alpha)-\psi_-(\alpha)=[\eta_1(\alpha)\eta_2(\alpha)]^{-1}g(\alpha). \tag{B3}$$

This is a standard Hilbert problem with solution

$$\psi(\alpha)=\frac{1}{2\pi i}\int_{\Gamma_1+\Gamma_2}\frac{g(t)\,dt}{\eta_1(t)\eta_2(t)(t-\alpha)}\,. \tag{B4}$$

To evaluate (B4), we first convert it to an integral *around* the branch cuts,

$$\psi(\alpha)=\frac{1}{\pi i}\int\frac{g(t)\,dt}{\eta_1(t)\eta_2(t)(t-\alpha)}\,. \tag{B5}$$

For $|t|\to\infty$, the integrand is $O(\log t/t^2)$ and the integral on a large circle vanishes. Hence the integral (B5) can be replaced by minus the contributions from the pole $t=\alpha$ and the four branch cut integrals from $\pm\alpha_1$ and $\pm\alpha_2$. On these contours, $g(t)$ becomes very simple and the integrations elementary. Omitting the details, we find that

$$\psi(\alpha)=[\eta_1(\alpha)\eta_2(\alpha)]^{-1}\log[1/\rho^2(\alpha)], \tag{B6}$$

where

$$\rho^2(\alpha)=(k^2-k_2^2)(k_2-k_1)^{-2}s(\alpha,\alpha_1)\times s(\alpha,-\alpha_1)s(\alpha,\alpha_2)s(\alpha,-\alpha_2) \tag{B7}$$

and

$$s(\alpha,\zeta)=\eta_1(\alpha)\eta_2(\zeta)+\eta_2(\alpha)\eta_1(\zeta). \tag{B8}$$

Finally using (B2), we get

$$\psi_1'(\alpha)/\psi_2'(\alpha)=\rho^{-2}(\alpha) \tag{B9}$$

and we can verify by direct substitution that (4.16) is satisfied.

2. Analytic properties of $\rho(\alpha)$

The only possible singularities of $\rho(\alpha)$ or $\rho^{-1}(\alpha)$ are at zeros of $s(\alpha,\alpha_j)$. Suppose $s(\alpha,-\alpha_1)=0$. On squaring, we have

$$(k_1+\alpha)(k_2-\alpha_1)=(k_2+\alpha)(k_1-\alpha_1), \tag{B10}$$

whose only solution is $\alpha=-\alpha_1$. But $s(-\alpha_1,-\alpha_1)=2\sqrt{(k_1-\alpha_1)(k_2+\alpha_1)}$, and this is nonzero, since $\alpha_1\neq k_1$ or $-k_2$ by the results of Appendix A, Parts 1 and 2. Therefore, $\rho(\alpha)$ and $\rho^{-1}(\alpha)$ are analytic in $u+l+C-\Gamma_1-\Gamma_2$.

As $|\alpha|\to\infty$ in $u+C$ we derive directly that

$$\rho(\alpha)\to\rho_\infty\alpha, \tag{B11}$$

where

$$\rho_\infty^2=(k^2-k_2^2)(k_1-k_2)^{-2}h(\alpha_1)h(-\alpha_1)h(\alpha_2)h(-\alpha_2) \tag{B12}$$

and

$$h(\alpha)=\eta_1(\alpha)+\eta_2(\alpha).$$

Lastly, we record the barrier equation satisfied by $\rho(\alpha)$ on Γ_1 and Γ_2,

$$\rho_-(\alpha)\rho_+(\alpha)=-k_2^2\Delta(\alpha). \tag{B13}$$

This can be verified directly from (B7) and shows that (4.18) is a solution to (4.16).

*S. Przeździecki is on leave of absence from the Institute of Fundamental Technological Research, Polish Academy of Sciences, Swietokrzyska 21, 00-049 Warsaw, Poland.
[1]S. Rosenbaum, Can. J. Phys. **45**, 3479 (1967).
[2]R.A. Hurd, Can. J. Phys. **45**, 2561 (1967).
[3]R.A. Hurd, Can. J. Phys. **54**, 775 (1976).
[4]S.R. Seshadri and A.K. Rajagopal, IEEE Trans. Antennas Propag. **11**, 497 (1963).
[5]E.V. Jull, Can. J. Phys. **42**, 1455 (1964).
[6]P. De Santis, Alta Freq. **35**, 112 (1966).
[7]L.B. Felsen, Proc. Inst. Electr. Eng. **111**, 454 (1964).
[8]B. Rulf, SIAM J. Appl. Math. **15**, 120 (1967).
[9]S. Przeździecki, J. Appl. Phys. **37**, 2768 (1966).
[10]W.E. Williams, J. Inst. Math. Its Appl. **2**, 186 (1966).
[11]S. Rosenbaum, Research Report No. PIBMRI-1320-66, Microwave Research Institute, Polytechnic Inst. Brooklyn (1966).
[12]S.R. Seshadri, Can. J. Phys. **46**, 1473 (1968).
[13]E. Arbel and L.B. Felsen in *Electromagnetic Theory and Antennas Part I*, edited by E.C. Jordan (Pergamon, Oxford, 1963), p. 391.
[14]For a cold electron plasma characterized by a plasma frequency ω_p and a gyro frequency ω_c, postulates (2.2) follow from assuming that the operating frequency ω satisfies $\omega_p < \omega < \omega_c$.
[15]Subsequently, it will be shown that the position of α_j is of no consequence, so long as it is not on the real axis ($\alpha_j = 0$ excepted).
[16]The solution (4.18) is not unique. Any function of the form $\exp[\eta_1(\alpha)^m \eta_2(\alpha)^n P(\alpha)]^{-2}(\alpha)$ satisfies (4.16) for an arbitrary polynomial $P(\alpha)$ and odd integers m and n. A requirement of algebraic behavior at infinity and avoidance of essential singularities at $\alpha + k_j = 0$ demands $P(\alpha) = 0$.

Bogoliubov inequality for unbounded operators and the Bose gas

M. Bouziane and Ph. A. Martin

Laboratoire de Physique Théorique, Ecole Polytechnique Fédérale, Lausanne, Switzerland
(Received 19 January 1976)

We provide the mathematical arguments which are needed to obtain a rigorous proof of the absence of condensation in a one- and two-dimensional Bose gas of particles having superstable interactions.

I. INTRODUCTION

The demonstration of the absence of condensation in a one- or two-dimensional interacting Bose gas relies on a well known argument[1] which makes use of inequalities originally due to Bogoliubov.[2] The Bogoliubov inequality is

$$\tfrac{1}{2}\beta\langle AA^* + A^*A\rangle\langle[[C,H],C^*]\rangle \geq |\langle[C,A]\rangle|^2, \tag{1}$$

where H is the Hamiltonian of the system in question, $\langle\circ\rangle$ denotes the thermal average with respect to the temperature $T=(k\beta)^{-1}$ and the Hamiltonian H. A and C are observables of the system which have to be conveniently chosen in view of specific applications.

A mathematically correct use of the Bogoliubov inequalities in statistical mechanics requires two steps. In the first step one establishes the existence of the averages and inequality (1) for a finite volume system, whereas the second one consists in a proper handling of the thermodynamic limit. This has been done for quantum systems whose finite volume description involves only a finite-dimensional Hilbert space (for instance, the proof of absence of long-range order in the one- and two-dimensional isotropic Heisenberg ferromagnet[3,4]), and also for a certain class of classical lattice systems.[5,6] Elegant proofs of the Bogoliubov inequalities have also been obtained in the framework of statistical mechanics of infinite systems for states satisfying the Kubo—Martin—Schwinger condition[7] (see also in Ref. 8). However, none of these existing proofs cover the case of the Bose gas.

The reason for this is that already at finite volume states of the Bose gas are described in an infinite-dimensional Hilbert space, and the operators A, C, H which are used in (1) are of an unbounded nature. In such a situation operator relations like (1) need to be treated with care, for it is well known that formal manipulations of unbounded operators may lead to paradoxes. {Set for instance $H=p^2/2m$, $C=p$, $A=q$ with $[q,p]=i\hbar$, then (1) gives formally $\hbar^2 \leq 0$!}

As far as an infinite volume description of the interacting Bose gas is concerned, the question of the existence of KMS states has not yet been settled and the validity of the formulation of Ref. 7 is far from obvious.

Since there are very few exact results on the interacting Bose gas, it seems to be desirable to put on a firm basis the available pieces of information. Thus the purpose of this note is to provide the mathematical arguments which are needed to make completely rigorous the proof of absence of condensation in a one- or two-dimensional Bose gas with superstable interactions. The superstability condition will be seen to play an essential role.

Technically, we extend slightly the study of some operator relations done by Ginibre in Ref. 9. For the application considered in Ref. 9 it was sufficient to assume that the operators occurring in the Bogoliubov inequalities were bounded by the number of particles operator N. The point is that this will not be the case here and questions of domain have to be examined carefully. However, uniform bounds which are needed to control the thermodyanic limit may be taken from Ref. 9.

In paragraph 2, we indicate in which sense the Bogoliubov inequalities have to be understood for a general class of unbounded operators, and the specific application to the Bose gas in given in paragraph 3.

II. BOGOLIUBOV INEQUALITY FOR UNBOUNDED OPERATORS

Let $\mathcal{H}$ be a separable Hilbert space and H be a self-adjoint operator on $\mathcal{H}$ with discrete spectrum. We shall assume throughout the following that $\exp(-\beta H)$ belongs to the trace class of operators $\mathcal{L}_1(\mathcal{H})$ for all $\beta>0$. We consider the set $\mathcal{C}_H$ of linear operators A on $\mathcal{H}$ with domain $\mathcal{D}(A)$ having the property

$$\mathcal{C}_H = \{A \,|\, \mathcal{D}(A)\cap\mathcal{D}(A^*)\supset\mathcal{D}(\exp(\beta H)) \text{ for all } \beta>0\}$$

$\mathcal{C}_H$ is a linear manifold. Moreover, if $A\in\mathcal{C}_H$, its adjoint A^* belongs also to $\mathcal{C}_H$. $\mathcal{D}(\exp(\beta H))$ being dense in $\mathcal{H}$, A and A^* are densely defined and hence closable. (For the relevant mathematical concepts see Ref. 10).

On $\mathcal{C}_H$ we introduce the two following sesquilinear forms for fixed positive β:

$$\langle A,B\rangle_N = \frac{1}{Z}\sum_{n=1}^{N}(A\varphi_n, B\varphi_n)\exp(-\beta\lambda_n), \tag{2}$$

$$(A,B)_N = \frac{1}{Z}\sum_{n=1}^{N}\sum_{m=1}^{N}(A\varphi_n,\varphi_m)(\varphi_m,B\varphi_n) \times \frac{\exp(-\beta\lambda_m)-\exp(-\beta\lambda_n)}{\lambda_n-\lambda_m},$$

where $\{\varphi_n, n=1,2,\ldots\}$ is an orthonormal basis of eigenvectors of H with corresponding eigenvalue λ_n, and $Z=\operatorname{Tr}\exp(-\beta H)$. The operators A and B are obviously well defined on the eigenvectors φ_n and the positive factor $[\exp(-\beta\lambda_m)-\exp(-\beta\lambda_n)]/(\lambda_n-\lambda_m)$ set equal to $\beta\exp(-\beta\lambda_n)$ when $\lambda_n=\lambda_m$.

Lemma 1:

If A and B belong to $\mathcal{C}_H$ then

$$\langle A,B\rangle=\lim_{N\to\infty}\langle A,B\rangle_N \text{ and } (A,B)=\lim_{N\to\infty}(A,B)_N$$

exist and defined sesquilinear forms on $\mathcal{C}_H$. (A,B) is a positive definite scalar product on $\mathcal{C}_H$.

The proof of the convergence relies on the fact that the sums $\sum_{n=1}^{\infty}(\varphi_n,\chi\varphi_n)$ are absolutely convergent when χ is a trace class operator. We remark first that $A\exp(-\beta H)$ is bounded. Indeed if $A\in\mathcal{C}_H$, $A\exp(-\beta H)$ is defined everywhere and closed since A^* is densely defined. Therefore by the closed graph theorem, $A\exp(-\beta H)$ is bounded and has a bounded adjoint. Moreover, since we can decompose $A\exp(-\beta H)$ $=A\exp(-\beta H/2)\exp(-\beta H/2)$, we conclude that $A\exp(-\beta H)$ belongs to $\mathcal{L}_1(\mathcal{H})$ for all $\beta>0$. Hence, $\langle A,B\rangle_N$ which can also be written as

$$\langle A,B\rangle_N=\frac{1}{Z}\sum_{n=1}^{N}\left(A\exp\left(-\frac{\beta H}{2}\right)\varphi_n, B\exp\left(-\frac{\beta H}{2}\right)\varphi_n\right)$$
$$=\frac{1}{Z}\sum_{n=1}^{N}\left(\varphi_n,\left[A\exp\left(-\frac{\beta H}{2}\right)\right]^* B\exp\left(-\frac{\beta H}{2}\right)\varphi_n\right)$$

converges as $N\to\infty$ when A and B belong to $\mathcal{C}_H$.

Using the inequality $[\exp(-\beta x)-\exp(-\beta y)]/(y-x)$ $\le\frac{1}{2}\beta(\exp(-\beta x)+\exp(-\beta y))$ and the fact that all terms are positive, we majorize

$$(A,A)_N \le\frac{\beta}{2Z}\sum_{n=1}^{N}\sum_{m=1}^{N}(A\varphi_n,\varphi_m)(\varphi_m,A\varphi_n)(\exp(-\beta\lambda_m)+\exp(-\beta\lambda_n))$$
$$\le\lim_{N\to\infty}\tfrac{1}{2}\beta(\langle A^*,A^*\rangle_N+\langle A,A\rangle_N)=\tfrac{1}{2}\beta(\langle A^*,A^*\rangle+\langle A,A\rangle). \quad (3)$$

$(A,A)_N$ is a bounded and monotonously increasing sequence and therefore it has a limit $(A,A)=\lim_{N\to\infty}(A,A)_N$. Since $\mathcal{C}_H$ is a linear manifold, $(A\pm B, A\pm B)_N$ and $(A\pm iB, A\pm iB)_N$ converge also, from which we conclude by the polarization formula that $(A,B)_N$ converges to a limit (A,B) as $N\to\infty$. From its definition (A,B) is clearly a scalar product. It is positive definite because $(A,A)=0$ implies $(\varphi_n,A\varphi_m)=0$ for all n and m and hence $A=0$.

The functional $\langle A,B\rangle$ is an extension of the usual trace $(1/Z)\,\mathrm{Tr}A^*B\exp(-\beta H)$ and reduces to it when A and B are bounded. We get the Bogoliubov inequality if we choose B of the form $[C,H]$ for some C. Precisely, we have the following lemma.

Lemma 2: Assume that

((i) A and C belong to $\mathcal{C}_H$,

(ii) C and C^* map $\mathcal{D}(\exp(\beta H))$ into $\mathcal{D}(H)$ for all $\beta>0$.

Then

$$|\langle A,C\rangle-\langle C^*:A^*\rangle|^2 \le\tfrac{1}{2}\beta(\langle[H,C],C\rangle+\langle C^*,[H,C^*]\rangle)(\langle A,A\rangle+\langle A^*,A^*\rangle). \quad (4)$$

If $\varphi\in\mathcal{D}(\exp(\beta H))$, then $\varphi=\exp(-\beta H)\psi$ for some $\psi\in\mathcal{H}$ and $CH\varphi=C\exp(-\beta H/2)H\exp(-\beta H/2)\psi$ is defined, $C\exp(-\beta H/2)$ and $H\exp(-\beta H/2)$ being both bounded. By assumption HC is also defined on $\mathcal{D}(\exp(\beta H))$. Hence $[C,H]$ and $[C^*,H]$ belong to $\mathcal{C}_H$. Setting $B=[C,H]$ in the definition of (A,B) we get immediately $(A,[C,H])$ $=\langle C^*,A^*\rangle-\langle A,C\rangle$. Equation (4) results from the Schwartz inequality $|(A,B)|^2\le(A,A)(B,B)$ and of the bound (3).

We shall now proceed to the verification of the conditions of lemma 2 in the case of interest for the Bose gas.

III. APPLICATION TO THE BOSE GAS

A. The Hamiltonian of the Bose gas

Let us recall the construction of the Hamiltonian of a Bose gas submitted to an external gauge symmetry breaking field.[11,12] We consider first a n-particles system of mass m enclosed in a cubic box Λ of side L in R^d, $d=1,2,3$, described in $\mathcal{H}^n(\Lambda)=(\mathcal{L}^2(\Lambda)^{\otimes n})_{\text{sym.}}$. Its kinetic energy is the dn-dimensional Laplacian

$$H_0^n=\frac{1}{2m}\sum_{j=1}^{n}\underline{\underline{p}}_j^2$$

with periodic boundary conditions. H_0^n is self-adjoint on its natural domain $\mathcal{D}(H_0^n)$:

$$\mathcal{D}(H_0^n)=\left\{\varphi(k_1,\ldots,k_n)\right.$$
$$\left.\times\left|\sum_{k_1,\ldots,k_n}\left(\frac{1}{2m}\sum_{j=1}^{n}k_j^2\right)^2|\varphi(k_1,\ldots,k_2)|^2<\infty\right.\right\}$$

with

$$p_j=\left\{p_j^r=-i\frac{\partial}{\partial x_j^r}, r=1\cdots d\right\},\qquad p_j^2=\sum_{r=1}^{d}(p_j^r)^2,$$
$$k_j=\left\{k_j^r=\frac{2\pi\nu^r}{L},\ r=1\cdots d,\ \nu^r \text{ integers}\right\}.$$

The interaction $U^n(x_1\cdots x_n)$, $x_j=\{x_j^r, r=1\cdots d\}$, satisfies the superstability condition

$$U^n(x_1\cdots x_n)\ge-bn+aV^{-1}n^2 \quad (5)$$

with $a>0$, $b>0$, $V=L^d$, and $\mathcal{D}(H_0^n)\cap\mathcal{D}(U^n)$ is assumed to be dense in $\mathcal{H}^n(\Lambda)$. Let $Q_{H_0^n}+Q_{U^n}$ be the sum of the quadratic forms associated with H_0^n and U^n on $\mathcal{D}(H_0^n)$ $\cap\mathcal{D}(U^n)$. This form is densely defined, bounded below, and hence closable. Its closure Q_{H^n} determines uniquely a self-adjoint operator H^n together with its domain $\mathcal{D}(H^n)$. One has the following relations between the domains of the involed operators and forms:

$$\mathcal{D}(H^n)\subset\mathcal{D}_Q(H^n)=\mathcal{D}_Q(H_0^n)\cap\mathcal{D}_Q(U^n)\supset\mathcal{D}(H_0^n)\cap\mathcal{D}(U^n).$$

H^n is the total energy of the n-particle system.

In order to treat the Bose gas in the grand canonical formalism, we introduce the Fock space $\mathcal{F}(\Lambda)=\sum_{n=0}^{\infty}{}^{\oplus}$ $\times\mathcal{H}^n(\Lambda)$ and on $\mathcal{F}(\Lambda)$ the direct sum Hamiltonian H $=\sum_{n=0}^{\infty}{}^{\oplus}H^n$. H is self-adjoint on its natural domain:

$$\mathcal{D}(H)=\{\phi=\{\phi^n, n=0,1,\cdots\}\in\mathcal{F}(\Lambda)\,|\,\phi^n\in\mathcal{D}(H^n)$$

and $\sum_{n=0}^{\infty} \|H^n\phi^n\|^2 < \infty\}$.

In the same way we define

$$H_\mu = \sum_{n=0}^{\infty} {}^{\oplus} (H^n - \mu n I^n)$$

for every real chemical potential μ on $\mathcal{D}(H_\mu)$:

$$\mathcal{D}(H_\mu) = \{\phi = \{\phi^n, n=0,1,\cdots\} \in \mathcal{F}(\Lambda) \mid \phi^n \in \mathcal{D}(H^n)$$

and $\sum_{n=0}^{\infty} \|(H^n - \mu n)\phi^n\|^2 < \infty\}$.

We note that

(a) $\mathcal{D}(H_\mu) \subset \mathcal{D}(N^2) \subset \mathcal{D}(N)$ for all μ where N is the number of particles operator and

(b) $\mathcal{D}(H_\mu) = \mathcal{D}(H)$ is independent of μ.

(a) follows from the superstability condition (5):

$$(\phi^n, (H^n - \mu n)\phi^n) \geq (-(\mu + b)n + aV^{-1}n^2)\|\phi^n\|^2, \phi^n \in \mathcal{D}(H^n). \tag{6}$$

For fixed μ and V, and for n large enough, the right-hand side of (6) is positive. One can find a number $C > 0$ (depending on μ and V) such that $(\phi^n, (H^n - \mu n)\phi^n) \geq Cn^2\|\phi^n\|^2$ and also $\|(H^n - \mu n)\phi^n\|^2 \geq C^2 n^4 \|\phi^n\|^2$. Therefore if ϕ belongs to $\mathcal{D}(H_\mu)$, we see that $n^4\|\phi^n\|^2$ is summable, which shows that ϕ belongs also to $\mathcal{D}(N^2) \subset \mathcal{D}(N)$. (b) is deduced of the identity

$$\|(H^n - \mu n)\phi^n\|^2 = \|H^n\phi^n\|^2 - 2\mu n(\phi^n, H^n\phi^n) + n^2\|\phi^n\|^2,$$

the summability of one of the series $\sum_n \|H^n\phi^n\|^2$ or $\sum_n \|(H^n - \mu n)\phi^n\|^2$ implying (a) and the summability of the other.

In order to get a nonvanishing order parameter at finite volume, it is necessary to add to H_μ a gauge symmetry breaking term (Bogoliubov's method of quasi-averages). We choose it of the form $\lambda\sqrt{V}(a_0 + a_0^*)$, where a_0 is the annihilation operator on $\mathcal{F}(\Lambda)$ of a particle in the $k=0$ momentum state. The $\sqrt{V}$ factor is needed for reasons of extensivity. We show that for every V, $\lambda\sqrt{V}(a_0 + a_0^*)$ is relatively bounded with respect to H_μ with relative bound less that one.

We remark first that by (a), a_0 and a_0^* are defined on $\mathcal{D}(H)$. We have to establish that

$$\|\lambda\sqrt{V}(a_0 + a_0^*)\phi\| \leq b\|H_\mu\phi\| + a\|\phi\|, \phi \in \mathcal{D}(H) \tag{7}$$

with $b < 1$ for each V. If $\phi \in \mathcal{D}(H)$ and z is an arbitrary complex number with $\mathrm{Im} z \neq 0$, we can write $\phi = (H_\mu - z)^{-1}\psi$ for some $\psi \in \mathcal{F}(\Lambda)$. Then

$$\begin{aligned}\|\lambda\sqrt{V}a_0\phi\| &= \|\lambda\sqrt{V}a_0(N+1)^{-1/2}(N+1)^{1/2}(H_\mu - z)^{-1}\psi\| \\ &\leq |\lambda|\sqrt{V}\,\|a_0(N+1)^{-1/2}\|\,\|(N+1)^{1/2}(H_\mu - z)^{-1}\| \\ &\quad \times (\|H_\mu\phi\| + |z|\,\|\phi\|).\end{aligned} \tag{8}$$

The norm $\|(N+1)^{1/2}(H_\mu - z)^{-1}\|$ is readily evaluated in the spectral representation of H_μ:

$$\begin{aligned}\|(N+1)^{1/2}(H_\mu - z)^{-1}\| &= \sup_n \sup_\epsilon \left|\frac{(n+1)^{1/2}}{\epsilon - \mu n - z}\right| \\ &= \sup_n \sup_\epsilon \left(\frac{(n+1)^{1/2}}{[(\epsilon - \mu n - \mathrm{Re} z)^2 + (\mathrm{Im} z)^2]^{1/2}}\right),\end{aligned} \tag{9}$$

where ϵ runs for each n on the spectrum of H^n. One deduces of (6) that for large n the spectrum of $H^n - \mu n$ is bounded below by $\epsilon - \mu n \geq Cn^2$, C depending only on μ and V. Consequently, for fixed μ and V, the quantity $|(n+1)^{1/2}/(\epsilon - \mu n - z)|$ can be made as small as we wish uniformly in ϵ and n by choosing $\mathrm{Im} z$ large enough. In view of (8), (9) and of a similar estimate for $\lambda\sqrt{V}a_0^*$ we see that (7) holds. Therefore, $H_{\mu\lambda} = H - \mu N + \lambda\sqrt{V}(a_0 + a_0^*)$ is self-adjoint on $\mathcal{D}(H_{\mu\lambda}) = \mathcal{D}(H)$ for all μ, λ, and V. This concludes the description of the full Hamiltonian $H_{\mu\lambda}$ and of its domain.

We do not repeat here the proof that $\exp(-\beta H_\mu)$ belongs to $\mathcal{L}_1(\mathcal{F}(\Lambda))$ for all μ.[11,12] The fact that $\exp(-\beta H_{\mu\lambda})$ is also of trace class for all μ and λ follows from the inequality

$$H_{\mu+|\lambda|} - |\lambda| V \leq H_{\mu\lambda} \leq H_{\mu-|\lambda|} + |\lambda| V. \tag{10}$$

B. Absence of condensation

The absence of condensation in a one- or two-dimensional Bose gas is then understood in the following sense. The average value of the order parameter $V^{-1/2}a_0$ is calculated with the gauge symmetry breaking field in the thermodynamic limit which is taken first. Then the field is removed and the order parameter is shown to vanish. This procedure motivates the following choice of the operators A and C on $\mathcal{F}(\Lambda)$ which will enter the Bogoliubov inequality: $A = V^{-1/2}a_k$ is the annihilation operator of a particle in momentum state k and $C = \sum_{n=0}^{\infty} \oplus C^n$ with

$$(C^n\varphi)(x_1, \ldots, x_n) = \left(\sum_{j=1}^{n} \exp(ikx_j)\right)\varphi(x_1, \ldots, x_n),$$

$$\varphi(x_1, \ldots, x_n) \in \mathcal{H}^n(\Lambda) \text{ and } k = \{2\pi\nu^r/L,\ r = 1 \ldots d, \nu_r \text{ integers}\}.$$

Clearly when $k=0$, A is precisely the order parameter and C reduces to the generator N of the gauge group.

Let us verify that A and C fulfill the conditions (i) and (ii) of Lemma 2 with respect to the Hamiltonian $H_{\mu\lambda}$. Since A, A^*, C and C^* are defined on $\mathcal{D}(N) \supset \mathcal{D}(H_{\mu\lambda}) \supset \mathcal{D}(\exp(\beta H_{\mu\lambda}))$, (i) holds true. For (ii) we proceed with the following steps. (α) C^n maps the form domain $\mathcal{D}_Q(H^n)$ into itself. Since by construction $\mathcal{D}_Q(H^n) = D_Q(H_0^n) \cap \mathcal{D}_Q(U^n)$, it is sufficient to show that C^n leaves $\mathcal{D}_Q(H_0^n)$ and $\mathcal{D}_Q(U^n)$ invariant. For $\mathcal{D}_Q(U^n)$ it is an obvious fact: C^n is defined by the multiplication by a bounded function of the coordinates. Now $\varphi \in \mathcal{D}_Q(H_0^n)$ if and only if $\varphi \in \mathcal{D}(p_j^r)$ for all $j = 1 \ldots n$, $r = 1 \ldots d$, and C^n leaves the $\mathcal{D}(p_j^r)$ invariant with

$$(p_j^r C^n\varphi)(x_1 \cdots x_n) = k^r \exp(ikx_j)\varphi(x_1 \cdots x_n) + (C^n p_j^r\varphi)(x_1 \cdots x_n)$$

[remember that $\exp(ikx_j)$, $k = \{2\pi\nu^r/L\}$, is periodic]. (β) $[C^n, H_0^n]$ is defined on $\mathcal{D}_Q(H^n)$. We have indeed on $\mathcal{D}_Q(H_0^n) \supset \mathcal{D}_Q(H^n)$

$$\begin{aligned}([C^n, H_0^n]\varphi)(x_1 \cdots x_n) &= \sum_{r=1}^{d}\sum_{j=1}^{n} k^r \exp(ikx_j)(p_j\varphi)(x_1 \cdots x_n) \\ &\quad + k^2(C^n\varphi)(x_1 \cdots x_n).\end{aligned} \tag{11}$$

Using the inequality $\sum_{mn}^q M_{mn} \leq q\sum_n^q M_{nn}$ valid for and q-dimensional positive matrix M and the superstability condition (6), we can majorize the square of the norm of the first term in the right-hand side of (11) by

$$\sum_{r,r'}^{d}\sum_{j,j}^{n} k^r k^{r'}(\exp(ikx_{j'})p_{j'}^{r'}\varphi, \exp(ikx_j)p_j^r\varphi)$$

$$\leq dn\sum_{r}^{d}\sum_{j}^{n}(k^r)^2(p_j^r\varphi, p_j^r\varphi) \leq 2mk^2 n\, dQ_{H_0}(\varphi,\varphi)$$

$$\leq 2m\, dk^2(nQ_{H^n}(\varphi,\varphi) + bn^2(\varphi,\varphi)),\ \varphi\in D_Q(H^n). \qquad (12)$$

(γ)C^n maps $\mathcal{D}(H^n)$ into itself. If $\psi\in\mathcal{D}(H_0^n)\cap\mathcal{D}(U^n)$ and $\varphi\in\mathcal{D}(H^n)$, we are allowed to write in view of (α) and (β)

$$Q_{H^n}(C^n\varphi,\psi) = (C^n\varphi, H^n\psi) = (\varphi, (C^n)^* H^n\psi)$$

$$= (\varphi, (H^n C^* + [(C^n)^*, H_0^n])\psi)$$

$$= ((C^n H^n + [H_0^n, C^n])\varphi, \psi). \qquad (13)$$

Since $\mathcal{D}(H_0^n)\cap\mathcal{D}(U^n)$ is a core of Q_{H^n} and by the very definition of $\mathcal{D}(H^n)$, (13) implies that $C^n\varphi\in\mathcal{D}(H^n)$ and

$$H^n C^n\varphi = C^n H^n\varphi + [H_0^n, C^n]\varphi. \qquad (14)$$

(δ) (ii) of lemma 2 is true. We remark first that $C(N+1)^{-1}$ maps $\mathcal{D}(H)$ into itself. If $\phi=\{\phi^n, n=1,2\cdots\}\in\mathcal{D}(H)$, then by (14)

$$(HC(N+1)^{-1}\phi)^n = (n+1)^{-1}C^n H^n\phi^n + (n+1)^{-1}[H_0^n, C^n]\phi^n,$$

$$\sum_{n=0}^{\infty}\left(\frac{1}{n+1}\right)^2 \|C^n H^n\phi^n\|^2 \leq \sum_{n=0}^{\infty}\|H^n\phi^n\|^2 < \infty,$$

and in virtue of (11) and (12)

$$\sum_{n=0}^{\infty}\left(\frac{1}{n+1}\right)^2 \|[H_0^n, C^n]\phi^n\|^2$$

is also finite. Now for every $\Psi\in\mathcal{F}(\Lambda)$, $N\exp(-\beta H_{\mu\lambda})\Psi$ belongs to $\mathcal{D}(H)$ since we can write

$$HN\exp(-\beta H_{\mu\lambda}) = NH_{\mu\lambda}\exp(-\beta H_{\mu\lambda})$$

$$+ N(\mu N - \lambda\sqrt{V}(a_0 + a_0^*))\exp(-\beta H_{\mu\lambda}),$$

the first term of the right-hand side being obviously defined everywhere as well as the second one in virtue of (a) $\mathcal{D}(H_{\mu\lambda}) = \mathcal{D}(H)\subset\mathcal{D}(N^2)$. Therefore we conclude that $C\exp(-\beta H_{\mu\lambda})\Psi = C(N+1)^{-1}(N+1)\exp(-\beta H_{\mu\lambda})\Psi$ belongs to $\mathcal{D}(H) = D(H_{\mu\lambda})$ and this is precisely (ii) of lemma 2.

We obtain a proof of the absence of condensation if we insert A and C in the Bogoliubov inequality in its form (4). After a direct calculation which we do not reproduce here[1] one finds for each $k=\{2\pi\nu^r/L, \nu^r$ integers$\}$

$$\frac{|\langle a_0\rangle|^2}{V} \leq \left(\frac{k^2}{m}\langle N\rangle + \lambda\sqrt{V}\langle a_0 + a_0^*\rangle\right)\beta\left(\frac{\langle a_k^* a_k\rangle}{V} + \frac{1}{2V}\right). \qquad (15)$$

Here the averages $\langle A\rangle = \mathrm{Tr}A\exp(-\beta H_{\mu\lambda})/\mathrm{Tr}\exp(-\beta H_{\mu\lambda})$ are well defined in virtue of the preceding analysis and they depend on the parameters β, V, μ, and λ. Using $\sqrt{V}(a_0 + a_0^*) \leq N + V$ and summing (15) on k up to $|k|\leq k_0 < \infty$, one gets

$$\frac{1}{V}\sum_{|k|\leq k_0}\left(\frac{1}{k^2/m + |\lambda|}\right)\frac{|\langle a_0\rangle|^2}{V}$$

$$\leq \beta\left(\frac{\langle N\rangle}{V} + 1\right)\frac{1}{V}\sum_{|k|\leq k_0}(\langle a_k^* a_k\rangle + \tfrac{1}{2})$$

$$\leq \beta\left(\frac{\langle N\rangle}{V} + 1\right)\left(\frac{\langle N\rangle}{V} + \sum_{|k|\leq k_0}\frac{1}{2V}\right). \qquad (16)$$

Furthermore for fixed μ and β the density $\langle N\rangle/V$ is bounded uniformly with respect to $|\lambda|\leq\lambda_0$, $V\geq V_0$ with $\lambda_0<\infty$, $V_0>0$. This follows from the following facts (lemma 2 of Ref. 9):

(1) The pressure $p(\lambda,\mu) = (1/\beta V)\lg\mathrm{Tr}\exp(-\beta H_{\mu\lambda})$ is bounded by a positive function of the form

$p(\lambda,\mu)\leq|\lambda| + \hat{p}(|\lambda| + \mu)$, $\hat{p}(\mu)$ independent of V.

(2) $p(\lambda,\mu)$ and $\hat{p}(\mu)$ are increasing and convex in μ. Hence, choosing a $\delta>0$, we have

$$\frac{\langle N\rangle}{V} = \frac{\partial}{\partial\mu}p(\lambda,\mu) \leq \frac{p(\lambda,\mu+\delta) - p(\lambda,\mu)}{\delta} \leq \frac{\lambda_0 + \hat{p}(\lambda_0 + \mu + \delta)}{\delta}.$$

With this and (16), we see that there exists a M independent of λ and V (but depending on β and μ) such that:

$$\lim_{V\to\infty}\frac{|\langle a_0\rangle|^2}{V} \leq M\left(\int_0^{k_0}\frac{|k|^{d-1}}{k^2/m + |\lambda|}\, d|k|\right)^{-1}.$$

Since the integral in the right-hand side tends to infinity as $\lambda\to 0$ for $d\leq 2$ we conclude that the order parameter vanishes:

$$\lim_{\lambda\to 0}\lim_{V\to\infty}\frac{\langle a_0\rangle}{\sqrt{V}} = 0$$

for all $\beta>0$ and all chemical potentials μ.

ACKNOWLEDGMENTS

We thank Dr. H. Kunz for suggesting to us to carry out this analysis and for discussions.

[1]P.C. Hohenberg, Phys. Rev. **158**, 383 (1967).
[2]N.N. Bogoliubov, Phys. Abh. S.U. 6, **1**, 229 (1962).
[3]N.D. Mermin and H. Wagner, Phys. Rev. Lett. **17**, 1133 (1966).
[4]D. Ruelle, *Statistical Mechanics* (Benjamin, New York, 1969).
[5]N.D. Mermin, J. Math. Phys. **8**, 1061 (1967).
[6]P.A. Vuillermot and M.V. Romerio, Commun. Math. Phys. **41**, 281 (1975).
[7]J.C. Garrison and J. Wong, Commun. Math. Phys. **26**, 1 (1972).
[8]J. Naudts, A. Verbeure, and R. Weder, Commun. Math. Phys. **44**, 87 (1975).
[9]J. Ginibre, Commun. Math. Phys. **8**, 26 (1968).
[10]T. Kato, *Perturbation Theory for Linear Operators* (Springer, New York, 1966).
[11]D. Ruelle, Helv. Phys. Acta **46**, 789 (1963).
[12]D.W. Robinson, in *The Thermodynamic Pressure in Quantum Statistical Mechanics*, *Lecture Notes in Physics* 9 (Springer, New York, 1971).

Feynman path integrals and quantum mechanics as $\hbar \to 0$

Aubrey Truman

Mathematics Department, Heriot-Watt University, Edinburgh, Scotland
(Received 30 December 1975)

When the space of paths is a certain Hilbert space H, we show how to extend the Feynman path integral $\mathcal{J}$ of DeWitt and Albeverio and Hoegh-Krohn. Our extension enables us to integrate a wider class of functionals on H. We establish a new representation for the wavefunction in nonrelativistic quantum mechanics—the quasiclassical representation. Using our extension of $\mathcal{J}$ and the quasiclassical representation, we discuss the problem of obtaining classical mechanics as the limiting case of quantum mechanics when $\hbar \to 0$.

1. INTRODUCTION

There is a saying attributed to Einstein that "mathematics is an exact science until one tries to apply it to the physical world." Nowhere has this statement been more valid than in the many attempts to give a rigorous formulation of the Feynman path integral $\mathcal{J}$ in nonrelativistic quantum mechanics. That is at least until recently when a number of papers appeared which give a workable definition of the Feynman path integral $\mathcal{J}$ for a restricted class of functionals.

Amongst the first of these papers were those of Cameron and Itô[1] giving a precise definition of $\mathcal{J}$ by various limiting procedures. The first definition of $\mathcal{J}$ not involving a limiting procedure was formulated by C. DeWitt[2] by invoking some of the ideas of Bourbaki on "promeasures" and generalizing them to "prodistributions." C. DeWitt's far reaching formulation was later elaborated upon by Albeverio and Hoegh-Krohn.[3] For spinless nonrelativistic quantum mechanical particles Albeverio and Hoegh-Krohn choose the path space to be a certain Hilbert space H. Using basic results from harmonic analysis on H, Albeverio and Hoegh-Krohn define $\mathcal{J}$ without recourse to "prodistributions." Unfortunately the class of integrable functionals on H in Albeverio and Hoegh-Krohn's treatment is rather restricted and the applications of the theory to nonrelativistic quantum mechanics although elegant are rather limited. It is clear that if the theory is to have greater applicability a new definition of $\mathcal{J}$ is required.

In this paper we give a concrete realization of the path space H and by considering a certain projection P_n on H we extend the definition of $\mathcal{J}$ to a wider class of functionals. We believe that for nonrelativistic quantum mechanics the extension of $\mathcal{J}$ given here is in some real sense maximal. As an application of our extension of $\mathcal{J}$ we discuss the celebrated problem of obtaining classical mechanics as the limiting case of quantum mechanics when $\hbar \to 0$. To this end we establish a new representation for the wavefunction solution of the Schrödinger equation—the quasiclassical representation.

To make our paper as self-contained as possible, we give a very brief outline of DeWitt's and Albeverio and Hoegh-Krohn's work in deriving the Feynman—Itô formula for the wavefunction. This Feynman—Itô formula is extended to give the quasiclassical representation. As we shall see our extension of $\mathcal{J}$ and the quasiclassical representation are very convenient in the discussion of the problem of obtaining classical mechanics from quantum mechanics. The problem of obtaining classical mechanics from quantum mechanics as $\hbar \to 0$ has been discussed by several authors. The most notable results are due to Hepp and Maslov, who use a completely different approach.[4]

The basic question posed by Feynman's work[5] is how to define the Feynman integral, $\mathcal{J}(f)$, written symbolically

$$\mathcal{J}(f) = \int_{x\in X} f(x)\, dw(x),$$

where f is a functional defined on the path space X and w is the "pseudomeasure" corresponding to the Feynman integral. Cameron[6] has shown that w is a poor additive set function. DeWitt defines w, therefore, not as an additive set function but as a distribution of rank zero. The "pseudomeasure" w is, in fact, determined by defining its Fourier transform to be $\exp(-i/2)W$, where W is a positive definite quadratic form on X' the dual of X. W is called the variance of w.

DeWitt's definition of $\mathcal{J}$ using "prodistributions" assumes that the path space X is merely a locally convex Hausdorf topological vector space. In the cases of interest to us the path space X will, in addition, be a separable and reflexive Banach space and we shall not require the machinery of "prodistributions." Using the Hahn—Banach theorem, the separability and reflexivity of X imply that X', the dual of X, is separable in the norm topology. We choose as a convenient σ field on X' the σ field generated by the subsets of X' open in the norm topology. Denoting the action of the functional x' on x by $\langle x', x\rangle$, for each $x \in X$, $\exp(-i\langle x', x\rangle)$ is continuous in x' in the norm topology and is therefore Borel measurable relative to our σ field. Let $f \in \mathcal{J}(X)$ be a functional defined on X such that

$$f(x) = \int \exp(-i\langle x', x\rangle)\, d\mu(x'),$$

where μ is a bounded complex measure on X'. Then, assuming the variance W is such that $\exp[-(i/2)W(x', x')]$ is measurable, the Feynman integral with variance W is defined by

$$\mathcal{J}(f) = \int \exp[-(i/2)W(x', x')]\, d\mu(x'),$$

whenever this integral exists.

We shall see that for a spinless nonrelativistic quantum mechanical particle we can choose $X = H$, a certain reflexive Hilbert space of paths. In this case the variance W_H appropriate for H turns out to be the Hilbert space inner product. The Borel sets in H are given by the σ field generated by the subsets of H open in the inner product $\| \ \|$ topology and trivially $\exp[-(i/2)W(x', x')] = \exp[-(i/2)\|x'\|^2]$ is Borel measurable relative to this σ field. In Sec. 2 we shall give the detailed structure of H by using elementary Fourier analysis. This detailed structure of H is used in developing our definition of the Feynmann path integral $\mathcal{J}$.

In the following sections we discuss DeWitt's and Albeverio and Hoegh–Krohn's definition of the Feynman integral $\mathcal{J}$ for the path space H, concluding Sec. 5 by deriving the Feynman–Itô formula. In Sec. 6, using the fact that H has a reproducing kernel, we extend the Feynman integral $\mathcal{J}$ for the path space H and calculate an important Feynman integral. In Sec. 7 we derive our quasiclassical representation by using the translational properties of $\mathcal{J}$. Finally in Sec. 8 we discuss the problem of obtaining classical mechanics as the limiting case of quantum mechanics when $\hbar \to 0$. This treatment is carried out in one dimension for ease of understanding, but it is clear that all the essential ideas are easily generalized to higher dimension than one.

2. THE PATH SPACE $X = H$

We give here a concrete realization of H the Hilbert space of paths for a nonrelativistic quantum mechanical particle in one dimension. Albeverio and Hoegh-Krohn define H to be the space of continuous functions $\gamma(\tau)$ defined on $[0, t]$ normalized so that $\gamma(t) = 0$ with $d\gamma/d\tau \in L^2\{\mathcal{R}, [0, t]\}$ and with norm given by the inner product

$$(\gamma, \gamma') = \int_0^t \frac{d\gamma}{d\tau}\frac{d\gamma'}{d\tau}\, d\tau.$$

There are problems with this definition in the interpretation of $d\gamma/d\tau$. The difficulty is that there are examples well known to analysts of monotonic continuous functions whose a. e. derivatives vanish. Hence, if we interpret $d\gamma/d\tau$ to be the a. e. derivative of γ, there are numerous pathological functions with norm zero. To form a sensible path space H with the above inner product we would have to factor out these pathological functions by forming a quotient space. The alternative is to interpret $d\gamma/d\tau$ to be the weak derivative of γ which is the viewpoint adopted here. This leads us to define H by Fourier series. The separability and reproducing kernel properties of H are then easily established.

Consider V, the real vector space of continuous functions $\gamma(\tau)$ on $[0, t]$, satisfying $\gamma(t) = 0$, defined by

$$\gamma(\tau) = \alpha_0(\tau - t) - \sum_{n=1}^{\infty} \frac{\alpha_n t}{2\pi n}\left[\sin\left(\frac{2\pi n\tau'}{t}\right)\right]_\tau^t + \sum_{n=1}^{\infty} \frac{\beta_n t}{2\pi n}\left[\cos\left(\frac{2\pi n\tau'}{t}\right)\right]_\tau^t, \qquad (1a)$$

where α_0, α_n, $\beta_n \in \mathcal{R}$, $\sum_1^\infty (\alpha_n^2 + \beta_n^2) < \infty$, and the dummy variable in the square brackets is τ'. Since we can always integrate a Fourier series,

$$\gamma(\tau) = -\int_\tau^t \frac{d\gamma_\omega}{d\tau'}(\tau')\, d\tau',$$

where

$$\frac{d\gamma_\omega}{d\tau}(\tau) = \alpha_0 + \sum_{n=1}^{\infty} \alpha_n \cos\frac{2\pi n\tau}{t} + \sum_{n=1}^{\infty} \beta_n \sin\frac{2\pi n\tau}{t}. \qquad (1b)$$

$(d\gamma_\omega/d\tau)(\tau) \in L^2[0, t]$ and the notation is derived from the fact that $d\gamma_\omega/d\tau$ is the weak derivative of γ, i. e., if we consider γ as a generalized function its derivative is $d\gamma_\omega/d\tau$.[7]

We define an inner product (,) on V by

$$(\gamma, \gamma') = \int_0^t \frac{d\gamma_\omega}{d\tau}\frac{d\gamma'_\omega}{d\tau}\, d\tau = t\alpha_0\alpha_0' + \frac{t}{2}\sum_1^\infty \alpha_n\alpha_n' + \frac{t}{2}\sum_1^\infty \beta_n\beta_n'. \qquad (2)$$

V is evidently a real separable Hilbert space H in the inner product norm $\|\gamma\|^2 = (\gamma, \gamma) = \int_0^t (d\gamma_\omega/d\tau)^2\, d\tau$. H is the Hilbert space of paths on which we shall define the "pseudomeasure" w.

When $d\gamma_\omega/d\tau$ is sufficiently smooth we deduce $d\gamma/d\tau = d\gamma_\omega/d\tau$. One case when this is true is

$$\frac{dG_\omega}{d\tau}(\sigma, \tau) = -\Theta(\tau - \sigma) \quad \text{a. e.} \quad [0, t],$$

Θ being the Heaviside function,

$$\frac{dG_\omega}{d\tau}(\sigma, \tau) = \frac{(\sigma - t)}{t} - \sum_1^\infty \frac{1}{\pi n}\left[\sin\left(\frac{2\pi n\tau'}{t}\right)\right]_\sigma^t \cos\left(\frac{2\pi n\tau}{t}\right) + \sum_1^\infty \frac{1}{\pi n}\left[\cos\left(\frac{2\pi n\tau'}{t}\right)\right]_\sigma^t \sin\left(\frac{2\pi n\tau}{t}\right). \qquad (3a)$$

Then

$$G(\sigma, \tau) = -\int_\tau^t \frac{dG_\omega}{d\tau'}(\sigma, \tau')\, d\tau' = t - \sigma^\vee\tau,$$

where $^\vee$ is the maximum.

We have $\forall\ \tau \in [0, t]$,

$$G(\sigma, \tau) = \left(\frac{\sigma - t}{t}\right)(\tau - t) + \sum_1^\infty \frac{t}{2\pi^2 n^2}\left[\sin\left(\frac{2\pi n\tau'}{t}\right)\right]_\sigma^t \times \left[\sin\left(\frac{2\pi n\tau'}{t}\right)\right]_\tau^t + \sum_1^\infty \frac{t}{2\pi^2 n^2}\left[\cos\left(\frac{2\pi n\tau'}{t}\right)\right]_\sigma^t \times \left[\cos\left(\frac{2\pi n\tau}{t}\right)\right]_\tau^t. \qquad (3b)$$

$G(\sigma, \tau)$ is the reproducing kernel for the Hilbert space H. For we have

$$(\gamma(\tau), G(\sigma, \tau)) = \int_0^t \frac{d\gamma_\omega}{d\tau}(\tau)\frac{dG_\omega}{d\tau}(\sigma, \tau)\, d\tau = -\int_\sigma^t \frac{d\gamma_\omega}{d\tau}\, d\tau = \gamma(\sigma), \quad \forall\ \gamma \in H. \qquad (3c)$$

This fact will be of crucial importance in what follows.

3. THE FEYNMAN PATH INTEGRAL WHEN $X = H$

In Sec. 5 we shall see that the appropriate variance W_H for a nonrelativistic quantum mechanical particle is given by the inner product

$$W_H(\gamma_1, \gamma_2) = (\gamma_1, \gamma_2). \qquad (4)$$

Definition 1: $\mathcal{M}(H)$ is the space of complex valued measures of bounded absolute variation on H, $\mu \in \mathcal{M}(H)$ iff $\|\mu\| = \int |d\mu| < \infty$. $\| \ \|$ is a norm on $\mathcal{M}(H)$.

Definition 2: The space of functionals $\mathcal{J}(H)$ is defined by $f\in \mathcal{F}(H)$ iff $f(\gamma)=\int \exp[-i(\gamma',\gamma)]\,d\mu(\gamma')$, $\mu\in\mathcal{M}(H)$.[8]

Definition 3: When $f\in\mathcal{F}(H)$,

$$f(\gamma)=\int \exp[-i(\gamma',\gamma)]\,d\mu(\gamma'), \tag{5}$$

$\mu\in\mathcal{M}(H)$, taking $W_H(\gamma_1,\gamma_2)=(\gamma_1,\gamma_2)$, the Feynman path integral $\mathcal{F}(f)$ is defined by

$$\mathcal{F}(f)=\int \exp\left[-\frac{i}{2}(\gamma',\gamma')\right]d\mu(\gamma'). \tag{6}$$

$\mu\in\mathcal{M}(H)$ ensures $\mathcal{F}(f)$ exists. For the continuous function $\exp[(-i/2)(\gamma',\gamma')]$ is Borel measurable and

$$|\mathcal{F}(f)|\leqslant\int|d\mu(\gamma')|=\|\mu\|\overset{\text{def}}{=}\|f\|_0<\infty. \tag{7}$$

It is not difficult to establish $\|\ \|_0$ is a norm on $\mathcal{F}(H)$. The properties of $\mathcal{F}(H)$ and $\mathcal{M}(H)$ are summarized in Theorems 1 and 2.

Definition 4: The convolution, $\mu*\nu$, of $\mu,\nu\in\mathcal{M}(H)$ is defined by

$$(\mu*\nu)(A)=\int \mu(A-\gamma)\,d\nu(\gamma), \tag{8}$$

for any Borel set $A\subset H$. $(\mu*\nu)$ is well defined because of Fubini's theorem.

Theorem 1: $\mathcal{M}(H)$ is a commutative Banach algebra in absolute variation norm under $*$. (H is a separable metric group under vector addition; it follows that $\mathcal{M}(H)$ is a commutative topological semigroup under $*$—see Parthasaraty.[9])

Proof: By Fubini's theorem for any bounded continuous functional f

$$\int f(\gamma)d(\mu*\nu)(\gamma)=\int f(\gamma+\gamma')\,d\mu(\gamma)\,d\nu(\gamma'). \tag{9}$$

Hence, we have

$$\mu*\nu=\nu*\mu,\quad \|\mu*\nu\|\leqslant\|\mu\|\,\|\nu\|. \tag{10}$$

The associativity of $*$ follows from Fubini's theorem. Completeness follows by standard arguments.

Theorem 2: $\mathcal{F}(H)$ is a Banach function algebra in $\|\ \|_0$. $\mathcal{F}(H)$ is closed under addition, multiplication, and composition with entire functions.

Proof: Putting $f(\gamma)=\exp[-i(\gamma,\delta)]$ in (9), we obtain

$$\int \exp[-i(\gamma,\delta)]d(\mu*\nu)(\gamma)$$
$$=\int \exp[-i(\gamma,\delta)\,d\mu(\gamma)\int \exp[-i(\gamma',\delta)]\,d\nu(\gamma'). \tag{11}$$

If the entire function $E(z)=\sum_{n=1}^{\infty}a_nz^n$ then $E(f)=\sum_{n=0}^{\infty}a_nf^n$ is the Fourier transform of $\sum_{n=0}^{\infty}a_n(\mu*\mu*\cdots*\mu)\in\mathcal{M}(H)$ if f is the Fourier transform of $\mu\in\mathcal{M}(H)$. This follows from Eq. (11) and

$$\left\|\sum_{n=0}^{\infty}a_n(\mu*\mu*\cdots*\mu)\right\|\leqslant\sum_{n=0}^{\infty}|a_n|\,\|\mu\|^n<\infty. \tag{12}$$

In the next section we see how the above definition of the Feynman integral $\mathcal{F}(f)$ enables us to evaluate $\mathcal{F}(f)$ for certain functionals f. The examples f that we choose illustrate simultaneously the power and the limitations of the definition of $\mathcal{F}$.

4. THE EVALUATION OF SOME FEYNMAN INTEGRALS

Consider first the Green's function $G(\sigma,\tau)$ of the operator $-d^2/d\tau^2$,

$$-\frac{d^2}{d\tau^2}G(\sigma,\tau)=\delta(\sigma-\tau), \tag{13}$$

$(dG/d\tau)(\sigma,0)=0$, $G(\sigma,t)=0$. Then $G(\sigma,\tau)=t-\sigma^{\vee}\tau$, $\sigma^{\vee}\tau=\max(\sigma,\tau)$. We have seen $G\in H$ and for any $\gamma\in H$

$$\gamma(\sigma)=(G(\sigma,\tau),\gamma(\tau)),\quad \sigma\in[0,t]. \tag{14}$$

Thus, H is a reproducing kernel Hilbert space. This is the key to the evaluation of $\mathcal{F}(f)$ in the examples below.

Example 1:

$$f[\gamma]=\exp[-i\alpha\gamma(\sigma)],\quad \alpha\in\mathcal{R}^1,\quad \sigma\in[0,t].$$

To evaluate $\mathcal{F}(f)$ we observe that

$$f[\gamma]=\int \exp[-i(\gamma',\gamma)]\,d\mu_H(\gamma'), \tag{15}$$

where $\mu_H\in\mathcal{M}(H)$ and for Borel $B\subset H$

$$\mu_H(B)=\begin{cases}1, & \text{if } \alpha G(\sigma,\tau)\in B,\\ 0, & \text{otherwise.}\end{cases} \tag{16}$$

It follows from Eq. (6) that

$$\begin{aligned}\mathcal{F}(f)&=\int \exp[-\tfrac{1}{2}i(\gamma',\gamma')]\,d\mu_H(\gamma')\\ &=\exp[-\tfrac{1}{2}i(\alpha G(\sigma,\tau),\alpha G(\sigma,\tau))]\\ &=\exp[-\tfrac{1}{2}i\alpha^2G(\sigma,\sigma)]=\exp[-\tfrac{1}{2}i\alpha^2(t-\sigma)].\end{aligned} \tag{17}$$

Example 2: $f[\gamma]=\int_{\mathcal{R}^1}\exp[-i\alpha\gamma(\sigma)]\,d\mu(\alpha)$,

$\int_{\mathcal{R}^1}|d\mu(\alpha)|<\infty$, $\sigma\in[0,t]$.

We have

$$f[\gamma]=\int \exp[-i(\gamma',\gamma)]\,d\mu_H(\gamma'), \tag{18}$$

where $\mu_H\in\mathcal{M}(H)$ and for Borel $B\subset H$

$$\mu_H(B)=\begin{cases}\mu(A), & \text{if } B\supset G(\sigma,\tau)A',\ \text{Borel } A'\subset\mathcal{R}^1, A=\cup A',\\ 0, & \text{otherwise.}\end{cases} \tag{19}$$

Then we have from Eq. (6)

$$\begin{aligned}\mathcal{F}(f)&=\int \exp[-\tfrac{1}{2}i(\gamma',\gamma')]\,d\mu_H(\gamma')\\ &=\int_{\alpha\in\mathcal{R}^1}\exp[-\tfrac{1}{2}i\alpha^2(t-\sigma)]\,d\mu(\alpha).\end{aligned} \tag{20}$$

Example 3:

$$f[\gamma]=f[\gamma(\sigma_0),\gamma(\sigma_1),\ldots,\gamma(\sigma_{n-1})],\quad 0=\sigma_0<\sigma_1<\sigma_2\cdots<\sigma_{n-1}<t,$$

and

$$f[\gamma]=\int_{\alpha\in\mathcal{R}^n}\exp\Big[-i\sum_{j=0}^{n-1}\alpha_j\gamma(\sigma_j)\Big]\,d\mu(\alpha),\quad \int_{\alpha\in\mathcal{R}^n}|d\mu(\alpha)|<\infty.$$

Proceeding as above

$$f[\gamma]=\int \exp[-i(\gamma',\gamma)]\,d\mu_H(\gamma'), \tag{21}$$

where $\mu_H\in\mathcal{M}(H)$ and for Borel $B\subset H$

$$\mu_H(B)=\begin{cases}\mu(A), & \text{if } B\supset\left\{\sum_{j=0}^{n-1}\alpha_jG(\sigma_j,\tau)\colon(\alpha_0,\ldots,\alpha_{n-1})\in\text{Borel } A'\subset\mathcal{R}^n\right\}\text{ and } A=\cup A',\\ 0, & \text{otherwise.}\end{cases} \tag{22}$$

Then from Eq. (6)

$$\begin{aligned}\mathcal{F}(f)&=\int_{\alpha\in\mathcal{R}^n}\exp\left(-\tfrac{1}{2}i\sum_{j,k=0}^{n-1}\alpha_j\alpha_kG(\sigma_j,\sigma_k)\right)d\mu(\alpha)\\ &=\int_{\alpha\in\mathcal{R}^n}\exp\left(-\tfrac{1}{2}i\sum_{j,k=0}^{n-1}\alpha_j\alpha_k(t-\sigma_j{}^{\vee}\sigma_k)\right)d\mu(\alpha).\end{aligned} \tag{23}$$

Example 4:

$$f[\gamma]=\int_0^t\cdots\int_0^t d\sigma_1\cdots d\sigma_n\int_{\alpha\in\mathcal{R}^{n+1}}\exp\left[-i\sum_{j=0}^n\alpha_j\gamma(\sigma_j)\right]d\mu(\alpha),$$

$$\int_{\alpha\in\mathcal{R}^{n+1}}|d\mu(\alpha)|<\infty.$$

In this case we have

$$f[\gamma]=\int\exp[-i(\gamma',\gamma)]\,d\mu_H(\gamma'),\tag{24}$$

where $\mu_H\in\mathcal{M}(H)$ and for Borel $B\subset H$

$$\mu_H(B)=\begin{cases}(\lambda\times\mu)(A), & \text{if } B\supset\left\{\sum_{j=0}^n\alpha_jG(\sigma_j,\tau)\right.\\ & :(\sigma_1,\ldots,\sigma_n,\alpha_0,\ldots,\alpha_n)\\ & \left.\in\text{Borel }A',\ A'\subset[0,t]^n\times\mathcal{R}^{n+1}\right\},\\ & A=\cup A',\\ 0, & \text{otherwise,}\end{cases}\tag{25}$$

λ being the Lebesgue measure on $[0,t]^n$, $(\lambda\times\mu)$ being the product measure on $[0,t]^n\times\mathcal{R}^{n+1}$. Then by definition

$$\mathcal{F}(f)=\int_0^t\cdots\int_0^t d\sigma_1\cdots d\sigma_n\int_{\alpha\in\mathcal{R}^{n+1}}\times\exp\left(-\tfrac{1}{2}i\sum_{j,k=0}^n\alpha_j\alpha_k(t-\sigma_j\vee\sigma_k)\right)d\mu(\alpha).\tag{26}$$

The functionals f in Examples 1, 2, and 3 depend upon $\gamma(\sigma)$ for only a finite number of values of $\sigma\in[0,t]$. Such functionals are usually called cylinder functionals. We shall work with a more restricted class of cylinder functionals than is usual.

Define the linear map $\pi_n:H\to S$, the step functions on $[0,t]$, by

$$(\pi_n\gamma)(\tau)=\gamma_j,\quad \frac{jt}{n}\le\tau<\frac{(j+1)t}{n},\tag{27}$$

$$\gamma_j=\gamma(jt/n),\quad j=0,1,\ldots,n-1.$$

Definition 5: The space of cylinder functionals $\mathcal{C}(H)$ is defined by $f\in\mathcal{C}(H)$ iff f is defined on S, $\exists\pi_n$ such that $(f\circ\pi_n)=f$, where $(f\circ\pi_n)$ denotes the composition $(f\circ\pi_n)(\gamma)=f[\pi_n\gamma]$, and

$$\mathcal{F}(f)=\int_{\alpha\in\mathcal{R}^n}\exp\frac{-i}{2}\left[\sum_{j,k=0}^{n-1}\alpha_j\alpha_kG\left(\frac{jt}{n},\frac{kt}{n}\right)\right]d\mu(\alpha_0,\ldots,\alpha_{n-1})\tag{28}$$

exists for μ defined by

$$f[\gamma]=(f\circ\pi_n)(\gamma_0,\ldots,\gamma_{n-1})=\int\exp\left[\left(-i\sum_{j=0}^{n-1}\alpha_j\gamma_j\right)\right]d\mu(\alpha_0,\ldots,\alpha_{n-1})\ .\tag{29}$$

Equations (28) and (29) define $\mathcal{F}(f)$ when $f\in\mathcal{C}(H)$.

Example 5: $f[\gamma]=(f\circ\pi_n)(\gamma)=\exp[-(i/2)\gamma^TF^{-1}\gamma]$, where F^{-1} is real symmetric nonsingular matrix and $\gamma^T=(\gamma_0,\ldots,\gamma_{n-1})\in\mathcal{R}^n$.

From the result

$$\int_{\mathcal{R}^n}\exp\left(-i\alpha^T\gamma-\frac{i}{2}\alpha^TF\alpha\right)d\alpha=\frac{(2\pi i)^{n/2}}{[\det(-F)]^{1/2}}\exp\left(\frac{i}{2}\gamma^TF^{-1}\gamma\right),\tag{30}$$

we have

$$\mathcal{F}(f)=(2\pi i)^{-n/2}[\det(-F)]^{1/2}\int_{\mathcal{R}^n}\exp(-\tfrac{1}{2}i\alpha^T(F+G)\alpha)\,d\alpha,\tag{31}$$

where $G_{ij}=G(it/n,jt/n)=t-(t/n)(i\vee j)$, $i,j=0,1,\ldots,n-1$, $\alpha^T=(\alpha_0,\ldots,\alpha_{n-1})$. A second application of (30) yields

$$\mathcal{F}[\exp(-\tfrac{1}{2}i\gamma^TF^{-1}\gamma)]=(\det G)^{-1/2}[\det(F^{-1}+G^{-1})]^{-1/2}.\tag{32}$$

We require this result in Sec. (5).

To conclude this section we see that $f\in\mathcal{C}(H)\cup\mathcal{F}(H)$, $\mathcal{F}(f)$ is *consistently* defined by Eqs. (5), (6), (28), and (29). In later sections we denote $\mathcal{C}(H)\cup\mathcal{F}(H)$ by $\mathcal{F}^c(H)$ and the above definition of $\mathcal{F}$ by $\mathcal{F}_{\text{DAH}}$. The most important application of $\mathcal{F}_{\text{DAH}}$ is the Feynman–Itô formula given in the next section.

5. THE FEYNMAN-ITÔ FORMULA

Theorem 3: The solution of the Schrödinger equation

$$i\frac{\partial\psi}{\partial t}=\frac{-1}{2}\ \frac{\partial^2\psi}{\partial X^2}+V[X]\psi\tag{33}$$

with Cauchy data $\psi(X,0)=\phi[X]=\int\exp(i\alpha X)\,d\nu(\alpha)\in L^2(\mathcal{R}^1)$, with a real-valued potential $V[X]=\int\exp(i\alpha X)\,d\mu(\alpha)$; μ,ν being of bounded absolute variation on $\mathcal{R}^1$, is

$$\psi(X,t)=\mathcal{F}[\exp(-i\int_0^tV[\Gamma(\tau)+X]\,d\tau)\phi[\Gamma(0)+X]],\tag{34}$$

$$\exp(-i\textstyle\int_0^t V[\Gamma(\tau)+X]d\tau)\phi[\Gamma(0)+X]\in\mathcal{F}(H).$$

Proof: The free particle Hamiltonian $H_0=-\frac{1}{2}\partial^2/\partial X^2$ is self-adjoint in $L^2(\mathcal{R}^1)$ on its natural domain. Because μ is of bounded absolute variation on $\mathcal{R}^1$, $V[X]$ is bounded and continuous. Hence, by the Kato–Rellich theorem $H=(H_0+V)$ is self-adjoint on the natural domain of H_0 and

$$\psi(X,t)=[\exp(-itH)\phi](X).\tag{35}$$

Consider now the linear differential equation

$$\frac{dy}{dt}=-iV(t)y,\tag{36}$$

$y\in E=\mathcal{L}(L^2(\mathcal{R}^1),L^2(\mathcal{R}^1))$, the Banach space of bounded linear transformations $y:L^2(\mathcal{R}^1)\to L^2(\mathcal{R}^1)$, where $V(t)=\exp(itH_0)V\exp(-itH_0)$. Then $V(t)$ considered as an element of $\mathcal{L}(E,E)$ is a regulated function of $t\in[0,\infty)$. It follows that there is an unique solution $y(t)\in E$, $t\in[0,\infty)$, satisfying (36) and $y(0)=1$.[10] Furthermore

$$y(t)=\operatorname*{s-lim}_{n\to\infty}y_n(t),\tag{37}$$

where $y_0=1$ and $y_n(t)=1-i\int_0^ty_{n-1}(t')V(t')\,dt'$. However, $y(t)=\exp(itH_0)\exp(-itH)\in E$, satisfies (36) and $y(0)=1$. It follows that

$$\exp(-itH)=\sum_{n=0}^\infty(-i)^n\int_{0\le t_1\le\cdots\le t}\cdots\int\exp[-i(t-t_n)H_0]V\times\exp[-i(t_n-t_{n-1})H_0]\cdots\exp[-i(t_2-t_1)H_0]V\times\exp(-it_1H_0)\,dt_1\cdots dt_n.\tag{38}$$

Also, we have

$$\exp(-itH_0)\exp(i\alpha X)=\exp(-it\alpha^2/2)\exp(i\alpha X)\tag{39}$$

and using the fact that V and ϕ are the Fourier transforms of μ and ν gives in (35)

$$\psi(X,t)=\sum_{n=0}^{\infty}(-i)^n\int_{0\le t_1\le\cdots\le t}\cdots\int\int\cdots\int\exp(-\tfrac{1}{2}i)\times[(t-t_n)(\alpha_0+\cdots+\alpha_n)^2+\cdots+(t_2-t_1)(\alpha_0+\alpha_1)^2+t_1\alpha_0^2]\exp i\sum_{j=0}^{n}\alpha_jX\,d\nu(\alpha_0)\prod_{j=1}^{n}d\mu(\alpha_j)\,dt_j. \quad (40)$$

Introducing $t_0=0$, we have

$$\psi(X,t)=\sum_{n=0}^{\infty}(-i)^n\int_{0\le t_1\le\cdots\le t}\cdots\int\int\cdots\int\exp\left(-\tfrac{1}{2}i\sum_{j,k=0}^{n-1}(t-t_j\nu t_k)\alpha_j\alpha_k\right)\exp\left(i\sum_{j=0}^{n}\alpha_jX\right)d\nu(\alpha_0)\times\prod_{j=1}^{n}d\mu(\alpha_j)\,dt_j, \quad (41)$$

which by symmetry gives

$$\psi(X,t)=\sum_{n=0}^{\infty}\frac{(-i)^n}{n!}\int_0^t\cdots\int_0^t\int\cdots\int\exp\left(\frac{-i}{2}\sum_{j,k=0}^{n-1}(t-t_j\nu t_k)\alpha_j\alpha_k\right)\exp\left(i\sum_{j=0}^{n}\alpha_jX\right)d\nu(\alpha_0)\times\prod_{j=1}^{n}d\mu(\alpha_j)\,dt_j. \quad (42)$$

However,

$$\mathcal{J}[V[\Gamma(t_1)+X]V[\Gamma(t_2)+X]\cdots V[\Gamma(t_n)+X]\phi[\Gamma(0)+X]]$$
$$=\mathcal{J}\left[\int\cdots\int\exp\left(i\sum_{j=0}^{n}\alpha_j\Gamma(t_j)\right)\times\exp\left(i\sum_{j=0}^{n}\alpha_jX\right)d\mu(\alpha_n)\cdots d\nu(\alpha_0)\right]$$
$$=\int\cdots\int\exp\left(-\tfrac{1}{2}i\sum_{j,k=0}^{n-1}(t-t_j\nu t_k)\alpha_j\alpha_k\right)\times\exp\left(i\sum_{j=0}^{n}\alpha_jX\right)d\mu(\alpha_n)\cdots d\nu(\alpha_0). \quad (43)$$

Therefore, from (42),

$$\psi(X,t)=\mathcal{J}\left[\sum_{n=0}^{\infty}\frac{(-i)^n}{n!}\int_0^t V[\Gamma(t_1)+X]\,dt_1\cdots\times\int_0^t V[\Gamma(t_n)+X]\,dt_n\phi[\Gamma(0)+X]\right]$$

and finally using Theorem 2

$$\psi(X,t)=\mathcal{J}[\exp(-i\int_0^t V[\Gamma(\tau)+X]\,d\tau)\phi[\Gamma(0)+X]]$$

We shall require this result in Sec. 7.

6. THE EXTENSION OF $\mathcal{J}_{\mathrm{DAH}}$

The space of functionals $\mathcal{J}^c(H)$ does not contain all the functionals we are required to integrate. To enlarge the class of integrable functionals we try to extend $\mathcal{J}$ to $\overline{\mathcal{J}^c(H)}$ the closure of $\mathcal{J}^c(H)$ in the topology defined by the seminorms $\{p\gamma\}_{\gamma\in H}$, $p\gamma(f)=|f(\gamma)|$. Clearly $\sup_{\gamma\in H}p\gamma(f)\le\|f\|_0$. Hence the topology defined by $\{p\gamma\}_{\gamma\in H}$ is weaker than the $\|\ \|_0$ topology. We call this topology the weak topology. We do not succeed in defining $\mathcal{J}$ on the whole of $\overline{\mathcal{J}^c(H)}$. We define $\mathcal{J}$ by linearity and continuity on the subspace $\mathcal{J}(P_\infty H)$. $\mathcal{J}(P_\infty H)$ contains all the functionals we require.

We define the linear map $P_n: H\to H$ by

$$(P_n\gamma)(\tau)=\sum_{j=0}^{n-1}\left[G\left(\frac{(j+1)t}{n},\tau\right)-\left(G\frac{jt}{n},\tau\right)\right][\gamma_{j+1}-\gamma_j]\left(\frac{t}{n}\right)^{-1}, \quad (45)$$

where $\gamma_j=\gamma(jt/n)$, $j=0,1,\ldots,n$. From the reproducing kernel property $\forall\ \gamma,\gamma'\in H$

$$(\gamma',P_n\gamma)=\sum_{j=0}^{n-1}\frac{(\gamma'_{j+1}-\gamma'_j)}{t/n}\frac{(\gamma_{j+1}-\gamma_j)}{t/n}\frac{t}{n}=(P_n\gamma',\gamma). \quad (46)$$

Also, we have

$$(P_n\gamma)(\tau)=\gamma_j+\left(\tau-\frac{jt}{n}\right)(\gamma_{j+1}-\gamma_j)\left(\frac{t}{n}\right)^{-1},\quad \frac{jt}{n}\le\tau<(j+1)\frac{t}{n},\quad j=0,1,\ldots,n-1.$$

Evidently then $P_n^2=P_n$. Since P_n is everywhere defined on H, the closed graph theorem implies that P_n is a projection.

Theorem 4: $V=\{\gamma\in H:\|P_n\gamma-\gamma\|\to 0$ as $n\to\infty\}=H$.

Proof: First V is a closed subspace of H. V is a subspace because P_n is linear. Let $V\ni\gamma_m$ and $\|\gamma_m-\gamma\|\to 0$ as $m\to\infty$.

$$\|P_n\gamma-\gamma\|=\|P_n\gamma-\gamma-P_n\gamma_m+\gamma_m+P_n\gamma_m-\gamma_m\|\le\|P_n(\gamma-\gamma_m)\|+\|\gamma-\gamma_m\|+\|P_n\gamma_m-\gamma_m\|. \quad (47)$$

Hence,

$$\|P_n\gamma-\gamma\|\le 2\|\gamma-\gamma_m\|+\|P_n\gamma_m-\gamma_m\|. \quad (48)$$

Given $\epsilon>0$, $\exists\ N_\epsilon$ such that $\|\gamma-\gamma_m\|<\epsilon/4$ when $m=N_\epsilon$. Also, $\exists\ N(m,\epsilon)$ such that $\|P_n\gamma_m-\gamma_m\|<\epsilon/2$ $n\ge N(m,\epsilon)$. From (48), when $n\ge N(N_\epsilon,\epsilon)$, $\|P_n\gamma-\gamma\|<\epsilon$. Thus, $\|P_n\gamma-\gamma\|\to 0$ as $n\to\infty$, so V is a closed subspace of H.

Let $\gamma\in H$. Then $\exists\ \alpha_0,\alpha_n,\beta_n\in\mathcal{R}$ such that

$$\left\|\frac{d\gamma_w}{d\tau}-S_N\right\|_{L_2}\to 0\ \text{ as } N\to\infty,$$

where

$$S_N(\tau)=\alpha_0+\sum_1^N \alpha_n\cos\left(\frac{2\pi n\tau}{t}\right)+\sum_1^N \beta_n\sin\left(\frac{2\pi n\tau}{t}\right). \tag{49}$$

Hence, defining

$$T_N(\tau)=-\int_\tau^t S_N(\tau')\,d\tau',\quad \|\gamma-T_N\|=\left\|\frac{d\gamma_w}{d\tau}-S_N\right\|_{L_2}\to 0$$

as $N\to\infty$. It is not difficult to show that for each N $T_N\in V$, thus, $\gamma\in V$.

Corollary:

$$\forall\gamma\in H,\quad \|P_n\gamma-\gamma\|^2=(\gamma,\gamma)-(\gamma,P_n\gamma)\to 0,\quad \text{as } n\to\infty. \tag{50}$$

When f is a continuous functional on H, in the sense that $(f\circ P_n)\to f$, in the weak topology, as $n\to\infty$, it is natural to define

$$\mathcal{J}(f)=\lim_{n\to\infty}\mathcal{J}_{\mathrm{DAH}}[f\circ P_n], \tag{51}$$

whenever this limit exists.

Definition 6: The space of functionals $\mathcal{J}(P_\infty H)$ is defined by $f\in\mathcal{J}(P_\infty H)$ iff

$$\mathcal{J}(f)=\lim_{n\to\infty}\mathcal{J}_{\mathrm{DAH}}[f\circ P_n] \tag{52}$$

exists.

Theorem 5: $\mathcal{J}(H)\subset\mathcal{J}(P_\infty H)$ and $\mathcal{J}$ is an extension of $\mathcal{J}_{\mathrm{DAH}}$.

Proof: First of all $\mathcal{J}(H)\subset\mathcal{J}(P_\infty H)$. Let $f\in\mathcal{J}(H)$ and $f(\gamma)=\int\exp[-i(\gamma',\gamma)]\,d\mu(\gamma')$. Then

$$(f\circ P_n)(\gamma)=\int_H\exp[-i(\gamma',P_n\gamma)]\,d\mu(\gamma')=\int_H\exp[-i(P_n\gamma',\gamma)]\,d\mu(\gamma')=\int_H\exp[-i(\gamma'',\gamma)]\,d\mu(P_n^{-1}\gamma''). \tag{53}$$

Then by definition

$$\mathcal{J}_{\mathrm{DAH}}[f\circ P_n]=\int_H\exp[-\tfrac12 i(\gamma'',\gamma'')]\,d\mu(P_n^{-1}\gamma'')=\int_H\exp[-\tfrac12 i(P_n\gamma',P_n\gamma')]\,d\mu(\gamma')=\int_H\exp[-\tfrac12 i(\gamma',P_n\gamma')]\,d\mu(\gamma'). \tag{54}$$

Consider now

$$\mathcal{J}_{\mathrm{DAH}}(f)=\int_H\exp[-\tfrac12 i(\gamma',\gamma')]\,d\mu(\gamma').$$

We have

$$|\mathcal{J}_{\mathrm{DAH}}[f\circ P_n]-\mathcal{J}_{\mathrm{DAH}}(f)|\le\int_H|\exp[-\tfrac12 i(\gamma',P_n\gamma')]-\exp[-\tfrac12 i(\gamma',\gamma')]|\,|d\mu(\gamma')|<2\int_H|d\mu(\gamma')|<\infty. \tag{55}$$

The result follows from the last corollary and the dominated convergence theorem.

Secondly $f\in\mathcal{C}(H)\cap\mathcal{J}(P_\infty H)\Rightarrow\mathcal{J}(f)=\mathcal{J}_{\mathrm{DAH}}(f)$. If $f\in\mathcal{C}(H)$ then $\exists\ \pi_n$ such that $f=(f\circ\pi_n)$. For any integer m

$$\pi_n\circ P_{mn}=P_{mn}\circ\pi_n=\pi_n. \tag{56}$$

$$\mathcal{J}(f)=\lim_{n\to\infty}\mathcal{J}_{\mathrm{DAH}}[f\circ P_n]\text{exists}\Rightarrow\mathcal{J}(f)=\lim_{m\to\infty}\mathcal{J}_{\mathrm{DAH}}[f\circ P_{mn}]=\lim_{m\to\infty}\mathcal{J}_{\mathrm{DAH}}[(f\circ\pi_n)\circ P_{mn}]=\mathcal{J}_{\mathrm{DAH}}[f].$$

$\mathcal{J}$ as defined by Eq. (52) clearly extends $\mathcal{J}_{\mathrm{DAH}}$ to $\mathcal{J}(P_\infty H)$.

The above extension of $\mathcal{J}_{\mathrm{DAH}}$ enables us to integrate a wider class of functionals. One of the most important is given below.

Example 6: Let $D(\tau)\in C[0,t)$ be the unique solution of $\ddot D(\tau)+\lambda p(\tau)D(\tau)=0$ with $p(\tau)\in C[0,t]$ and $\dot D(0)=0$, $D(0)=1$, $D(t)\ne 0$.

Then, if

$$f[\gamma]=\exp\left(\frac{-i\lambda}{2}\right)\int_0^t p(\tau)\gamma^2(\tau)\,d\tau,\quad \lambda\in\mathcal{R},\quad \mathcal{J}[f]=[D(t)]^{-1/2}. \tag{57}$$

First of all

$$(f\circ P_n)(\gamma)=\exp\left(\frac{-i\lambda}{2}\right)\sum_{j=0}^{n-1}\left[p_j\int_{jt/n}^{(j+1)t/n}\left(\gamma_j+\left(\tau-\frac{jt}{n}\right)(\gamma_{j+1}-\gamma_j)\frac{n}{t}\right)^2 d\tau\right]=\exp\left(-\frac{i\lambda}{2}\right)\sum_{j=0}^{n-1}\left(\frac{pj}{3}(\gamma_j^2+\gamma_j\gamma_{j+1}+\gamma_{j+1}^2)\frac{t}{n}\right), \tag{58}$$

where $pj=p(\tau_j')$, for some $\tau_j'\in[jt/n,(j+1)t/n]$. Then, using the notation $\gamma^T=(\gamma_0,\ldots,\gamma_{n-1})$,

$$(f\circ P_n)(\gamma)=\exp(-\tfrac12 i\gamma^T F^{-1}\gamma), \tag{59}$$

where

$$F^{-1}=-\lambda t/n\begin{bmatrix} p_0/3 & p_0/6 & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_0/6 & (p_0+p_1)/3 & p_1/6 & \cdot & \cdot & \cdot & \cdot \\ \cdot & p_1/6 & (p_1+p_2)/3 & p_2/6 & \cdot & \cdot & \cdot \\ - & - & - & - & - & - & - \\ \cdot & \cdot & \cdot & \cdot & p_{n-2}/6 & (p_{n-2}+p_{n-1})/3 & p_{n-1}/6 \\ \cdot & \cdot & \cdot & \cdot & \cdot & p_{n-1}/6 & (p_{n-1}+p_n)/3 \end{bmatrix}. \tag{60}$$

From Example 5, using

$$(G_{ij})=\left(G\left(\frac{it}{n},\frac{jt}{n}\right)\right)=t\begin{bmatrix} 1 & 1-1/n & 1-2/n & 1-3/n & \cdot & \cdot & \cdot & 1/n \\ 1-1/n & 1-1/n & 1-2/n & 1-3/n & \cdot & \cdot & \cdot & 1/n \\ 1-2/n & 1-2/n & 1-2/n & 1-3/n & \cdot & \cdot & \cdot & 1/n \\ 1-3/n & 1-3/n & 1-3/n & 1-3/n & \cdot & \cdot & \cdot & 1/n \\ - & - & - & - & - & - & - & - \\ 1/n & 1/n & 1/n & - & - & - & - & 1/n \end{bmatrix}, \tag{61}$$

$$(G_{ij}^{-1})=\left(\frac{t}{n}\right)^{-1}\begin{bmatrix} 1 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & -1 & 2 & -1 & \cdot & \cdot & \cdot \\ - & - & - & - & - & - & - \\ \cdot & \cdot & \cdot & \cdot & -1 & 2 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & -1 & 2 \end{bmatrix}, \tag{62}$$

it follows that

$$\mathcal{J}[f\circ P_n]=(\det G)^{-1}[\det(F^{-1}+G^{-1})]^{-1/2}=(\det D_{ij}^n)^{-1/2}, \tag{63}$$

where

$$\Delta t=\frac{t}{n},\quad q_j=p_j+p_{j-1},\quad j=0,1,\ldots,n,\quad (p_{-1}=0),$$

$$(D_{ij}^n)=\begin{bmatrix} 1-\frac{\lambda(\Delta t)^2}{3}q_0 & -1-\frac{\lambda(\Delta t)^2}{6}p_0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1-\frac{\lambda(\Delta t)^2}{6}p_0 & 2-\frac{\lambda(\Delta t)^2 q_1}{3} & -1-\frac{\lambda(\Delta t)^2 p_1}{6} & \cdot & \cdot & \cdot & \cdot \\ \cdot & -1-\frac{\lambda(\Delta t)^2 p_1}{6} & 2-\frac{\lambda(\Delta t)^2 q_2}{3} & -1-\frac{\lambda(\Delta t)^2 p_2}{6} & \cdot & \cdot & \cdot \\ ---- & ---- & ---- & ---- & ---- & ---- & ---- \\ ---- & ---- & ---- & ---- & ---- & ---- & ---- \\ \cdot & \cdot & \cdot & \cdot & -1-\frac{\lambda(\Delta t)^2 p_{n-2}}{6} & 2-\frac{\lambda(\Delta t)^2 q_{n-1}}{3} & -1-\frac{\lambda(\Delta t)^2 p_{n-1}}{6} \\ \cdot & \cdot & \cdot & \cdot & \cdot & -1-\frac{\lambda(\Delta t)^2 p_{n-1}}{6} & 2-\frac{\lambda(\Delta t)^2 q_n}{3} \end{bmatrix}. \tag{64}$$

Determinants similar to the one above arise in the study of Wiener measure. Our evaluation borrows from the paper of Gel'fand and Yaglom.[11]

Let D_k^n be the minor of order $(k+1)$ in the top left-hand corner of D^n. Then we have

$$\frac{(D_{k+1}^n-2D_k^n+D_{k-1}^n)}{(\Delta t)^2}+\lambda\frac{(p_k+p_{k+1})}{3}D_k^n+\frac{\lambda p_k}{3}D_{k-1}^n=\frac{\lambda^2(\Delta t)^2p_k^2D_{k-1}^n}{36} \tag{65}$$

for $2\leq k\leq n-1$ and

$$D_0^n=1-\frac{\lambda(\Delta t)^2p_0}{3},\quad \frac{D_1^n-D_0^n}{\Delta t}=\frac{-\lambda\Delta t(3p_0+p_1)}{3}+\frac{\lambda^2(\Delta t)^3p_0}{9}\left(\frac{3p_0+4p_1}{4}\right). \tag{66}$$

For fixed n we can in principle solve the above difference equations for $D_{n-1}^n=\det(D_{ij}^n)$. However, we only require $\lim_{n\to\infty}D_{n-1}^n$. Defining $D^n(\tau)\in C[0,t]$ by $D^n(k\Delta t)=D_k^n$, we see that, as $n\to\infty$, $D^n(\tau)\to D(\tau)$ where $D(\tau)$ is the unique solution of $\ddot{D}(\tau)+\lambda p(\tau)D(\tau)=0$, $D(0)=1$, $\dot{D}(0)=0$, $D(\tau)\in C[0,t]$. Hence, if $D(t)\neq 0$,

$$\det D_{ij}^n=D_{n-1}^n\to D(t),\quad \text{as } n\to\infty. \tag{67}$$

We conclude this section with our definition of $\mathcal{J}$ for the Hilbert space of paths, H.

Definition 7: Let G be the $(n\times n)$ matrix in Eq. (61), P_n the projection defined on H by Eq. (45) and $f\in\mathcal{J}(P_\infty H)$. Then $\mathcal{J}(f)$ is defined by

$$\mathcal{J}(f)=\lim_{n\to\infty}\int \exp(-\tfrac{1}{2}i\alpha^T G\alpha)\,d\mu_n(\alpha), \tag{68}$$

where μ_n is the complex measure

$$[f\circ P_n](\gamma)=\int \exp(-i\alpha^T\gamma)\,d\mu_n(\alpha),\quad \alpha^T=(\alpha_0,\ldots,\alpha_{n-1}),\quad \gamma^T=(\gamma_0,\ldots,\gamma_{n-1}). \tag{69}$$

7. THE QUASICLASSICAL REPRESENTATION

Theorem 6: Let $\psi(x,t)$ be the solution of

$$i\hbar\frac{\partial\psi}{\partial t}=-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}+V(x)\psi \tag{70}$$

with Cauchy data $\psi(x,0)=\phi(x)\in L_2$, where V and ϕ are the Fourier transforms of complex measures of bounded absolute variation. Let $X_{\rm cl}(\tau)\in C^2[0,t]$ be the real solution of

$$m\ddot{X}_{\rm cl}(\tau)=-\frac{\partial V}{\partial x}[X_{\rm cl}(\tau)],\quad \tau\in[0,t], \tag{71}$$

$X_{\rm cl}(t)=x$ and define the classical action $S_{\rm cl}$, corresponding to the trajectory $X_{\rm cl}(\tau)$, by

$$S_{\rm cl}=\int_0^t\frac{m}{2}\dot{X}_{\rm cl}^2(\tau)\,d\tau-\int_0^t V[X_{\rm cl}(\tau)]\,d\tau.$$

Then

$$\psi(x,t)=\exp\frac{iS_{\rm cl}}{\hbar}\,\mathcal{J}\left[\exp\left(-\frac{i}{\hbar}\int_0^t\Delta^2V\left[X,\left(\frac{\hbar}{m}\right)^{1/2}\Gamma\right]\right)d\tau\right.$$
$$\times\exp\left(-i\left(\frac{m}{\hbar}\right)^{1/2}\frac{dX_{\rm cl}}{d\tau}(0)\Gamma(0)\right)\phi\left[\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(0)\right.$$
$$\left.\left.+X_{\rm cl}(0)\right]\right], \tag{72}$$

where

$$\Delta^2V\left[X_{\rm cl},\left(\frac{\hbar}{m}\right)^{1/2}\Gamma\right]=V\left[X_{\rm cl}(\tau)+\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(\tau)\right]$$
$$-V[X_{\rm cl}(\tau)]-\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(\tau)\frac{\partial V}{\partial x}[X_{\rm cl}(\tau)].$$

Proof: The result above depends upon the transformation properties of $\mathcal{J}$ when the underlying Hilbert space undergoes a parallel translation $H\ni\gamma\to\gamma+a$, fixed $a\in H$. Since all the functionals f involved are such that $f\in\mathcal{J}(H)$ we work with $\mathcal{J}=\mathcal{J}_{\rm DAH}$.

Let

$$f(\gamma)=\int\exp[-i(\gamma',\gamma)]\,d\mu(\gamma').$$

Then

$$f(\gamma+a)=\int\exp[-i(\gamma',\gamma)]\exp[-i(\gamma',a)]\,d\mu(\gamma').$$

By definition

$$\mathcal{J}[f(\gamma+a)]=\int\exp[-\tfrac{1}{2}i(\gamma',\gamma')]\exp[-i(\gamma',a)]\,d\mu(\gamma')$$
$$=\exp[\tfrac{1}{2}i(a,a)]\int\exp[-\tfrac{1}{2}i(\gamma'+a,\gamma'+a]\,d\mu(\gamma'). \tag{73}$$

Hence,

$$\mathcal{J}[f(\gamma+a)]=\exp[\tfrac{1}{2}i(a,a)]\int\exp[-\tfrac{1}{2}i(\gamma'',\gamma'')]\,d\mu(\gamma''-a). \tag{74}$$

However, we have

$$\int\exp[-i(\gamma',\gamma)]\,d\mu(\gamma'-a)$$
$$=\int\exp[-i(\gamma'-a,\gamma)]\exp[-i(a,\gamma)]\,d\mu(\gamma'-a)$$
$$=\exp[-i(a,\gamma)]f(\gamma). \tag{75}$$

Thus, we have

$$\mathcal{J}[f(\gamma+a)]=\exp[\tfrac{1}{2}i(a,a)]\mathcal{J}[\exp(-i(a,\gamma)f(\gamma))]. \tag{76}$$

The equation

$$i\hbar\frac{\partial\psi}{\partial t}=-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}+V(x)\psi, \tag{77}$$

with Cauchy data $\psi(x,0)=\phi(x)$ is equivalent to the equation

$$i\frac{\partial\psi}{\partial t}=-\frac{1}{2}\frac{\partial^2\psi}{\partial X^2}+\frac{1}{\hbar}V\left[\left(\frac{\hbar}{m}\right)^{1/2}X\right]\psi,$$

where $X=(m/\hbar)^{1/2}x$ with Cauchy data

$$\psi(X,0)=\phi[(\hbar/m)^{1/2}X]. \tag{78}$$

The Feynman–Itô formula then gives

$$\psi(x,t)=\mathcal{J}\left[\exp\left(-\frac{i}{\hbar}\int_0^t V\left[\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(\tau)+x\right]d\tau\right)\right.$$
$$\left.\times\phi\left[\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(0)+x\right]\right]. \tag{79}$$

We now make the parallel translation $\Gamma\to\Gamma+(m/\hbar)^{1/2}\times\Gamma_{\rm cl}$, where $\Gamma_{\rm cl}\in H$ and $x+\Gamma_{\rm cl}=X_{\rm cl}$. From translational invariance

$$\psi(x,t)=\exp\left(\frac{im}{2\hbar}(X_{\rm cl},X_{\rm cl})\right)\mathcal{J}\left[\exp\left\{-\frac{i}{\hbar}\left(\int_0^t V\left[\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(\tau)\right.\right.\right.\right.$$
$$\left.\left.\left.+X_{\rm cl}(\tau)\right]d\tau-(m\hbar)^{1/2}\int_0^t\frac{d\Gamma_\omega}{d\tau}\frac{dX_{\rm cl}}{d\tau}d\tau\right)\right\}$$
$$\left.\times\phi\left[\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(0)+X_{\rm cl}(0)\right]\right]. \tag{80}$$

Integrating by parts, we have

$$\int_0^t\frac{d\Gamma_\omega}{d\tau}\frac{dX_{\rm cl}}{d\tau}d\tau$$
$$=-\Gamma(0)\frac{dX_{\rm cl}(0)}{d\tau}+\int_0^t\frac{\Gamma(\tau)}{m}\frac{\partial V}{\partial x}[X_{\rm cl}(\tau)]\,d\tau, \tag{81}$$

where we have used $\Gamma(t)=0$, which follows because $\Gamma_{\rm cl}(t)=0$. However,

$$X_{\rm cl}\in C^2[0,t]\Rightarrow(X_{\rm cl},X_{\rm cl})=\int_0^t\left(\frac{dX_{\rm cl}}{d\tau}\right)^2d\tau.$$

Finally, putting

$$S_{\rm cl}=\int_0^t\frac{m}{2}\dot{X}_{\rm cl}^2(\tau)\,d\tau-\int_0^t V[X_{\rm cl}(\tau)]\,d\tau,$$

in toto,

$$\psi(x,t)=\exp\left(\frac{iS_{cl}}{\hbar}\right)\mathcal{J}\left[\exp\left\{-\frac{i}{\hbar}\left(\int_0^t V\left[X_{cl}(\tau)+\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(\tau)\right]\right.\right.\right.$$
$$\left.\left.-V[X_{cl}(\tau)]-\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(\tau)\frac{\partial V}{\partial x}[X_{cl}(\tau)]d\tau\right)\right\}$$
$$\left.\times\exp\left(-i\left(\frac{m}{\hbar}\right)^{1/2}\frac{dX_{cl}(0)}{d\tau}\Gamma(0)\right)\phi\left[X_{cl}(0)+\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(0)\right]\right]. \tag{82}$$

This is the quasiclassical representation.

In the above representation we can choose $dX_{cl}(0)/d\tau$ for convenience depending upon the exact form of the initial wavefunction ϕ. When V and ϕ are sufficiently regular we can use the above expression to obtain a power series expansion of ψ in ascending powers of $\hbar^{1/2}$.

8. QUANTUM MECHANICS IN THE LIMIT AS $\hbar\to 0$

We now investigate the relationship which obtains between the solution $\psi(x,t)$ of the Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t}=-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}+V(x)\psi, \tag{83}$$

with $\psi(x,0)=\phi(x)$ and the solution $x=x(x_0,p_0,t)$ of the classical equation of motion

$$m\frac{d^2x}{dt^2}=-\frac{\partial V}{\partial x}[x] \tag{84}$$

satisfying $m(dx/dt)(0)=p_0$, $x(0)=x_0$, where V and ϕ satisfy the conditions of Theorem 6 and $V\in C^2$.

It is possible to carry out this investigation for a variety of initial wavefunctions ϕ. Here we shall restrict our attention to the physically important case $\phi(x)=\exp(ip_0x/\hbar)\psi_0(x)$, where p_0 and ψ_0 are independent of $\hbar$. In the limit as $\hbar\to 0$, this boundary condition is equivalent to giving the quantum mechanical particle an initial momentum p_0.

We shall assume the classical problem satisfies:

(1) The solution $x=x(x_0,p_0,t)$ exists and is unique for $t\in[0,T]$.

(2) The equation $x=x(x_0,p_0,t)$ can be solved uniquely to yield $x_0=x_0(x,p_0,t)$, $t\in[0,T]$.

(3) There is a unique $X(p_0,x,t,\tau)$ such that

$$m\frac{d^2X}{d\tau^2}(p_0,x,t,\tau)=-\frac{\partial V}{\partial x}[X(p_0,x,t,\tau)],\quad \tau\in[0,t]$$

$$m\frac{dX}{d\tau}(p_0,x,t,0)=p_0,\quad X(p_0,x,t,t)=x,\quad t\in[0,T].$$

(The above conditions are simultaneously satisfied for sufficiently small T when $\partial^2V/\partial x^2$ is bounded.[12]) Evidently the above uniqueness assumptions imply $X[p_0,x(x_0,p_0,t),t,\tau]=x(x_0,p_0,\tau)$ and $X(p_0,x,t,\tau)=x[x_0(x,p_0,t),p_0,\tau]$ This last condition is used below.

Theorem 7: Let $\psi(x,t)$ be the solution of the Schrödinger equation (83) with Cauchy data $\psi(x,0)=\exp(ip_0x/\hbar)\psi_0(x)$. Let $\tilde S_{cl}(p_0,x,t)$ be the unique solution of the Hamilton–Jacobi equation, satisfying

$$\tilde S_{cl}(p_0,x,0)=p_0x,\quad\left(\frac{(\partial\tilde s/\partial x)^2}{2m}\right)+V(x)+\frac{\partial\tilde s}{\partial t}=0,$$

so that

$$\tilde S_{cl}(p_0,x,t)=p_0x_0(x,p_0,t)+\int_0^t\frac{m}{2}\dot X^2(p_0,x,t,\tau)-V[X(p_0,x,t,\tau)]d\tau.$$

Then, if the above conditions are valid and if $\mathcal{J}$ satisfies a dominated convergence theorem,

$$\lim_{\hbar\to 0}\exp\left[-\frac{i}{\hbar}\tilde S_{cl}(p_0,x,t)\right]\psi(x,t)=\psi_0[x_0(x,p_0,t)]\left(\frac{\partial x_0}{\partial x}(x,p_0,t)\right)^{1/2},$$

for $t\in[0,T]$.

Proof: Choose $m(dX_{cl}/d\tau)(0)=p_0$ in Theorem 6. Then we have

$$\exp\left[-\frac{i}{\hbar}\tilde S_{cl}(p_0,x,t)\right]\psi(x,t)=\mathcal{J}\left[\exp\left(-\frac{i}{\hbar}\int_0^t\Delta^2V\left[X_{cl},\left(\frac{\hbar}{m}\right)^{1/2}\right]d\tau\right)\times\psi_0\left[x_0(x,p_0,t)+\left(\frac{\hbar}{m}\right)^{1/2}\Gamma(0)\right]\right]. \tag{85}$$

We then deduce

$$\lim_{\hbar\to 0}\exp\left[-\frac{i}{\hbar}\tilde S_{cl}(p_0,x,t)\right]\psi(x,t)=\mathcal{J}\left[\exp\left(-\frac{i}{2m}\int_0^t\gamma^2(\tau)V''[X(p_0,x,t,\tau)]\,d\tau\right)\right]\times\psi_0[x_0(x,p_0,t)]. \tag{86}$$

From Example 6 we have

$$\lim_{\hbar\to 0}\exp\left(-\frac{i}{\hbar}\tilde S_{cl}(p_0,x,t)\right)\psi(x,t)=[D(p_0,x,t,t)]^{-1/2}\psi_0[x_0(x,p_0,t)], \tag{87}$$

where

$$\frac{d^2D}{d\tau^2}(p_0,x,t,\tau)+\frac{V''}{m}[X(p_0,x,t,\tau)]D(p_0,x,t,\tau)=0,$$
$$D(p_0,x,t,0)=1,\quad \dot D(p_0,x,t,0)=0.$$

We now find $D(p_0,x,t,\tau)$. The momentum $P=m(dX/d\tau)(p_0,x,t,\tau)$ satisfies the equation

$$\frac{dP}{d\tau}+V'[X(p_0,x,t,\tau)]=0. \tag{88}$$

However, from uniqueness we have $x[x_0(x,p_0t),p_0,\tau]=X(p_0,x,t,\tau)$. Differentiating the last equation with respect to x_0 gives

$$\frac{d^2}{\partial\tau^2}\left[\frac{\partial x}{\partial x_0}[x_0(x,p_0,t),p_0,\tau]\right]+\frac{V''}{m}[X(p_0,x,t,\tau)]\times\frac{\partial x}{\partial x_0}[x_0(x,p_0,t),p_0,\tau]=0. \tag{89}$$

Hence, $D=(\partial x/\partial x_0)[x_0(x,p_0t),p_0,\tau]$ satisfies the correct equation. D also satisfies the correct boundary conditions. Therefore, using the implicit function theorem,

$$D(p_0,x,t,t)=\frac{\partial x}{\partial x_0}[x_0(x,p_0,t),p_0,t]$$
$$=\left[\frac{\partial x_0}{\partial x}(x,p_0,t)\right]^{-1}, \tag{90}$$

the result follows.

A simple consequence of Theorem 7 is that

$$\lim_{\hbar\to 0}\int_a^b|\psi(x,t)|^2\,dx=\int_{x(x_0,p_0,t)\in[a,b]}|\psi(x_0,0)|^2\,dx_0, \tag{91}$$

Judicious use of the principle of stationary phase yields

$$\lim_{\hbar\to 0}\int_a^b|\tilde\psi(p,t)|^2\,dp=\int_{m\dot x(x_0,p_0,t)\in[a,b]}|\psi(x_0,0)|^2\,dx_0. \tag{92}$$

Similar results to these were first obtained by Maslov using an entirely different approach.[13]

It follows that if $\psi(x,0)=\exp(ip_0x/\hbar)\psi_0(x)$ is nonzero only in the neighborhood of some point x_0 then the probability of the quantum mechanical particle being in the neighborhood of the point (p,x) in phase space at time t will differ from zero as $\hbar\to 0$ only if $p=m\dot x(x_0,p_0,t)$ and $x=x(x_0,p_0,t)$. In this sense quantum mechanics $\to$ classical mechanics as $\hbar\to 0$.

It is interesting also to consider Wigner's quasi-probability density $\rho_\hbar^w(p,x,t)$. For the pure state ψ, $\rho_\hbar^w$ is defined by

$$\rho_\hbar^w(p,x,t)=(2\pi)^{-1}\int\exp[-(i\eta p)]\psi^*(x-\tfrac12\hbar\eta,t)\psi(x+\tfrac12\hbar\eta,t)\,d\eta. \tag{93}$$

For continuous bounded potentials $V, H=(H_0+V)$ is self-adjoint and $\exp(-iHt/\hbar)$ is a continuous unitary group.

Thus, $\|\psi\|_{L_2}=\|\phi\|_{L_2}$ and the integral on the right-hand side of (93) is absolutely convergent $\phi\in L_2$. Formally

$$\int\rho_\hbar^w(p,x,t)\,dp=|\psi(x,t)|^2,$$
$$\int\rho_\hbar^w(p,x,t)\,dx=|\tilde\psi(p,t)|^2, \tag{94}$$

but ρ is not positive definite. We now discuss $\lim_{\hbar\to 0}\rho_\hbar^w(p,x,t)$ for the pure state $\psi(x,t)$ with $\psi(x,0)=\exp(i\pi_0x/\hbar)\psi_0(x)$.

Theorem 8: Assuming the conditions in Theorem 7, for the pure state $\psi(x,t)$ with $\psi(x,0)=\exp(i\pi_0x/\hbar)\psi_0(x)$, in the topology of $\mathcal{D}'$, as $\hbar\to 0$,

$$\rho_\hbar^w(p,x,t)\to|\psi_0[x_0(x,\pi_0,t)]|^2\left\|\frac{\partial x_0}{\partial x}(x,\pi_0,t)\right\|$$
$$\times\delta(p-m\dot x[x_0(x,\pi_0,t),\pi_0,t]),$$

where

$$\frac{\partial x}{\partial\tau}[x_0,\pi_0,\tau]\Big|_{\tau=t}=\dot x[x_0,\pi_0,t].$$

Proof: From Theorem 7

$$\rho_\hbar^{1/2}(x,t)=\exp\left(-\frac{i}{\hbar}\tilde S_{\rm cl}(x,\pi_0,t)\right)\psi(x,t)$$
$$\to\psi_0[x_0(x,\pi_0,t)]\left[\frac{\partial x_0}{\partial x}(x,\pi_0,t)\right]^{1/2}=\rho^{1/2}(x,t), \tag{95}$$

pointwise, as $\hbar\to 0$. We assume this convergence is uniform, $\forall\, x\in K$, compact subsets of $\mathbb{R}$. Then we have

$$\rho_\hbar^{1/2}\left(x+\frac{\hbar\eta}{2},t\right)\to\rho^{1/2}(x,t), \tag{96}$$

as $\hbar\to 0$, $x\in K$.

Let $F(p,x)\in\mathcal{D}$. Then we define $\langle F(p,x)\rangle_t$ by

$$\langle F(p,x)\rangle_t=\int\rho_\hbar^w(p,x,t)F(p,x)\,dp\,dx. \tag{97}$$

$$\therefore\langle F(p,x)\rangle_t=(2\pi)^{-1}\int\exp\left\{\frac{i}{\hbar}\left[\tilde S\left(\pi_0,x+\frac{\hbar\eta}{2},t\right)-\tilde S\left(\pi_0,x-\frac{\hbar\eta}{2},t\right)\right]\right\}\rho_\hbar^{1/2}\left(x+\frac{\hbar\eta}{2},t\right)$$
$$\times\rho_\hbar^{1/2*}\left(x-\frac{\hbar\eta}{2},t\right)\exp[-i\eta\cdot p]F(p,x)\,d\eta\,dx\,dp. \tag{98}$$

However,

$$|F(p,x)\rho_\hbar^{1/2}(x+\hbar\eta/2,t)\rho_\hbar^{1/2*}(x-\hbar\eta/2,t)|$$
$$\le\tfrac12|F(p,x)|[|\rho_\hbar(x+\hbar\eta/2,t)|+|\rho_\hbar(x-\hbar\eta/2,t)|]. \tag{99}$$

Hence,

$$\int|F(p,x)||\rho_\hbar^{1/2}(x+\hbar\eta/2,t)\rho_\hbar^{1/2*}(x-\hbar\eta/2,t)|\,d\eta\,dx\,dp$$
$$\le 2(\|\psi\|_{L_2}^2/\hbar)\int|F(p,x)|\,dx\,dp<\infty. \tag{100}$$

Fubini's theorem now gives

$$\langle F(p,x)\rangle_t=(2\pi)^{-1/2}\int\exp\{(i/\hbar)[\tilde S(\pi_0,x+\hbar\eta/2,t)-\tilde S(\pi_0,x-\hbar\eta/2,t)]\}\hat F(\eta,x)\rho_\hbar^{1/2}(x+\hbar\eta/2,t)$$
$$\times\rho_\hbar^{1/2*}(x-\hbar\eta/2,t)\,d\eta\,dx, \tag{101}$$

where $\hat F(\eta,x)$ is the Fourier transform with respect to p of $F(p,x)$. However,

$$|\hat F(\eta,x)|\left|\rho_\hbar^{1/2}\left(x+\frac{\hbar\eta}{2},t\right)\rho_\hbar^{1/2*}\left(x-\frac{\hbar\eta}{2},t\right)\right|$$
$$\le\tfrac12\sup_{\eta,x}|(1+\eta^2)\hat F(\eta,x)|\,\frac{[|\rho_\hbar(x+\hbar\eta/2,t)|+|\rho_\hbar(x-\hbar\eta/2,t)|]}{(1+\eta^2)} \tag{102}$$

and

$$\sup_{\eta,x}|(1+\eta^2)\hat F(\eta,x)|$$
$$\le(2\pi)^{-1/2}\int\frac{dp}{(1+p^2)}\sup_{p,x}\left|(1+p^2)\left(1-\frac{d^2}{dp^2}\right)F(p,x)\right|$$
$$=M<\infty. \tag{103}$$

Hence

$$|\hat F(\eta,x)\rho_\hbar^{1/2}(x+\hbar\eta/2,t)\rho_\hbar^{1/2*}(x-\hbar\eta/2,t)|$$
$$\le\{M/[2(1+\eta^2)]\}[|\rho_\hbar(x+\hbar\eta/2,t)|+|\rho_\hbar(x-\hbar\eta/2,t)|] \tag{104}$$

and

$$\int(1+\eta^2)^{-1}[|\rho_\hbar(x+\hbar\eta/2,t)|+|\rho_\hbar(x-\hbar\eta/2,t)|]\,dx\,d\eta$$
$$\le 2\|\psi\|_{L_2}^2\int d\eta/(1+\eta^2)<\infty. \tag{105}$$

We can then apply Lebesgue's dominated convergence theorem in Eq. (101). Since $\tilde S$ is continuously differentiable with respect to x, we have

$$\langle F(p,x)\rangle_t \to (2\pi)^{-1/2}\int \exp\left[i\eta\frac{\partial\tilde{s}}{\partial x}(\pi_0,x,t)\right]\hat{F}(\eta,x)\times|\rho(x,t)|\,d\eta\,dx = \int F\left[\frac{\partial\tilde{s}}{\partial x}(\pi_0,x,t),x\right]|\rho(x,t)|\,dx, \quad \text{as } \hbar\to 0. \tag{106}$$

A straightforward differentiation yields

$$\frac{\partial\tilde{s}}{\partial x}(\pi_0,x,t)=m\dot{X}(\pi_0,x,t,t)=m\dot{x}[x_0(x,\pi_0,t),\pi_0,t]. \tag{107}$$

The result follows.

To see the classical limit of quantum mechanics we must choose the initial data to correspond to a particle with momentum π_0 at position ξ_0. Hence, we put $|\psi_0(y)|^2 = |\psi\alpha(y)|^2 \to \delta(y-\xi_0)$, as $\alpha\to 0$.

Finally then

$$\lim_{\alpha\to 0}\lim_{\hbar\to 0}\langle F(p,x)\rangle_t = F(m\dot{x}(\xi_0,\pi_0,t),x(\xi_0,\pi_0,t)), \tag{108}$$

or, in $\mathcal{D}'$,

$$\lim_{\alpha\to 0}\lim_{\hbar\to 0}\rho^w_\hbar(p,x,t)=\delta[p-m\dot{x}(\xi_0,\pi_0,t)]\delta[x-x(\xi_0,\pi_0,t)] \tag{109}$$

—the correct classical limit.

9. CONCLUSION

We have given here a rigorous extension of $\mathcal{J}$ the Feynman path integral for a nonrelativistic quantum mechanical particle moving in 1 dimension on the paths in the path space H. This definition of $\mathcal{J}$ has enabled us to integrate a wider class of functionals on H than was previously possible. Further we have obtained a new quasiclassical representation for the wavefunction solution of the Schrödinger equation for a quantum mechanical particle in a continuous bounded potential $V\in C^1$. We have seen that the first term in a formal power series expansion in $(\hbar/m)^{1/2}$ in our quasiclassical representation corresponds to the correct classical mechanical limit of quantum mechanics as $\hbar/m\to 0$. When the potential $V\in C^2$, this classical mechanical limit will be attained as $\hbar/m\to 0$, if $\mathcal{J}$ obeys some sort of dominated convergence theorem.

It is clear that the above ideas are easily generalized to a nonrelativistic quantum mechanical particle in n-dimensional Euclidean space. In this connection we expect $\mathcal{J}(P_\infty H)$ will at least include enough functionals to enable us to treat the quantum mechanical anharmonic oscillator (see Example 6). Whether this treatment will generalize to a quantum mechanical particle free to move in an n-dimensional Riemannian manifold is, however, not clear.[14] The definition of our projection P_n in terms of the reproducing kernel G of the Hilbert space H and subsequent definition of $\mathcal{J}$ critically depends upon the beautifully simple form of G for a free particle in Euclidean space.[15] It is hard to imagine, except in very special cases, that the corresponding kernel for the Hilbert space of paths for a particle in a n-dimensional Riemannian manifold will be so simple.

ACKNOWLEDGMENT

I am grateful to Professor J. T. Lewis of the Dublin Institute of Advanced Study for first interesting me in functional integration and providing me with useful references when this work was in its early stages.

[1](a) R.H. Cameron, J. Math. Phys. 39, 126–40 (1960); (b) K. Itô, Proc. Fifth Berkeley Symposium on Math. Statistics and Probability (University of California Press, California, 1961), Vol. II, Part 1, p. 145–161.

[2](a) Cécile Morette DeWitt, Commun. Math. Phys. 28, 47–67 (1972); (b) Cécile Morette DeWitt, Commun. Math. Phys. 37, 63–81 (1974).

[3]S. Albeverio and R. Hoegh-Krohn, "Mathematical Theory of Feynman Path Integrals," University of Oslo Preprint (1974).

[4](a) K. Hepp, Commun. Math. Phys. 35, 265–77 (1974); (b) V.P. Maslov, Zh. Vychisl. Mat. 1, 638–63 (1961); (c) V.P. Maslov, Zh. Vychisl. Mat. 1, 113–28 (1961); (d) V.P. Maslov, *Théorie des perturbations et methods asymtotiques* (Dunod, Paris, 1972).

[5]R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

[6]R.H. Cameron, J. Math. and Phys. 39, 126–40 (1960).

[7]In fact, γ is the unique continuous function on $(0,t)$ with weak derivative $d\gamma_w/d\tau$ and $\gamma(t)=0$.

[8]The separability of H implies that if μ exists it is uniquely determined by f.

[9]K. R. Parthasaraty, *Probability Measures on Metric Spaces* (Academic, New York, 1967). See Theorem 1.1, p. 57.

[10]J. Dieudonné, *Foundations of Modern Analysis* (Academic, New York, 1969), Vol. 1, Chap. X. 6.

[11]I. M. Gelfand and A.M. Yaglom, J. Math. Phys. 1, 48–69 (1960).

[12]V.P. Maslov, Ref. 4 (c), Sec. 4.

[13]V.P. Maslov, Ref. 4 (b).

[14]For a new and elegant approach to this problem see Cécile Morette DeWitt. "The Semi-classical Expansion," University of Texas Preprint.

[15]In a private communication Dr. Alan Carey of the University of Adelaide has recently advocated greater use of reproducing kernel techniques in quantum field theory.

Quantization problem and variational principle in the phase-space formulation of quantum mechanics*

Leon Cohen†

Department of Physics and Astronomy, University of North Carolina, Chapel Hill, North Carolina 27514
(Received 3 May 1976)

The problem of quantization in the phase-space formulation of quantum mechanics is considered. An integral equation for the phase-space eigenfunctions is derived which is equivalent to the standard eigenvalue equation for a quantum mechanical operator. A differential form is also given. A variational principle is derived for quasiprobability distributions. It is shown that the expected value of the classical Hamiltonian calculated with a trial quasiprobability distribution will be greater than the ground state energy only if the distribution is chosen from a certain class of functions. The notion of ψ-representability is introduced to classify these functions. They represent distributions which correspond to possible quantum mechanical states. Also, a general relation is given between different distribution functions.

1. INTRODUCTION

The phase space formulation of quantum mechanics has been used advantageously in a number of different fields. The basic idea is to calculate expectation values via phase space integration rather than through the operator formalism of quantum mechanics. The phase space most commonly used is that of position and momentum where both are considered as ordinary variables rather than operators. To accomplish this one introduces a probability distribution function $F(q,p)$ and a classical function $H(q,p)$ such that the quantum mechanical result[1]

$$\langle \mathbf{H}\rangle = \int \psi^* \mathbf{H}\psi, \tag{1.1}$$

yields the same value as phase space averaging

$$\langle \mathbf{H}\rangle = \int\int H(q,p)\, F(q,p)\, dp\, dp, \tag{1.2}$$

where $\mathbf{H}$ is the quantum mechanical operator corresponding to the classical function $H(q,p)$.

The probability distribution function has to be chosen so that it yields the correct quantum mechanical marginal probability distributions of position and momentum,

$$\int F(q,p)\, dp = |\psi(q)|^2, \tag{1.3}$$

$$\int F(q,p)\, dq = |\phi(p)|^2. \tag{1.4}$$

As $F(q,p)$ does not, in general, behave as a proper probability distribution (e.g., they may be negative or imaginary) they are often called "quasiprobabilities."

An explicit formula for the set of all possible distribution functions which satisfy Eqs. (1.3) and (1.4) has been given by the author,[2]

$$F(q,p) = \frac{1}{4\pi^2}\int\int\int \exp(-i\theta q - i\tau p + i\theta u) \times f(\theta,\tau)\, \psi^*(u - \tfrac{1}{2}\tau\hbar)\, \psi(u + \tfrac{1}{2}\tau\hbar)\, d\theta\, d\tau\, du, \tag{1.5}$$

where $f(\theta,\tau)$ is any function satisfying

$$f(0,\tau) = f(\theta,0) = 1. \tag{1.6}$$

There have been many particular distributions given. The most commonly used distribution is that by Wigner,[3]

$$F(q,p) = \frac{1}{2\pi}\int \psi^*(q - \tfrac{1}{2}\tau\hbar)\exp(-i\tau p)\, \psi(q + \tfrac{1}{2}\tau h)\, d\tau \tag{1.7}$$

which can be obtained from Eq. (1.5) by taking $f(\theta,\tau) = 1$. It can be derived by using the Weyl rule of association.[4,5]

Another commonly used distribution is the case where $f(\theta,\tau) = \cos\frac{1}{2}\theta\tau h$. It was given by Takabayasi[6] [but not in the form of Eq. (1.5)] and the unsymmetrized version [$f = \exp(\frac{1}{2}i\theta\tau\hbar)$] was used by Von Roos[7] to study quantum plasmas. Margenau and Hill[8] and Mehta[9] have derived this distribution by using the symmetrization rule of association.

To insure that Eq. (1.1) and (1.2) yield the same result the following relationship must hold between the quantum mechanical operator $\mathbf{H}$ and the classical function $H(q,p)$:

$$\begin{aligned}\mathbf{H} &= \int\int \gamma(\theta,\tau) f(\theta,\tau)\exp(i\theta\mathbf{q} + i\tau\mathbf{p})\, d\theta\, d\tau \\ &= \int\int \exp(i\theta\tau\hbar/2)\,\gamma(\theta,\tau) f(\theta,\tau)\exp(i\theta\mathbf{q})\exp(i\tau\mathbf{p})\, d\theta\, d\tau,\end{aligned} \tag{1.8}$$

where

$$\gamma(\theta,\tau) = \frac{1}{4\pi^2}\int\int H(q,p)\exp(-i\theta q - i\tau p)\, dq\, dp. \tag{1.9}$$

The problem of formulating the quantization problem in the phase space representation has been considered by a number of authors. Moyal[5] has derived an equation, for the case of the Wigner distribution, which the phase space eigenfunctions must satisfy. In the next section we present a simple and straightforward formulation of the quantization problem for an arbitrary quasiprobability distribution.

In Sec. 3 we derive a variational principle in the phase space formulation similar to the one in quantum mechanics. It will be shown that not all trial functions are acceptable.

Before proceeding we show how different distributions are related to each other. Suppose we have two distributions $F_1(q,p)$ and $F_2(q,p)$ characterized by $f_1(\theta,\tau)$ and $f_2(\theta,\tau)$. The moment generating function[2] is defined by

$$M(\theta,\tau) = \int\int F(q,p)\exp(i\theta q + i\tau p)\, dq\, dp, \tag{1.10}$$

and from Eq. (1.5) equals

$$M(\theta,\tau) = f(\theta,\tau)\int \psi^*(u - \tfrac{1}{2}\tau h)\exp(i\theta u)\,\psi(u + \tfrac{1}{2}\tau h)\, du. \tag{1.11}$$

Hence, for two different distributions

$$\frac{M_1(\theta,\tau)}{f_1(\theta,\tau)}=\frac{M_2(\theta,\tau)}{f_2(\theta,\tau)}, \qquad (1.12)$$

or in terms of the distribution

$$F_1(q,p)=\frac{1}{4\pi^2}\int \frac{f_1(\theta,\tau)}{f_2(\theta,\tau)} \exp(i\theta(q'-q)+i\tau(p'-p)) \times F_2(q',p')\,d\theta\,d\tau\,dq'\,dp' \qquad (1.13)$$

which can also be written as

$$F_1(q,p)=\left[f_1\left(i\frac{\partial}{\partial q}, i\frac{\partial}{\partial p}\right)\Big/ f_2\left(i\frac{\partial}{\partial q}, i\frac{\partial}{\partial p}\right)\right]F_2(q,p). \qquad (1.14)$$

For the case of F_2 being the Wigner distribution and $f_1=\cos\frac{1}{2}\theta\tau h$, this relation has been previously derived.[6,7,9]

We further point out that, in general, different correspondence rules give different quantum mechanical operators.[10] That is, for a given $H(q,p)$, different choices of $f(\theta,\tau)$ will lead, using relations (1.8) and (1.9), to different **H**. In the case of where $H(q,p)$ is the sum of functions which are only functions of q and p, then the same **H** will result. Two different $H(q,p)$'s lead to the same H if their Fourier transforms are related by

$$\gamma_1(\theta,\tau)f_1(\theta,\tau)=\gamma_2(\theta,\tau)f_2(\theta,\tau), \qquad (1.15)$$

or

$$H_1(q,p)=\frac{1}{4\pi^2}\int \frac{f_2(\theta,\tau)}{f_1(\theta,\tau)} H_2(q',p') \times\exp(i\theta(q-q')+i\tau(p-p'))\,d\theta\,d\tau\,dq'\,dp'. \qquad (1.16)$$

2. QUANTIZATION

For the Wigner distribution Moyal[5] has introduced phase space eigenfunctions to formulate the quantization problem. For a general quasiprobability distribution these eigenfunctions are defined[2] as

$$h_{nm}(q,p)=\frac{1}{4\pi^2}\int \exp(-i\theta q-i\tau p+i\theta u)f(\theta,\tau) \times\varphi_n^*(u-\tfrac{1}{2}\tau\hbar)\,\varphi_m(u+\tfrac{1}{2}\tau\hbar)\,d\theta\,d\tau\,du, \qquad (2.1)$$

where φ_n is a complete orthogonal set. The h_{nm}'s are then also a complete orthogonal set in the space of q and p and satisfy the following relations[2,5]:

$$\iint h_{nm}h_{n'm'}\,dq\,dp=\frac{1}{2\pi\hbar}\,\delta_{nn'}\,\delta_{mm'}, \qquad (2.2)$$

$$\sum_{nm} h_{nm}(q,p)\,h_{nm}(q',p')=\frac{1}{2\pi\hbar}\,\delta(q-q')\,\delta(p-p'), \qquad (2.3)$$

$$\iint h_{nm}(q,p)\,dq\,dp=\delta_{nm}, \qquad (2.4)$$

$$\sum_n h_{nn}(q,p)=\frac{1}{2\pi\hbar}. \qquad (2.5)$$

To obtain the equation which the eigenfunctions must satisfy we shall start with

$$\mathbf{H}\varphi_m=E_m\varphi_m \qquad (2.6)$$

and convert it into an equation for h_{nm}. For clarity and simplicity of algebra we first treat the case of $f=1$ and then use Eq. (1.13) and (1.15) to obtain the equation for the general case.

Substituting **H** as given by Eq. (1.8) (taking $f=1$) into Eq. (2.6) we have

$$\iint \exp(i\theta\tau\hbar/2)\,\gamma(\theta,\tau)\exp(i\theta q)\,\varphi_m(q+\tau\hbar)=E_m\varphi_m(q). \qquad (2.7)$$

Letting q go into $q+\frac{1}{2}\tau'h$, multiplying by $\exp(-i\tau'p)\,\varphi_n^*(q-\frac{1}{2}\tau'h)$, and integrating with respect to τ', we have

$$\int \exp(i\theta(\tau+\tau')\hbar/2)\,\gamma(\theta,\tau)\exp(i\theta q-i\tau'p)\,\varphi_n^*(q-\tfrac{1}{2}\tau'\hbar) \times\varphi_m(q+\tfrac{1}{2}\tau'\hbar+\tau\hbar)\,d\theta\,d\tau\,d\tau'=2\pi\,E_m h_{nm}(q,p), \qquad (2.8)$$

but from Eq. (2.1),

$$\varphi_n^*(x)\,\varphi_m(y)=\int \exp(i(y-x)p/\hbar)\,h_{nm}((x+y)/2,p)\,dp \qquad (2.9)$$

and substituting into Eq. (2.8) we obtain

$$\int \exp[\tfrac{1}{2}i\theta\tau\hbar+i\theta q+i\tau p'+i\tau'(p'-p+\tfrac{1}{2}\theta\hbar)]\,\gamma(\theta,\tau)\, h_{nm}(q+\tfrac{1}{2}\tau\hbar,p')\,d\theta\,d\tau\,d\tau'\,dp'=2\pi\,E_m h_{nm}(q,p). \qquad (2.10)$$

Integrating with respect to τ' and then p' yields

$$\iint \gamma(\theta,\tau)\exp(i\theta q+i\tau p)\,h_{nm}(q+\tfrac{1}{2}\tau\hbar,p-\tfrac{1}{2}\theta\hbar)\,d\theta\,d\tau =E_m h_{nm}(q,p). \qquad (2.11)$$

Using relations (1.13) and (1.15) the general case can now be derived. It is

$$\iint K(q,p;x,y)\,h_{nm}(x,y)\,dx\,dy=E_m h_{nm}(q,p), \qquad (2.12)$$

where

$$K(q,p;x,y)=\frac{1}{4\pi^2}\int \frac{\gamma(\theta,\tau)f(\theta,\tau)f(\theta',\tau')}{f(\theta+\theta',\tau+\tau')} \times\exp[i\theta'(q-x+\tfrac{1}{2}\tau\hbar)-i\tau'(p+y-\tfrac{1}{2}\theta\hbar)] \times\exp(i\theta x+i\tau y)\,d\theta\,d\tau\,d\theta'\,d\tau'. \qquad (2.13)$$

This can also be written as an operator equation

$$f\left(-i\frac{\partial}{\partial q_H},-i\frac{\partial}{\partial p_H}\right)f\left(i\frac{\partial}{\partial q}+i\frac{\partial}{\partial q_H}, i\frac{\partial}{\partial p}+i\frac{\partial}{\partial p_H}\right) \times\left[f\left(i\frac{\partial}{\partial q},i\frac{\partial}{\partial p}\right)\right]^{-1} H\left(q+\tfrac{1}{2}i\frac{\partial}{\partial p},\,p-\tfrac{1}{2}i\frac{\partial}{\partial q}\right) \times h_{nm}(q,p)=E_m h_{nm}(q,p). \qquad (2.14)$$

For the Wigner case,

$$H\left(q+\tfrac{1}{2}i\frac{\partial}{\partial p},\,p-\tfrac{1}{2}i\frac{\partial}{\partial q}\right)h_{nm}(q,p)=E_m h_{nm}(q,p). \qquad (2.15)$$

Moyal's[5] equation, for the case $f=1$ can be obtained from Eq. (2.11) or (2.15).

Eq. (2.14) can be considered as the quantization equation in the phase space formulation analogous to the operator equation, Eq. (2.6). It is likely that for particular problems, solving Eq. (2.14) with a judicious choice of f may be simpler then solving Eq. (2.6).

We now give a procedure for obtaining the state function when a phase space distribution is given. In the case of a phase space eigenfunction one would take $F(q,p)=h_{nm}(q,p)$. It can be readily verified, that up to an arbitrary constant phase factor, the state function is given by

$$\psi(q)=\frac{1}{2\pi R(q_0)}\times\int\frac{F(x,y)\exp[i\theta(x-(q+q_0)/2)+i(q-q_0)y/\hbar]}{f(\theta,(q-q_0)/\hbar)}\times d\theta\,dx\,dy, \tag{2.16}$$

where

$$R^2(q_0)=\int F(q_0,p)\,dp, \tag{2.17}$$

and q_0 is any number.

3. VARIATIONAL PRINCIPLE

The variational principle of standard quantum mechanics states that for any normalized wavefunction the expectation value of the Hamiltonian is greater than or equal to the ground state energy, E_0,

$$\langle\psi|\mathbf{H}|\psi\rangle\geq E_0. \tag{3.1}$$

We now ask whether the expectation value of the classical Hamiltonian obtained via phase space integration using an arbitrary equasiprobability distribution $F(q,p)$ also satisfies

$$\int\int H(q,p)F(q,p)\,dq\,dp\geq E_0. \tag{3.2}$$

In general, Eq. (3.2) is not true for an arbitrary $F(q,p)$, but if one choses an $F(q,p)$ from a certain class of function (discussed below) then indeed Eq. (3.2) holds for members of that class. We call such functions ψ representable. ψ representable distributions are those which can be obtained from some quantum mechanical state function. If a distribution is not ψ representable then there exists no corresponding quantum state and the use of such an F could lead to erroneous results.

We note that we are considering the full N-body distribution function. A much more difficult, and as yet unsolved problem, is to determine the conditions a reduced distribution must satisfy if it came from a proper N-body quantum mechanical distribution. In the density matrix formulation this problem is known as N representability.[11]

Since Eqs. (1.1) and (1.2) give the same results if Eq. (1.5) holds, it is clear that an F is ψ representable if there is some function $\psi(x)$ so that indeed (1.5) is satisfied. Otherwise there is no quantum state. In particular, from Eq. (1.5) we have that

$$\psi^*(x)\,\psi(y)=\frac{1}{2\pi}\int\frac{F(q,p)\exp[i\theta(q-(x+y)/2)+i(y-x)p/\hbar]}{f(\theta,(y-x)/2)}\times d\theta\,dq\,dp. \tag{3.3}$$

Hence, to test whether a given $F(q,p)$ is ψ representable one may evaluate the right-hand side of Eq. (3.3) and examine whether the result can be written in the form $\psi^*(x)\,\psi(y)$ for *some* function ψ. For the Wigner distribution, Eq. (3.3) becomes

$$\psi^*(x)\,\psi(y)=\int F((x+y)/2,\,p)\exp(i(y-x)p/\hbar))\,dp. \tag{3.4}$$

Of course, once it is established that F is ψ representable it is clear that (3.2) follows because (3.1) is true and Eqs. (1.1) and (1.2) give the same answer. But to illustrate more clearly where the breakdown occurs if F is not ψ representable, we give a derivation of (3.2) totally within the phase space formulation. We use the eigenfunction equations derived in Sec. 2.

Expanding $F(q,p)$ in terms of the eigenfunctions we have

$$F(q,p)=\sum_{nm}A_{nm}h_{nm}(q,p), \tag{3.5}$$

$$A_{nm}=2\pi\hbar\int F(q,p)\,h^*_{nm}(q,p)\,dq\,dp. \tag{3.6}$$

If $F(q,p)$ is normalized to one, then using Eq. (2.5),

$$\int F(q,p)\,dq\,dp=1=\sum_n A_{nn}. \tag{3.7}$$

Operating on (3.5) with the integral operator as defined by Eq. (2.13) and integrating with respect to q and p, yields

$$\frac{1}{4\pi^2}\int\frac{\gamma(\theta,\tau)f(\theta,\tau)f(\theta',\tau')}{f(\theta+\theta);\tau+\tau')}\exp[i(\theta+\theta')x+i(\tau+\tau')y]\times\exp[-i\theta'(q+\tfrac{1}{2}\tau\hbar)-i\tau'(p-\tfrac{1}{2}\theta\hbar)]\,d\theta\,d\theta'\,d\tau\,d\tau'\,dx\,dy\,dq\,dp=\sum_n E_nA_{nn}. \tag{3.8}$$

Using Eqs. (1.8) and (1.5), this becomes

$$\int\int H(q,p)F(q,p)\,dq\,dp=\sum E_nA_{nn}. \tag{3.9}$$

For an arbitrary $F(q,p)$ the A_{nn}'s could become negative and the variational principle would not follow. But if $F(q,p)$ is of such a functional form that it can be expressed in the form given by Eq. (3.3) for some ψ, then indeed the A_{nn}'s are always positive,

$$A_{nn}=\frac{\hbar}{2\pi}\int\psi^*(q-\tfrac{1}{2}\tau\hbar)\,\psi(q+\tfrac{1}{2}\tau h)\,\varphi_n(q-\tfrac{1}{2}\tau'\hbar)\,\varphi^*_n(q+\tfrac{1}{2}\tau'\hbar)\exp(-i(\tau-\tau')p)\,d\tau\,d\tau'\,dq\,dp=\left|\int\psi(q)\,\varphi^*_n(q)\,dq\right|^2\geq 0. \tag{3.10}$$

The usual arguments therefore lead to Eq. (3.2).

4. CONCLUSION

The main reason for the use of the phase space formulation is that very often it is mathematically easier to use than the operator formalism of quantum mechanics. Also since the formalism is "similar" to that of classical phase space it very often suggests methods of approximation or expansion used in classical mechanics. This is particularly true in the application to quantum statistical mechanics. We have shown that one must be particularly careful in assuming ψ representability of the distribution function. Also, as mentioned previously it is possible that in certain cases the quantization problem may be more tractable if one attempts to solve the eigenvalue problem in the phase space formalism. Since the phase space expectation values are sometimes easier to carry out in the operator formalism, the variational principle may possibly also be applied to real problems with profit.

In conclusion we would like to give a simple problem illustrating ψ representability. Consider the estimation of the ground state energy of the harmonic oscillator using the trial function,

$$F_1(q,p)=\frac{\sqrt{2}}{\pi\hbar}\exp(-\alpha^2q^2)-\frac{2}{\hbar^2\alpha^2}p^2, \qquad (4.1)$$

where α is the parameter to be varied. Calculating the energy with the classical Hamiltonian,

$$\langle H\rangle=\iint(p^2/2m+\tfrac{1}{2}m\omega^2q^2)F_1(q,p)\,dq\,dp$$
$$=\frac{1}{4}\left[\frac{m\omega^2}{\alpha^2}+\frac{\hbar^2}{2m}\alpha^2\right]. \qquad (4.2)$$

Minimizing with respect to α yields

$$\alpha^4=2\omega^2m^2/\hbar^2, \qquad (4.3)$$

and substituting into Eq. (4.3) we have for the estimated ground state energy

$$1/\sqrt{2}\,\tfrac{1}{2}\hbar\omega, \qquad (4.4)$$

which is *lower* than the ground state energy of the harmonic oscillator.

The reason for the failure is that $F_1(q,p)$ is not ψ representable. It is straightforward to show that for $F_1(q,p)$ there does not exist any function ψ such that Eq. (4.1) is satisfied. Moreover, as can be readily verified, the A_{nm}'s are not all positive for this case.

On the other hand, if we consider the trial function

$$F_2(q,p)=(\alpha\beta/\pi)\exp(-\alpha^2q^2)-\beta^2p^2, \qquad (4.5)$$

then direct calculation of the right-hand side of Eq. (3.3) shows that ψ representability forces us to take

$$\beta=1/\hbar\alpha. \qquad (4.6)$$

$F_2(q,p)$ now yields the correct ground state energy.

*Work done during tenure of a grant from The City University Faculty Research Award Program.

†Permanent address: Hunter College of the City University of New York, New York, N.Y. 10021.

[1]We consider one-dimensional systems. Generalization to higher dimensions is straightforward. All integrals go from $-\infty$ to ∞.

[2]L. Cohen, J. Math. Phys. **7**, 781 (1966).
[3]E. Wigner, Phys. Rev. **40**, 749 (1932).
[4]H. J. Groenewold, Physica **12**, 405 (1946).
[5]J. E. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).
[6]T. Takabayasi, Prog. Theor. Phys. (Kyoto) **11**, 341 (1954).
[7]O. von Roos, Phys. Rev. **119**, 1174 (1960).
[8]H. Margenau and R. N. Hill, Prog. Theor. Phys. **26**, 722 (1961).
[9]C. L. Mehta, J. Math. Phys. **5**, 677 (1964).
[10]J. R. Shewell, Am. J. Phys. **27**, 16 (1959).
[11]A. J. Coleman, Rev. Mod. Phys. **35**, 3, 668 (1963).

Prolongation structures and a generalized inverse scattering problem

Hedley C. Morris

School of Mathematics, Trinity College Dublin, Dublin 2, Ireland
(Received 15 September 1975)

The prolongation structure method of Wahlquist and Estabrook is used to determine a generalized inverse scattering problem for the equation $u_{tt}=u_{xx}+6(uu_x)_x+u_{xxxx}$ which describes the motion of shallow-water waves under gravity. The relevant Gel'fand–Levitan equation is solved for the single soliton solutions.

1. INTRODUCTION

Since its original discovery by Gardener, Green, Kruskal, and Muira[1] the inverse scattering method has been considerably developed. There are now over a score of physically significant equations for which exact solutions can be determined by the method. The search for such equations and inverse scattering problems continues and recently Wahlquist and Estabrook[2] introduced a new approach to the determination of such problems. The examples considered by Wahlquist and Estabrook led to the standard Schrödinger equation form of the inverse scattering problem. In this paper, we wish to show that their method can be used to determine alternative generalized eigenvalue problems of non-Schrödinger type which we will refer to as generalized inverse scattering problems. We will consider the equation

$$\tfrac{3}{4}u_{tt}+\tfrac{1}{4}(u_{xxxx})+\tfrac{3}{2}(uu_x)_x=0. \tag{1.1}$$

The numerical values of the coefficients can be altered by scaling of both dependent and independent variables. Our choice is that taken by Zakharov and Shabat[3] in an alternative approach to the same problem.

If we introduce $U\overset{\text{def}}{=}u-\frac{1}{6}$ and $\tau=i/\sqrt{3}t$, the equation becomes

$$U_{\tau\tau}-U_{xx}-U_{xxxx}-6(UU_x)_x=0 \tag{1.2}$$

which describes the motion of shallow-water waves under gravity.[4] For details of the prolongation structure method, which involves the theory of differential forms,[5] we refer to Wahlquist and Estabrook.[2]

2. THE ASSOCIATED SET OF FORMS

If we introduce the notation

$$u_x=p, \tag{2.1}$$

$$p_x=r, \tag{2.2}$$

and the potential w defined by

$$w_x=-\tfrac{3}{4}u_t, \tag{2.3}$$

$$w_t=\tfrac{1}{4}(u_{xxx}++6(uu_x)=\tfrac{1}{4}(r_x+6up) \tag{2.4}$$

then Eqs. (2.1)–(2.4) are completely equivalent to the higher-order equation (1.1). These first-order partial differential equations can be associated with the set of 2-forms $\alpha_1,\alpha_2,\alpha_3,\alpha_4$ defined by

$$\alpha_1=du\wedge dt-p\,dx\wedge dt, \tag{2.5}$$

$$\alpha_2=dp\wedge dt-r\,dx\wedge dt, \tag{2.6}$$

$$\alpha_3=du\wedge dx-\tfrac{4}{3}\,dw\wedge dt, \tag{2.7}$$

$$\alpha_4=dw\wedge dx+\tfrac{3}{2}up\,dx\wedge dt+\tfrac{1}{4}\,dr\wedge dt. \tag{2.8}$$

These 2-forms comprise a closed ideal and consequently by Cartans[5] theory are completely equivalent to Eqs. (2.1)–(2.4). If $\tilde{\alpha}_i$ $(i=1,\ldots,4)$ are the forms obtained by sectioning into a solution manifold of (2.1)–(2.4) then

$$\tilde{\alpha}_i=0\quad(i=1,\ldots,4). \tag{2.9}$$

The method proceeds by seeking sets of n 1-forms $\Omega^k=d\xi^k+F^k\,dx+G^k\,dt$ $(k=1,\ldots,n)$ having the property that the prolonged ideal $(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\Omega^1,\ldots,\Omega^n)$ is closed. This means that we must be able to find a set of n^2 1-forms η_i^k and a set of $4n$ functions f^{ki} such that

$$d\Omega^k=\sum_{i=1}^{4}f^{ki}\alpha_i+\sum_{i=1}^{n}\eta_i^k\wedge\Omega^i. \tag{2.10}$$

That requirement leads us to the equation

$$-pG_u^k-rG_p^k+(\tfrac{3}{2}up)F_w^k+[F,G]^k=0, \tag{2.11}$$

where we have introduced the notation

$$[F,G]^k=(F^iG^k_{,\xi^i}-G^iF^k_{,\xi^i}) \tag{2.12}$$

which has all the normal properties of a Lie bracket. For notational convenience we will cease to show the suffices on F and G and related quantities for the remainder of this work.

General consistency requirements on Eqs. (2.10)–(2.11) show that F and G must have the forms

$$F=X_1+uX_2+wX_3+4uwX_{12}+u^2X_{13}, \tag{2.13}$$

$$G=X_4+uX_5+u^2X_6-\tfrac{4}{3}wX_2+pX_8+\tfrac{1}{4}rX_3-\tfrac{8}{3}w^2X_{12}-\tfrac{8}{3}uwX_{13}+urX_{12}+2u^3X_{12}-\tfrac{1}{2}p^2X_{12}. \tag{2.14}$$

For the purposes of this paper we do not require the most general form and for simplicity put $X_{12}=X_{13}=0$. Substitution of (2.13) and (2.14) into (2.11) then yield the Lie bracket relations

$$\begin{aligned}&[X_1,X_8]=X_5,\\&[X_2,X_8]=2X_6-\tfrac{3}{2}X_3,\\&[X_1,X_4]=0,\\&[X_2,X_6]=0,\end{aligned} \tag{2.15}$$

$$[X_3, X_5] = 0,$$
$$[X_3, X_8] = 0,$$
$$[X_1, X_3] = 4X_8,$$

together with the relations

$$\begin{aligned} &[X_1, X_5] + [X_2, X_4] = 0, \\ &[X_1, X_6] + [X_2, X_5] = 0, \\ &[X_3, X_4] = \tfrac{4}{3}[X_1, X_2]. \end{aligned} \tag{2.16}$$

In order to determine a representation of this algebraic structure we will complete it into a Lie algebra. First we note that some simplification results if we make the identification

$$X_6 = kX_3 \tag{2.17}$$

which is consistent with the existing relations (2.15).

We introduce new generators X_{10} and X_{11} defined by

$$[X_1, X_2] = X_{10}, \tag{2.18}$$

$$[X_1, X_5] = X_{11} \tag{2.19}$$

in order to simplify the constraining relations (2.16). If we then ask that $(X_1, X_2, X_3, X_4, X_5, X_8, X_{10}, X_{11})$ close under the Lie bracket operation to form a Lie algebra, we find that when $k = 9/16$ this is possible and the following algebra results

$$\begin{aligned}
&[X_1, X_2] = X_{10}, \quad [X_1, X_3] = 4X_8, \\
&[X_1, X_5] = X_{11}, \quad [X_1, X_8] = X_5, \\
&[X_1, X_{10}] = \tfrac{3}{4}X_1 + \tfrac{9}{4}\mu X_3, \quad [X_1, X_{11}] = \tfrac{3}{2}X_4 - 3\mu X_2, \\
&[X_2, X_4] = -X_{11}, \quad [X_2, X_5] = -\tfrac{9}{4}X_8, \quad [X_2, X_8] = -\tfrac{3}{8}X_3, \\
&[X_2, X_{10}] = -\tfrac{3}{4}X_2, \quad [X_2, X_{11}] = -\tfrac{9}{4}X_8, \\
&[X_2, X_8] = -\tfrac{3}{8}X_3, \\
&[X_2, X_{10}] = -\tfrac{3}{4}X_2, \quad [X_2, X_{11}] = -\tfrac{9}{4}X_5, \\
&[X_3, X_4] = \tfrac{4}{3}X_{10}, \quad [X_3, X_{10}] = -\tfrac{3}{2}X_3, \quad [X_3, X_{11}] = X_2, \\
&[X_4, X_5] = -3\mu X_8, \quad [X_4, X_8] = -\tfrac{3}{4}\mu X_3 - \tfrac{1}{4}X_1, \\
&[X_4, X_{10}] = \tfrac{3}{2}X_4 - 3X_2, \quad [X_4, X_{11}] = -3\mu X_5, \\
&[X_5, X_8] = \tfrac{1}{4}X_2, \quad [X_5, X_{11}] = -\tfrac{9}{16}X_1 - \tfrac{9}{16}X_3, \\
&[X_8, X_{10}] = -\tfrac{3}{4}X_8, \quad [X_8, X_{11}] = -\tfrac{1}{4}X_{10}, \\
&[X_{10}, X_{11}] = -\tfrac{3}{4}X_{11}
\end{aligned} \tag{2.20}$$

with all other bracket relations zero. The constant μ is arbitrary and so we have a single parameter family of Lie algebras associated with our original equations. We will see shortly that μ is the eigenvalue in an inverse scattering equation which can be associated to the equation.

3. A THREE-DIMENSIONAL REPRESENTATION

We can determine the following three-dimensional representation of the algebra:

$$\begin{aligned}
&X_1 = -\mu\xi^1 b_3 - (\xi^2 b_1 + \xi^3 b_2), \\
&X_2 = \tfrac{3}{4}(\xi^1 b_2 + \xi^2 b_3), \\
&X_3 = \xi^1 b_3, \\
&X_4 = \xi^3 b_1 + \mu(\xi^1 b_2 + \xi^2 b_3), \\
&X_5 = \tfrac{1}{4}(\xi^1 b_1 + \xi^3 b_3 - 2\xi^2 b_2), \\
&X_8 = \tfrac{1}{4}(\xi^1 b_2 - \xi^2 b_3), \\
&X_{10} = \tfrac{3}{4}(\xi^1 b_1 - \xi^3 b_3), \\
&X_{11} = \tfrac{3}{4}(\xi^3 b_2 - \xi^2 b_1),
\end{aligned} \tag{3.1}$$

where $b_i = \partial/\partial\xi^i$.

The corresponding Pfaffian forms are

$$\begin{aligned}
&\Omega^1 = d\xi^1 - \xi^2\,dx + (\xi^3 + \tfrac{1}{4}u\xi^1)\,dt, \\
&\Omega^2 = d\xi^2 - (\xi^3 - \tfrac{3}{4}u\xi^1)\,dx - (\tfrac{1}{2}u\xi^2 + [(w-\mu) - \tfrac{1}{4}p]\,\xi^1)\,dt, \\
&\Omega^3 = d\xi^3 + (\tfrac{3}{4}u\xi^2 + (w-\mu)\,\xi^1)\,dx \\
&\qquad + [(\tfrac{1}{4}r + \tfrac{9}{16}u^2)\,\xi^1 - (w - \mu + \tfrac{1}{4}p)\,\xi^2 + \tfrac{1}{4}u\xi^3)]\,dt.
\end{aligned} \tag{3.2}$$

On a solution manifold of the prolonged ideal $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \Omega^1, \Omega^2, \Omega^3)$

$$\tilde{\Omega}^1 = \tilde{\Omega}^2 = \tilde{\Omega}^3 = 0, \tag{3.3}$$

which yields the equations

$$\xi^1_x = \xi^2, \tag{3.4}$$

$$\xi^2_x = \xi^3 - \tfrac{3}{4}u\xi^1, \tag{3.5}$$

$$\xi^3_x = -\tfrac{3}{4}u\xi^2 - (w-\mu)\,\xi^1, \tag{3.6}$$

$$\xi^1_t = -\xi^3 - \tfrac{1}{4}u\xi^1, \tag{3.7}$$

$$\xi^2_t = \tfrac{1}{2}u\xi^2 + (w - \mu - \tfrac{1}{4}p)\,\xi^1, \tag{3.8}$$

$$\xi^3_t = -(\tfrac{1}{4}r + \tfrac{9}{16}u^2)\,\xi^1 + (w - \mu + \tfrac{1}{4}p)\,\xi^2 - \tfrac{1}{4}u\xi^3. \tag{3.9}$$

Eliminating ξ^2 and ξ^3 from (3.4)–(3.6), we obtain the equation

$$\xi^1_{xxx} + \tfrac{3}{2}u\xi^1_x + (\tfrac{3}{4}u_x + w)\,\xi^1 = \mu\xi^1. \tag{3.10}$$

If we couple this with Eq. (3.7) which can be written

$$\xi^1_t + \xi^1_{xx} + u\xi^1 = 0, \tag{3.11}$$

then these two equations (3.10) and (3.11) specify a generalized inverse scattering problem associated with the original equation (1.1).

4. THE GEL'FAND-LEVITAN EQUATION AND SOLITON SOLUTIONS

The Gel'fand-Levitan equation appropriate to this generalized scattering problem is

$$F(x, s\,|\,t) + K(x, s\,|\,t) + \int_x^\infty K(x, p\,|\,t)\,F(p, s\,|\,t)\,dp = 0, \tag{4.1}$$

where $F(x, s\,|\,t)$ is a solution to the equations

$$\frac{\partial^3 F}{\partial x^3}(x, s\,|\,t) + \frac{\partial^3 F}{\partial s^3}(x, s\,|\,t) = 0, \tag{4.2}$$

$$\frac{\partial}{\partial t}F(x, s\,|\,t) + \frac{\partial^2}{\partial x^2}F(x, s\,|\,t) - \frac{\partial^2 F}{\partial s^2}F(x, s\,|\,t) = 0. \tag{4.3}$$

The solution function $u(x, t)$ of (1.1) is related to the kernel $K(x, s\,|\,t)$ by

$$u(x, t) = 2\frac{d}{dx}K(x, x\,|\,t). \tag{4.4}$$

A solution to (4.1) and (4.3) is given by

$$F(x,s\,|\,t)=\alpha\exp[-q(x-ws)-q^2(1-w^2)\,t], \qquad (4.5)$$

where α, q are constants and w is a cube root of unity. We will assume that $w\neq 1$ as $w=1$ leads to the trivial zero solution for $u(x,t)$.

Substituting this solution (4.5) into (4.1), we easily obtain the following form for the kernel $K(x,s\,|\,t)$:

$$K(x,s\,|\,t)=\frac{-\alpha\exp[-q(x-ws)-q^2(1-w^2)\,t]}{[1+[\alpha/q(1-w)]\exp[-q(1-w)x-q^2(1-w^2)\,t]}. \qquad (4.6)$$

This gives

$$K(x,x\,|\,t)=\left[\frac{-\alpha\exp(-Q\xi)}{1+(\alpha/Q)\exp(-Q\xi)}\right], \qquad (4.7)$$

where

$$\xi=x+[(1+w)/(1-w)]\,Qt \text{ and } Q=q(1-w). \qquad (4.8)$$

Hence, we obtain

$$u(x,t)=-2\frac{d}{dx}\left[\frac{\alpha\exp(-Q\xi)}{1+(\alpha/Q)\exp(-Q\xi)}\right]$$

$$=\frac{2\alpha Q\exp(-Q\xi)}{[1+(\alpha/Q)\exp(-Q\xi)]^2}. \qquad (4.9)$$

If we define η by

$$\alpha/Q=\exp(-Q\eta) \qquad (4.10)$$

then we can write (4.9) as

$$u(x,t)=\tfrac{1}{2}Q^2\operatorname{sech}^2\tfrac{1}{2}Q(\xi+\eta). \qquad (4.11)$$

As w is a cube root of unity

$$(1+w)/(1-w)=(\pm i\cot\pi/3)=\pm i/\sqrt{3}\,t$$

and so the single soliton solutions are

$$u^{\pm}(x,t)=\tfrac{1}{2}Q^2\operatorname{sech}^2\tfrac{1}{2}Q[(x\pm(i/\sqrt{3})\,Qt)+\eta]. \qquad (4.12)$$

The inverse scattering problem for Eq. (1.2) is obtained by making the substitutions $U=u-\frac{1}{6}$, $\tau=i/\sqrt{3}$ into (3.10) and (3.11). An almost identical solution to the relevant Gel'fand-Levitan equation gives the left and right propagating soliton solutions

$$U^{\pm}(x,t)=\tfrac{1}{2}Q^2\operatorname{sech}^2\tfrac{1}{2}Q[(x\pm Q(1+Q^2\tau)^{1/2})]. \qquad (4.13)$$

Similarly, using the linear suposition of the solutions for the kernel $F(x,s)$, the N-soliton solutions previously determined by Hirota[6] may be easily obtained.

[1]C.S. Gardener, J.M. Green, M. Kruskal, and R. Muira, Phys. Rev. Lett. **19**, 1095 (1967).

[2]H.D. Wahlquist and F.B. Estabrook, J. Math. Phys. **16**, 1 (1975).

[3]V.E. Zakharov and A.B. Shabat, Funct. Anal. Appl. **8**, 226 (1974).

[4]F. Ursell, Proc. Camb. Philos. Soc. **49**, 685 (1953).

[5]W. Slebodzinski, *Exterior Forms and Applications* (Polish Scientific Publishers, Warsaw, 1970).

[6]Ryogo Hirota, J. Math. Phys. **14**, 810 (1973).

Prolongation structures and nonlinear evolution equations in two spatial dimensions

H. C. Morris

School of Mathematics, Trinity College, Dublin, Ireland
(Received 1 December 1975)

The prolongation structure approach of Wahlquist and Estabrook is used to determine nonlinear evolution equations in two spatial dimensions for which an inverse scattering formulation exists. The equations of nonlinear wave–envelope interactions and the Kadomtsev–Petviashvilli–Dryuma equation are considered in detail.

1. INTRODUCTION

The success of the inverse scattering method in solving many two-dimensional problems of physical significance is by now well known. It is natural to wonder whether those equations which have been treated are capable of generalization to three or more dimensions in such a manner that a generalized form of the inverse scattering method may be applied to them also. Recently the prolongation structure method of Wahlquist and Estabrook[1] has shown that the study of certain ideals containing differential forms related to a nonlinear evolution equation is capable of determining the inverse scattering problem relevant to that equation, should it exist. The prolongation structure approach seems a natural one to use in order to extend existing results into higher dimensions. The purpose of this note is not to provide an exhaustive or general study but to illustrate with specific examples the ease and relative simplicity with which the method can do this without exhausting its generality. In Sec. 2. we discuss the general method and derive our basic equations labelled (2.7), (2.11), and (2.12). We then continue in Secs. 3 and 4 by showing how these equations can provide inverse scattering problems for the equation of nonlinear wave-envelope interactions in two spatial dimensions[2,3] and also the Kadomtsev—Petviashvili—Dryuma[4,5] equation. We note that a reverse procedure starting from the inverse scattering equations rather than the nonlinear evolution equation, and therefore complimentary to our own, has been recently published by Ablowitz and Haberman.[6]

2. EXTENDING AN EXISTING PROLONGATION STRUCTURE

Let us suppose that we have a two-dimensional evolution equation which can be expressed in terms of a closed set of 2-forms $\{\alpha_i\}$, $i=1,\ldots,N$, which possess a linear prolongation structure $\{\alpha_i, \Omega^\beta\}$, $i=1,\ldots,N$, $\beta=1,\ldots,M$, in which the 1-forms Ω^β are expressed

$$\Omega^\beta=\sum_{\alpha=1}^{M}(F^\beta_\alpha\, dx+G^\beta_\alpha\, dt)\,\zeta^\alpha+d\zeta^\beta. \tag{2.1}$$

This means that there exist MN function $f^{\beta i}$ and M^2 1-forms $\eta^\beta{}_\gamma$ such that

$$d\Omega^\beta=\sum_{i=1}^{N} f^{\beta i}\alpha_i+\sum_{\gamma=1}^{M}\eta^\beta{}_\gamma\wedge\Omega^\gamma. \tag{2.2}$$

Some of the forms $\{\alpha_i\}$, $i=1,\ldots,K$, essentially define the variables needed to reduce the nonlinear evolution equations to a set of first order partial differential equations. An example of this is $\alpha_1=du\wedge dt-p dx\wedge dt$ which defines p to be u_x in the sectioned ideal. We will refer to these forms as *linearization forms*. The remaining forms $\{\alpha_j\}$, $j=K+1,\ldots,N$, express the original equation directly and we will refer to these as the *dynamic forms* of the ideal. When we consider a higher dimensional form of an evolution equation, there will generally be additional linearizing forms and the dynamic forms will be modified by additional terms. In this paper we will consider the simplest possibility. This is the situation in which no additional linearizing forms are introduced by the generalization but simply modifications to the dynamic forms. This means that no derivatives of degree higher than one in the new spatial variable y will be present in the generalized equations. Thus the generalization we consider can be expressed by the linearizing forms

$$\bar\alpha_i=\alpha_i\wedge dy,\quad i=1,\ldots,K, \tag{2.3a}$$

and the dynamic forms

$$\bar\alpha_j=\alpha_j\wedge dy+\beta_j,\quad j=K+1,\ldots,N, \tag{2.3b}$$

where the β_j are a set of $(N-K)$ 3-forms. Consider the 2-forms $\bar\Omega^\beta$ of the form

$$\bar\Omega^\beta=\Omega^\beta\wedge dy+\sum_{\gamma=1}^{M}H^\beta{}_\gamma\zeta^\gamma dx\wedge dt+\sum_{\gamma=1}^{M}(A^\beta{}_\gamma dx+B^\beta{}_\gamma dt)\wedge d\zeta^\gamma, \tag{2.4}$$

where A and B are constant $(M\times M)$ matrices. It is easily shown that

$$d\bar\Omega^\beta=\sum_{k=1}^{N} f^{\beta i}\bar\alpha_i+\sum_{\gamma=1}^{M}\eta^\beta{}_\gamma\wedge\bar\Omega^\gamma, \tag{2.5}$$

provided that the matrix H is given by

$$\mathbf{H}=\mathbf{GA}-\mathbf{FB} \tag{2.6}$$

and

$$\sum_{i=K+1}^{N} f^{\beta i}\beta_i=[(dGA-dFB)\zeta]^\beta\wedge dx\wedge dt. \tag{2.7}$$

If we section $\bar\Omega^\beta$ into a solution manifold of the original, we obtain

$$\zeta_x=-F\zeta-A\zeta_y, \tag{2.8}$$

$$\zeta_t=-G\zeta-B\zeta_y, \tag{2.9}$$

$$A\zeta_t - B\zeta_x = -H\zeta. \quad (2.10)$$

For consistency these equations require that for A and B to be nontrivial

$$[A,B]=0, \quad (2.11)$$

$$[G,A]+[B,F]=0. \quad (2.12)$$

Each nonlinear solution of Eqs. (2.7), (2.11), and (2.12) yields a possible generalization of the original equation.

3. THE NONLINEAR WAVE-ENVELOPE INTERACTION EQUATIONS IN TWO SPATIAL DIMENSIONS[2,3]

The nonlinear wave—envelope interaction equations in one spatial dimension can be written in the matrix form

$$N_t = (\alpha N)_x + [\alpha N, N], \quad (3.1)$$

where N is an $(n\times n)$ matrix and $(\alpha N)_{ij} \overset{\text{def}}{=} \alpha_{ij} N_{ij}$. An equivalent set of closed 2-forms are θ_{ij} given by

$$\theta_{ij} = dN_{ij}\wedge dx + d(\alpha N)_{ij}\wedge dt + [\alpha N, N]_{ij}\, dx\wedge dt. \quad (3.2)$$

The 1-forms Ω_k defined by

$$\Omega_k = \sum_{j=1}^{n} (E-N)_{kj}\zeta_j\, dx - \sum_{j=1}^{n} (\alpha N)_{kj}\zeta_j\, dt + d\zeta_k, \quad (3.3)$$

where E is an arbitrary diagonal constant matrix provide a prolongation structure for Eqs. (3.1).

Equations (2.11), (2.12), and (2.7) become in this case

$$[A,B]=0, \quad (3.4)$$

$$[\alpha N, A]+[B,N]=0, \quad (3.5)$$

$$\beta_{ij} = \sum_{k=1}^{n} dN_{ik}\wedge dx\wedge dt(B_{kj} - \alpha_{ik}A_{kj}). \quad (3.6)$$

From Eq. (3.4) we see that both A and B can be simultaneously diagonalized. We therefore satisfy (3.4) by taking

$$A_{ij} = a_i\delta_{ij} \text{ and } B_{ij} = b_i\delta_{ij}. \quad (3.7)$$

If we substitute these into (3.5), we obtain the equation

$$N_{ik}(\alpha_{ik}(a_k - a_i) + (b_i - b_k)) = 0, \quad (3.8)$$

which we can satisfy by choosing the diagonal elements of A and B so that

$$\alpha_{ik} = (b_i - b_k)/(a_i - a_k). \quad (3.9)$$

Equation (3.6) becomes

$$\beta_{ij} = (b_j - \alpha_{ij}a_j)\, dN_{ij}\wedge dx\wedge dt. \quad (3.10)$$

Therefore, the generalization we have obtained is

$$N_t = (\alpha N)_x + (\gamma N)_y + [\alpha N, N], \quad (3.11)$$

where

$$\gamma_{ij} = (b_j - \alpha_{ij}a_j). \quad (3.12)$$

If we choose $E=-\lambda C$ where C is an arbitrary constant diagonal matrix, the sectioning of the forms Ω_k gives rise to the inverse scattering problem

$$\zeta_x - N\zeta = \lambda C\zeta, \quad (3.13)$$

$$\zeta_t - \alpha N\zeta = 0. \quad (3.14)$$

4. GENERALIZING THE BOUSSINESQ EQUATION TO THE KADOMTSEV-PETVIASHVILI-DRYUMA EQUATION[4,5]

It has been shown[7] that the forms Ω^1, Ω^2, Ω^3 defined by

$$\Omega^1 = d\zeta^1 - \zeta^2 dx + (\zeta^3 + \tfrac14 u\zeta^1)\, dt, \quad (4.1)$$

$$\Omega^2 = d\zeta^2 - (\zeta^3 - \tfrac34 u\zeta^1)\, dx - [\tfrac12 u\zeta^2 + (w - \mu - \tfrac14 p)\zeta^1]\, dt, \quad (4.2)$$

$$\Omega^3 = d\zeta^3 + [\tfrac34 u\zeta^2 + (w-\mu)\zeta^1]\, dx + [\tfrac14 r + \tfrac{9}{16}u^2)\zeta^1 - (w - \mu + \tfrac14 p)\zeta^2 + \tfrac14 u\zeta^3]\, dt \quad (4.3)$$

provide a prolongation structure for the Boussinesq equation

$$\tfrac34 u_{tt} + \tfrac14 u_{xxxx} + \tfrac32 (uu_x)_x = 0 \quad (4.4)$$

expressed in terms of the forms

$$\begin{aligned} \alpha_1 &= du\wedge dt - p\,dx\wedge dt, \\ \alpha_2 &= dp\wedge dt - r\,dx\wedge dt, \\ \alpha_3 &= du\wedge dx - \tfrac43 dw\wedge dt, \\ \alpha_4 &= dw\wedge dx + \tfrac32 up\,dx\wedge dt + \tfrac14 dr\wedge dt. \end{aligned} \quad (4.5)$$

In this case $K=3$ and α_4 is the only dynamic form. Equation (2.7) now becomes

$$\begin{bmatrix} 0&0&0\\0&0&0\\1&0&0 \end{bmatrix} \beta_4 = \begin{bmatrix} \tfrac14 du\wedge dx\wedge dt, & 0, & 0 \\ (\tfrac14 dp - dw)\wedge dx\wedge dt, & -\tfrac12 du\wedge dx\wedge dt, & 0 \\ (\tfrac14 dr + \tfrac98 u du)\wedge dx\wedge dt, & -(\tfrac14 dp + dw)\wedge dx\wedge dt, & \tfrac14 du\wedge dx\wedge dt \end{bmatrix} A - \begin{bmatrix} 0 & 0 & 0 \\ \tfrac34 du\wedge dx\wedge dt & 0 & 0 \\ dw\wedge dx\wedge dt, & \tfrac34 du\wedge dx\wedge dt & 0 \end{bmatrix} B. \quad (4.6)$$

This equation has many solutions but they all lead to the same form for β_4 which is

$$\beta_4 = \tfrac34 du\wedge dx\wedge dt. \quad (4.7)$$

The set of forms $\bar\alpha_1$, $\bar\alpha_2$, $\bar\alpha_3$, and $\bar\alpha_4 = \alpha_4\wedge dy + \tfrac34 du\wedge dx\wedge dt$ corresponds to the equation

$$\tfrac34 (u_{tt} + u_{xy}) + \tfrac14 (u_{xxx} + 6uu_x)_x = 0,$$

which is the Kadomtsev—Petviashvili—Dryuma equation. A pair of matrices which yield this 3-form and also satisfy Eqs. (2.11) and (2.12) is

$$A = 3\begin{bmatrix} 0&0&0\\0&0&0\\1&0&0 \end{bmatrix}, \quad B = \begin{bmatrix} 0&0&0\\1&0&0\\0&1&0 \end{bmatrix}. \quad (4.8)$$

With this choice of A and B Eqs. (2.8) and (2.9) become

$$\zeta^1_x = \zeta^2, \quad (4.9)$$

$$\zeta^2_x = \zeta^3 - \tfrac34 u\zeta^1, \quad (4.10)$$

$$\zeta^3_x = -\tfrac34 u\zeta^2 - (w-\mu)\zeta^1 - 3\zeta^1_y, \quad (4.11)$$

$$\zeta^1_t = -\zeta^3 - \tfrac14 u\zeta^1, \quad (4.12)$$

$$\zeta_t^2 = \tfrac{1}{2}u\zeta^2 + (w - \mu - \tfrac{1}{4}p)\zeta^1 - \zeta_y^1, \tag{4.13}$$

$$\zeta_t^3 = -(\tfrac{1}{4}r + \tfrac{9}{16}u^2)\zeta^1 + (w - \mu + \tfrac{1}{4}p)\zeta^2 - \tfrac{1}{4}u\zeta^3 - \zeta_y^2. \tag{4.14}$$

Equations (4.9), (4.10), and (4.11) together with (4.12) yield the equations

$$3\zeta_y^1 + \zeta_{xxx}^1 + \tfrac{3}{2}u\zeta_x^1 + (\tfrac{3}{4}u_x + w)\zeta^1 = \mu\zeta^1, \tag{4.15}$$

$$\zeta_t^1 + \zeta_{xx}^1 + u\zeta^1 = 0, \tag{4.16}$$

which provide an inverse scattering equation for the Kadomtsev–Petviashvili–Dryuma equation.

[1]H. Wahlquist and F. Estabrook, J. Math. Phys. 16, 1 (1975).
[2]V. Zakharov and S. Manakov, Zh. Eksp. Teor. Fiz. Pis. Red. 18, 413 (1973) [JETP Lett. 18, 243 (1973)].
[3]D. J. Benney and A. C. Newell, J. Math. & Phys. 46, 133 (1967).
[4]B. Kadomtsev and V. Petviashvili, Sov. Phys. Dokl. 15, 539 (1970).
[5]V. Dryuma, Zh. Eksp. Teor. Fiz. Pis. Red. 19, 753 (1974) [JETP Lett. 19, 387 (1974)].
[6]M. Ablowitz and R. Haberman, J. Math. Phys. 16, 2301 (1975).
[7]H. Morris, Trinity College Preprint TCD-1975-6, Trinity College (1975).

Path integrals and ordering rules

J. S. Dowker

Department of Theoretical Physics, The University, Manchester, England
(Received 1 December 1975)

It is shown again that path-integral quantization has no preference for any particular ordering rule for the Hamiltonian.

I. INTRODUCTION

It has recently[1] been claimed that the Weyl correspondence rule has a special role to play in Feynman quantization. In particular it is said that Feynman's postulate (that each path in phase space contributes an amount to the phase equal to the action) is entirely equivalent to using Weyl's rule in constructing the quantum Hamiltonian. The purpose of the present paper is to examine this statement.

Phase space functional integrals have a reasonably large literature which we do not wish to fully detail here. Some important references will be found in our earlier works.[2-4]

The calculation which we wish to discuss further here is contained in Ref. 2. There, following Tobocman, Klauder, Katz, Rosen, Martin, Garrod and Arthurs, a phase-space integral for the propagator of Schrödinger's equation,

$$i\frac{\partial}{\partial t}-\hat{H}\,\psi=0 \tag{1}$$

was constructed. The point of the calculation was to extend the work of the above authors by incorporating the ideas of Cohen,[5,6] on operator ordering.

II. DERIVATION OF PHASE-SPACE PATH INTEGRAL

The calculation is essentially straightforward, and so we only give the bare bones. The propagator $\langle q'',t''|q',t'\rangle$ is written, in standard fashion, as a folding of many short-time propagators, $\langle q_{j+1},t_{j+1}|q_j,t_j\rangle$ ($t_{j+1}=t_j+\epsilon$). Complete sets of momentum eigenstates, $|p_{j+1/2},t_{j+1/2}\rangle$, (with $t_{j+1/2}=t_j+\frac{1}{2}\epsilon$) are then inserted in these short time propagators and the $|q,t\rangle$ states on the outside are translated in time to the middle time, $t_{j+1/2}$, by Schrödinger's equation, assuming ϵ to be small, i.e., $\exp(-i\hat{H}\epsilon/2)=1-\frac{1}{2}i\hat{H}\epsilon$.

In order to evaluate the resulting brackets like $\langle q|\hat{H}|p\rangle$ the dependence of $\hat{H}(\hat{p},\hat{q})$ on $\hat{p}$ and $\hat{q}$ is exhibited by the general correspondence rule[5]

$$\begin{aligned}\hat{H}(\hat{p},\hat{q})&=\int\int d\xi\,d\eta\exp(i\hat{p}\xi+i\hat{q}\eta)F(\xi,\eta)\tilde{H}_c(\xi,\eta),\\ \tilde{H}_c(\xi,\eta)&=(2\pi)^{-2}\int\int dp\,dq\exp(-ip\xi-iq\eta)H_c(p,q),\end{aligned} \tag{2}$$

where $F(\xi,\eta)$ is restricted by $F(0,\eta)=1=F(\xi,0)$ and also, if $\hat{H}$ is Hermitian, by $F^*(\xi,\eta)=F(-\xi,-\eta)$. The standard Weyl, symmetrization and Born–Jordan rules are given by $F=1$, $F=\cos(\frac{1}{2}\xi\eta)$ and $F=\sin(\frac{1}{2}\xi\eta)/(\frac{1}{2}\xi\eta)$, respectively. (If we have a number of p's and q's we have only to interpret these, and the ξ and η, algebraically as vectors so that, e.g., $p\xi$ is just $p_\alpha\xi^\alpha$.)

Expression (2) is sandwiched between $\langle q|$ and $|p\rangle$ states and the operators $\hat{p}$ and $\hat{q}$ converted into eigenvalues. The resulting $\langle q|p\rangle$ brackets then combine with the terms of order ϵ, which have been reinstated in the exponent, to give the action, and we finally arrive at the formal limit

$$\begin{aligned}\langle q'',t''|q',t'\rangle&=\lim_{\substack{n\to\infty\\ \epsilon\to\infty}}\int\prod_{k=0}^{n}\frac{dp_{k+1/2}}{2\pi}\prod_{l=1}^{n}dq_l\\ &\times\exp\, i\sum_{j=0}^{n}[p_{j+1/2}(q_{j+1}-q_j)-H(p_{j+1/2},q_{j+1},q_j)\epsilon],\end{aligned} \tag{3}$$

where $H(p,q'',q')$ is given by

$$H(p,q'',q')=F(q''-q',-i\bar{\partial})H_c(p,\bar{q}). \tag{4}$$

The average coordinate, $\bar{q}$, equals $\frac{1}{2}(q''+q')$ and $\bar{\partial}=\partial/\partial\bar{q}$.

We should note at this point that we are assuming the coordinates q and the momenta p to range from $-\infty$ to $+\infty$. If the manifold of the q's is compact, then further considerations, not relevant to the point at issue here, will be needed.

Expression (3) can be interpreted as the skeletonized form of the path integral,

$$\langle q'',t''|q',t'\rangle=\mathcal{N}\int\int d[p]d[q]\exp[iA_F(q'',t''|q't')], \tag{5}$$

where A_F is the "classical" action. Thus, formally, as a shorthand we can write

$$A_F(q''t''|q't')=\int_{t',q'}^{t'',q''}[p\,dq-H_c(p,q)\,dt]$$

but the subscript F reminds us that we are to take the short-time form of A_F as

$$\begin{aligned}&A_F(q_{j+1},t_{j+1}|q_j,t_j)\\ &=\epsilon\{p_{j+1/2}\epsilon^{-1}[q_{j+1}-q_j]-H(p_{j+1/2},q_{j+1},q_j)\},\end{aligned} \tag{6}$$

$H(p,q'',q')$ being given by (4).

It should be apparent that there is no logical preference for the value $F=1$, i.e., for Weyl ordering. Feynman's postulate does not come ready provided with an algorithm for evaluating the functional integral. So far as the postulate is concerned, all orderings are equally good.

This also means that there is no preferred position for the "midpoint rule," which cannot in fact be extended to arbitrary Hamiltonians. This point is explained in Ref. 4. If the Hamiltonian contains a term $A(q)p$, e.g., from electromagnetic coupling, then gauge covariance (or, equivalently, hermiticity) says that we should take this as $A(\frac{1}{2}(q_{j+1}+q_j))p_{j+1/2}$ in the skeletonized path integral.[7] However, the path integral by itself does not lead unambiguously to this, or any other, form. The requirements of hermiticity or covariance have to be separately imposed, and, furthermore, any expression symmetric in q_{j+1} and q_j, such as[8] $\frac{1}{2}[A(q_{j+1})+A(q_j)]p_{j+1/2}$ would do equally as well as the midpoint form. This just reflects the well-known fact that the Hermitian ordering of $pA(q)$ is uniquely $\frac{1}{2}[\hat{p},A(\hat{q})]_+$. In fact any function F satisfying $\partial F(\xi,\eta)/\partial\xi|_{\xi=0}=0$ would produce this ordering.

This is in contrast to the situation for $p^2f(q)$ for which there is no unique Hermitian ordering. Such a situation occurs when the path integrals for a particle on a Riemannian space are being set up.[9] We have discussed this before[2-4] with the ordering question specifically in mind. The relevant results are rapidly explained.

It is convenient to assume that the quantum Hamiltonian is given and equals $-\frac{1}{2}\Delta_2$, in coordinate representation, Δ_2 being the covariant Laplace operator. Equation (2) can be inverted to give the effective "classical" Hamiltonian $H_c(p,q)$, and then $H(p,q'',q')$ is deduced from (4). We find

$$H_c(p,q)=\tfrac{1}{2}p_\alpha p_\beta g^{\alpha\beta}(q)+a\partial_\alpha\partial_\beta g^{\alpha\beta}+\tfrac{1}{2}F_0'p_\alpha\partial_\beta g^{\alpha\beta}+Q, \tag{7}$$

where

$$Q=-\tfrac{1}{2}g^{1/4}\Delta_2 g^{-1/4} \quad\text{and}\quad a=\tfrac{1}{8}[1+2(F_0')^2-F_0'']$$

with

$$F_0'\equiv\left.\frac{dF(x)}{dx}\right|_{x=0},\quad F_0''=\left.\frac{d^2F(x)}{dx^2}\right|_{x=0}$$

$$F(\tfrac{1}{2}\xi\eta)\equiv F(\xi,\eta).$$

We have taken $F(\xi,\eta)$ to be a function of $\xi^\alpha\eta_\alpha$ only. This covers most reasonable orderings. A more general discussion is contained in Ref. 10.

The expression for $H(p,q'',q')$ derived from (7) and (4) is not particularly instructive and so will not be written out. It is seen that the effective classical Hamiltonian equals the actual classical one, i.e., H_c for $\hbar=0$, plus an effective potential of order $\hbar$. This is just the quantum potential. The noncovariance of this potential is explained by noting that the operation of skeletonization is not a manifestly covariant one.[3,11]

In Ref. 3 it was shown explicitly for the special case of symmetrized ordering ($F_0'=0$, $F_0''=-1$) that after the integration over p had been done, to give a Lagrangian type path integral, the result agreed with the one derived by DeWitt,[9] after an averaging process had been performed on this.

It is also possible to discuss the Heisenberg equations of motion and the canonical commutation rules within this formalism.[3,10]

III. CONCLUSION

It has been re-emphasized that Feynman quantization shows no preference for any particular operator orderings of the Hamiltonian.[6]

This is proved by explicit construction of a phase-space path integral for the propagator of Schrödinger's equation in terms of an *arbitrary* ordering rule.

It is the freedom of choice of the lattice approximation that is the formalism's method of reflecting the factor ordering problem.[6,12]

There therefore seems to be no justification for claiming that Weyl's rule provides a royal road from Schrödinger's equation to the Feynman path integral.

[1] M.M. Mizrahi, J. Math. Phys. **16**, 2201 (1975).
[2] I.W. Mayes and J.S. Dowker, Proc. Roy. Soc. (Lond.) A **237**, 131 (1972).
[3] I.W. Mayes and J.S. Dowker, J. Math. Phys. **14**, 434 (1973).
[4] J.S. Dowker, "Covariant Schrödinger Equations," in *Proceedings of the International Conference on Functional Integration and Its Application* (London, 1974) (Clarendon Press, Oxford, 1975).
[5] L. Cohen, J. Math. Phys. **7**, 781 (1966).
[6] L. Cohen, J. Math. Phys. **11**, 3296 (1970).
[7] D.W. McLaughlin and L.S. Schulman, J. Math. Phys. **12**, 2520 (1971).
[8] R.P. Feynman, Rev. Mod. Phys. **20**, 327 (1948).
[9] B.S. DeWitt, Rev. Mod. Phys. **29**, 377 (1957).
[10] I.W. Mayes, Ph.D. thesis, Manchester University, 1971.
[11] J.S. Dowker and I.W. Mayes, Nucl. Phys. B **29**, 259 (1971).
[12] F.J. Testa, J. Math. Phys. **12**, 471 (1971).

Octonionic Hilbert spaces, the Poincaré group and SU(3)

Murat Günaydin*

Institut für Theoretische Kernphysik der Universität Bonn, Nussallee 14-16, D-5300 Bonn, West Germany
(Received 23 December 1975)

A formalism based on real octonions is developed in order to construct an octonionic Hilbert space for the description of colored quark states. The various possible forms of scalar products and related scalars are discussed. The choice of a direction in the space of octonion units leads naturally to a representation of the Poincaré group in terms of complex scalar products and complex scalars. The remaining octonion directions span the color degrees of freedom for quarks and anti-quarks. In such a Hilbert space, product states associated with color singlets are shown to form a physical quantum mechanical Hilbert space for the description of hadrons. Color triplets, on the other hand, correspond to unobservable parafermion states of order three.

1. INTRODUCTION

Even though the quark structure of hadrons has received indirect support from experiment for more than a decade now, the quarks themselves have so far not been observed. This unobservability problem became more acute when it was shown that quarks would act like partons in a gauge field theory of strong interactions in which the gauge bosons (gluons) are associated with an exact non-Abelian gauge group.[1] Because now one has to account not only for the unobservability of quarks but also of the gauge bosons in the theory which are needed to insure the Bjorken scaling observed in deep inelastic lepton nucleon scattering.

The quark models that account for the observed SU(6) multiplets naturally and also explain the $\pi^\circ \to 2\gamma$ decay in a quark parton model are the Han—Nambu model[2] and the color quark scheme of Gell-Mann,[3] which is equivalent to the paraquark scheme of Greenberg if one assumes exact color SU(3) invariance.[4,5] In the Han—Nambu model, quarks have integral charges and are assumed to be observable but because of their high masses they have not yet been seen. Whereas in the color quark scheme of Gell-Mann and Fritzsch the quarks are regarded as "mathematical objects" lying in a "fictitious" Hilbert space with an exact SU(3) (color) group. The color gauge bosons corresponding to this SU(3) group operate in this fictitious Hilbert space. Only the color singlet states are observable and form an observable subspace of this fictitious Hilbert space.

An exceptional paraquark scheme unifying the various three triplet quark models has been proposed by Gürsey and this author.[6–8] In this scheme the quark fields are regarded as transverse octonionic parafields. The color octet of gauge fields and the quark fields act in a split octonionic Hilbert space,[9] which has SU(3) as an algebraic automorphism group. On the basis of the propositional calculus of observable states as developed by Birkhoff and von Neumann this split octonionic Hilbert space is separated into an observable (longitudinal) and a nonobservable (transverse) subspace. The transverse octonionic quark fields create states that lie in the unobservable subspace. However, from these unobservable quark fields one can form product fields corresponding to observable states. The observable product states are all singlets under the algebraic automorphism group SU(3) of the Hilbert space, and the gauge fields corresponding to this SU(3) group do not couple to the fields corresponding to observable states. Thus by interpreting this SU(3) group as the color SU(3) one gets a natural realization of the proposal of Gell-Mann and Fritzsch. In a somewhat different vein Domokos *et al.*[10] have studied quark fields as elements of a *local* octonionic module. By constructing actions which are invariant under a local group G_2 and then quantizing the theory, they find that only color singlet states propagate, whereas quarks are "confined." To extend the scheme of Ref. 6 to include leptons, Gürsey suggested that the exceptional Jordan algebra be used to represent the charge space of elementary particles, quarks as well as leptons.[11] The resulting scheme leads naturally to the doubling of the number of quarks and to additional heavy leptons[12] which may find verification from the e^+e^- experiments at SLAC[13] and the neutrino experiments at Fermilab.[14]

In Refs. 6 and 9 the underlying algebra was taken as the split octonion algebra. In this article our aim is to show that one can obtain the same results over the real octonion algebra which is a division algebra. The plan of the paper is as follows. We first study the possible bilinear forms over the real octonions and establish their invariance groups. On physical grounds we choose the real octonionic Hilbert space with complex scalar products and study its properties. This Hilbert space has an algebraic automorphism group SU(3) and a gauge group U(4). We construct the representations of the Poincaré group over this Hilbert space à la Wigner. The space—time labels of the Wigner basis do not form a complete set over this Hilbert space. Additional "internal" labels have to be taken from the algebraic automorphism group SU(3). We then consider real octonionic transverse quark fields and show that by interpreting the algebraic automorphism group SU(3) of the Hilbert space as the exact SU(3) group of the 3-triplet quark models one obtains the same results as in Ref. 6. Finally we introduce an "algebra of colors" which is a six-dimensional Malcev algebra formed by color carrying octonion units and which has the color SU(3) as an automorphism group.

2. OCTONION ALGEBRA AS EXTENSIONS OF COMPLEX NUMBERS AND QUATERNIONS

The real octonion algebra $\mathbb{O}$ is an eight dimensional division algebra which is neither commutative nor associative.[15,16] A basis of this algebra can be chosen as[16]

$$1, e_A, \quad A = 1, 2, \ldots, 7,$$

where the elements e_A satisfy the following multiplication rule:

$$\begin{aligned} &e_1e_2 = e_3, \quad e_6e_7 = e_3, \\ &e_4e_7 = e_1, \quad e_6e_2 = e_4, \quad e_4e_3 = e_5 \\ &e_5e_1 = e_6, \quad e_5e_7 = e_2, \end{aligned} \tag{2.1}$$

and

$$e_Ae_B + e_Be_A = -2\delta_{AB}, \quad A, B = 1, 2, \ldots, 7.$$

As can easily be verified from the multiplication table, the real octonion algebra is an alternative algebra, i.e., the associator of three elements $X, Y, Z \in \mathbb{O}$ is an alternating function of its arguments:

$$\begin{aligned} [X, Y, Z] &\equiv (XY)Z - X(YZ) \\ &= [Z, X, Y] = [Y, Z, X] = -[Y, X, Z]. \end{aligned}$$

The norm N of an element $X \in \mathbb{O}$ is defined as

$$N(X) \equiv \bar{X}X = X\bar{X}, \tag{2.2}$$

where

$$X = X_0 + \sum_{A=1}^{7} X_A e_A, \qquad X_A, X_0 \in \mathbf{R}.$$
$$\bar{X} = X_0 - \sum_{A=1}^{7} X_A e_A,$$

The overbar denotes the octonion conjugation.

As is well known, one can obtain the real octonion algebra by a complex extension of the quaternions. To see this, decompose an element $X \in \mathbb{O}$ as

$$\begin{aligned} X &= (X_0 + X_1e_1 + X_2e_2 + X_3e_3) + e_7(X_7 + X_4e_1 + X_5e_2 + X_6e_3) \\ &= Q_1 + e_7Q_2 \end{aligned}$$

where Q_1, Q_2 belong to the quaternion subalgebra $\mathbb{H}$ generated by e_i $(i = 1, 2, 3)$. In this form, the octonion multiplication is given by

$$\begin{aligned} XY &= (Q_1 + e_7Q_2)(R_1 + e_7R_2) \\ &= (Q_1R_1 - R_2\tilde{Q}_2) + e_7(\tilde{Q}_1R_2 + R_1Q_2), \end{aligned} \tag{2.3}$$

where $\tilde{Q}$ is the quaternion conjugate of Q obtained by replacing e_i in Q by $-e_i$ $(i = 1, 2, 3)$. In general the operation ~ on an octonion X will be defined as:

$$X = Q_1 + e_7Q_2, \quad \tilde{X} \equiv \tilde{Q}_1 + e_7\tilde{Q}_2.$$

This conjugation operation ~ is not an automorphism of the octonion algebra.[17]

The real octonion algebra can also be written as a "quaternionic extension" of complex numbers. To do this, rewrite an octonion as

$$\begin{aligned} X &= X_0 + \sum_{A=1}^{7} X_A e_A \\ &= (X_0 + e_7X_7) + (X_1 + e_7X_4)e_1 + (X_2 + e_7X_5)e_2 + (X_3 + e_7X_6)e_3 \end{aligned}$$

or

$$X = x_0 + x_1e_1 + x_2e_2 + x_3e_3,$$

where $x_0 = X_0 + e_7X_7$, $x_i = X_i + e_7X_{i+3}$ $(i = 1, 2, 3)$, i.e., x_0, x_i belong to the complex subalgebra $\mathbb{C}$ generated by the imaginary unit e_7. Then the octonion multiplication takes the form

$$\begin{aligned} XY &= \left(x_0 + \sum_{i=1}^{3} x_ie_i\right)\left(y_0 + \sum_{i=1}^{3} y_ie_i\right) \\ &= \left(x_0y_0 - \sum_{i=1}^{3} x_iy_i^*\right) \\ &\quad + \sum_{k=1}^{3}(x_0y_k + y_0^*x_k + \sum_{i,j=1}^{3} \epsilon_{ijk}x_i^*y_j^*)e_k, \end{aligned} \tag{2.4}$$

where * operation denotes complex conjugation $(e_7 \to -e_7)$ within the complex subalgebra generated by e_7 and is an automorphism of the octonion algebra.

3. BILINEAR FORMS AND THEIR INVARIANCE GROUPS OVER THE REAL OCTONION ALGEBRA

There are four possible "bilinear forms"[18] that can be defined over the real octonion algebra which induce the usual octonionic norm and satisfy the composition law

$$N(XY) = N(X)N(Y), \quad X, Y \in \mathbb{O}.$$

These bilinear forms are:

(I) The bilinear form $(X, Y)_R$ which is real:

$$(X, Y)_R \equiv \tfrac{1}{2}(\bar{X}Y + \bar{Y}X) = \sum_{a=0}^{7} X_aY_a. \tag{3.1}$$

This bilinear product has the invariance group SO(8), which is also the invariance group of the octonionic norm.

(II) The bilinear product $(X, Y)_C$ which is complex:

$$(X, Y)_C \equiv \tfrac{1}{2}\{(\bar{X}Y) + \widetilde{(\bar{X}Y)}\}, \tag{3.2}$$

where ~ denotes quaternion conjugation defined above. In terms of the real components X_a, Y_a one has

$$\begin{aligned} (X, Y)_C = \sum_{a=0}^{7} X_aY_a &+ e_7(X_0Y_7 - X_7Y_0 + X_4Y_1 - X_1Y_4 \\ &+ X_5Y_2 - X_2Y_5 + X_6Y_3 - X_3Y_6). \end{aligned}$$

Now the real part of this product is invariant under SO(8) and the imaginary part has the invariance group Sp(8). Since SO(8) ∩ Sp(8) ≈ U(4) we see that the invariance group of this bilinear product is U(4). This can be seen more clearly by using the form of real octonion algebra as quaternionic extension of complex numbers, i.e.,

$$X = \sum_{a=0}^{7} X_ae_a = x_0 + \sum_{i=1}^{3} x_ie_i,$$
$$Y = \sum_{a=0}^{7} Y_ae_a = y_0 + \sum_{i=1}^{3} y_ie_i.$$

Then

$$(X, Y)_C = x_0^*y_0 + \sum_{i=1}^{3} x_iy_i^*, \quad x_0, y_0, x_iy_i \in \mathbb{C}.$$

In this form of the product, its invariance group U(4) is more obvious.

(III) The bilinear form that is a quaternion can be defined as

$$(X, Y)_H \equiv \tfrac{1}{2}\{(\bar{X}Y) + (\bar{X}Y)^*\} = \tfrac{1}{2}(\bar{X}Y + \bar{X}^* Y^*), \tag{3.3}$$

where * denotes complex conjugation ($e_7 \to -e_7$). Writing X and Y as doubled quaternions,

$$X = Q_1 + e_7 Q_2, \quad Y = R_1 + e_7 R_2,$$

we have

$$(X, Y)_H = \tilde{Q}_1 R_1 + R_2 \tilde{Q}_2.$$

This product has the invariance group SO(3)⊗SO(3) induced by the multiplications

$$\mathrm{SO}(3): Q_1, R_1 \xrightarrow{P} PQ_1, PR_1, \quad \text{where } P\tilde{P} = 1, \; P \in \mathbb{H},$$

and

$$\mathrm{SO}(3): Q_2, R_2 \xrightarrow{Q} Q_2 Q, R_2 Q, \quad \text{where } Q\tilde{Q} = 1, \; Q \in \mathbb{H}.$$

Note also that because the scalar product is not of the form $(\tilde{Q}_1 R_1 + \tilde{Q}_2 R_2)$, the invariance group is not the symplectic group Sp(2) ≈ SO(5).

(IV) The bilinear form that is an octonion is simply given by

$$(X, Y)_0 \equiv \bar{X} Y. \tag{3.4}$$

The invariance group of this form is the trivial multiplication by ± 1.[19]

4. THE REAL OCTONIONIC HILBERT SPACE WITH COMPLEX SCALAR PRODUCTS

In the next section we shall construct the representations of the Poincaré group over a real octonionic Hilbert space in such a way that the Hilbert space will carry non-Abelian automorphism and gauge groups. We shall require this Hilbert space to have two essential features:

(a) The Hilbert space must contain as a special case the standard complex Hilbert space of quantum mechanics with the gauge invariance group U(1).

(b) The Hilbert space should have SU(3) as an algebraic automorphism group. This requires that the scalar products be complex in the Hilbert space.

One can satisfy these conditions in two ways, i.e., by constructing the Hilbert space over the split octonion algebra as was done in Refs. 6 and 9 or over the real octonion algebra which we shall give below. As stated in Ref. 6, the automorphism group SU(3) of the Hilbert space must be an exact symmetry and hence cannot be identified as the broken unitary spin group. Therefore, one ought to identify this SU(3) group with the exactly conserved SU(3) group of three triplet quark models.[6]

Now denoting the real octonionic Hilbert space by $\mathcal{H}$ we can decompose every state vector $|Z\rangle$ of $\mathcal{H}$ as

$$|Z\rangle = |Z\rangle_0 + \sum_{A=1}^{7} |Z\rangle_A e_A \tag{4.1}$$

where $|Z\rangle_a$ ($a = 0, 1, 2, \ldots, 7$) are vectors with real components. The vector $|Z\rangle$ can also be decomposed as

$$|Z\rangle = |z)_0 + \sum_{i=1}^{3} |z)_i e_i, \tag{4.2}$$

where $|z)_\lambda$ ($\lambda = 0, 1, 2, 3$) are vectors with complex components, i.e.,

$$|z)_0 = |Z\rangle_0 + e_7 |Z\rangle_7, \quad |z)_i = |Z\rangle_i + e_7 |Z\rangle_{i+3}.$$

The combinations of such vectors with real octonion coefficients also belong to the space $\mathcal{H}$. If the index z is continuous, one can construct wavepacket states $|F\rangle$ such that

$$|F\rangle = |f_0)_0 + \sum_{i=1}^{3} |f_i)_i e_i$$
$$= \sum_{\lambda=0}^{3} |f_\lambda)_\lambda e_\lambda, \quad \lambda = 0, 1, 2, 3,$$

where $e_0 = 1$ and

$$|f_\lambda)_\lambda = \int f_\lambda(z) \, |z)_\lambda \, d\mu(z)$$

where $d\mu(z)$ is some measure associated with the label z. The corresponding bra vector will be defined as

$$\langle F| \equiv (\overline{|F\rangle})^T = {}_0(f_0| - \sum_{i=1}^{3} e_i \, {}_i(f_i|$$
$$= {}_0(f_0| - \sum_{i=1}^{3} {}_i(f_i^*| \, e_i, \tag{4.3}$$

where

$${}_\lambda(f_\lambda| \equiv (|f_\lambda)_\lambda)^\dagger$$

and the super † operation denotes the usual Hermitian conjugation of complex quantum mechanics.

Let $|F\rangle$ and $|G\rangle$ be two such states; then the octonionic bilinear product $\langle G|F\rangle$ will be given as

$$\langle G|F\rangle = \{{}_0(g_0| - \sum_{i=1}^{3} e_i \, {}_i(g_i|\}\{|f_0)_0 + \sum_{j=1}^{3} |f_j)_j e_j\}$$
$$= {}_0(g_0|f_0)_0 + \sum_{i=1}^{3} {}_i(f_i|g_i)_i$$
$$+ \sum_{k=1}^{3} \{{}_0(g_0|f_k)_k - {}_0(f_0|g_k)_k - \sum_{i,j=1}^{3} \epsilon_{ijk} \, {}_i(g_i|f_j^*)_j\} e_k, \tag{4.4}$$

where ${}_\lambda(g_\lambda|f_\nu)_\nu$ is the usual complex scalar product of quantum mechanics.

The complex bilinear product of these two states will simply be :

$$(G, F) \equiv \tfrac{1}{2}\{\langle G|F\rangle + \widetilde{\langle G|F\rangle}\}$$

or

$$(G, F) = {}_0(g_0|f_0)_0 + \sum_{i=1}^{3} {}_i(f_i|g_i)_i \tag{4.5}$$

From here on we shall consider only complex bilinear products and take as the scalar product in $\mathcal{H}$ this complex bilinear form. Under this scalar product the states of the form $|f_\lambda)_\lambda e_\lambda$ and $|f_\nu)_\nu e_\nu$ are orthogonal for $\lambda \neq \nu$. This follows from the definition of the scalar product and is independent of the complex bilinear product ${}_\lambda(f_\lambda|g_\nu)_\nu$.

Let α be a complex number belonging to the complex subalgebra $\mathbb{C}$ generated by e_7. Then from the definition of the scalar product we have

$$(G, F\alpha) = \alpha(G, F) = (G, F)\alpha \quad (4.6a)$$

and

$$(G\alpha, F) = \alpha^*(G, F) = (G, F)\alpha^*. \quad (4.6b)$$

Therefore, our scalar product is "right sesquilinear" in the complex field $\mathbb{C}$. However, it is not left sesquilinear in $\mathbb{C}$ since

$$(G, \alpha F) \neq \alpha(G, F), \quad (\alpha G, F) \neq \alpha^*(G, F).$$

But we still have

$$(\alpha^* G, F) = (G, \alpha F), \quad (4.7)$$

and therefore one can define the Hermitian conjugate $\Omega^\dagger$ of an operator Ω acting in our Hilbert space in the usual way, i.e.,

$$(\Omega^\dagger G, F) \equiv (G, \Omega F). \quad (4.8)$$

For a Hermitian operator H we will have

$$(HG, F) \equiv (G, HF), \quad (4.9)$$

and we define a unitary operator U as usual:

$$(UF, UG) \equiv (F, G). \quad (4.10)$$

Now, allowing only multiplications of state vectors by complex numbers generated by e_7, we see that the vectors of the form

$$\{\alpha|f_\lambda)_\lambda + \beta|g_\lambda)_\lambda\}e_\lambda, \quad \lambda = 0, 1, 2, 3, \quad \alpha, \beta \in \mathbb{C}$$

form four mutually orthogonal subspaces of our Hilbert space under the scalar product we defined. The subspace generated by the elements of the form:

$$\alpha|f_0)_0 + \beta|g_0)_0, \quad \alpha, \beta \in \mathbb{C}$$

corresponds to the usual complex Hilbert space of quantum mechanics. This complex subspace has the gauge invariance group U(1) corresponding to multiplication of the state vectors by a phase, which leave the scalar product invariant.

We shall call the automorphism group of the underlying composition algebra of a Hilbert space the "intrinsic covariance group." For the complex subspace the intrinsic covariance group is the cyclic group C_2 generated by complex conjugation ($e_7 \to -e_7$). The full octonionic Hilbert space has the intrinsic covariance group G_2, the automorphism group of octonions. The complex conjugation ($e_7 \to -e_7$), which is a discrete transformation for complex numbers, gets extended to an element of the group G_2.[16]

Now the intersection of the intrinsic covariance group C_2 and the gauge group U(1) of the complex Hilbert space is the trivial identity mapping:

$$C_2 \cap \mathrm{U}(1) = 1.$$

On the other hand for the real octonionic Hilbert space the gauge group U(4) and the intrinsic covariance group G_2 have a common SU(3) subgroup:

$$G_2 \cap \mathrm{U}(4) \approx \mathrm{SU}(3).$$

To prove this, it suffices to show that an SU(3) subgroup of the gauge group U(4) induces transformations which correspond to the SU(3) automorphisms of the octonion algebra leaving an imaginary unit, in our case e_7, invariant.

Consider the transformation of the complex component vectors $|f_i)_i$ ($i = 1, 2, 3$) of the octonionic state vector $|F\rangle = |f_0)_0 + \sum_{i=1}^3 |f_i)_i e_i$ under the SU(3) group:

$$\mathrm{SU}(3): |f_i)_i \to |f_i')_i = \sum_{j=1}^{3} U_{ij}|f_j)_j \quad (4.11)$$

where U_{ij} is the 3×3 special unitary matrix taken over the complex field $\mathbb{C}$ generated by e_7. This transformation induces the mappings

$$\mathrm{SU}(3): |F\rangle \to |F'\rangle = |f_0)_0 + \sum_{i,j} U_{ij}|f_j)_j e_i,$$

$$\mathrm{SU}(3): |G\rangle \to |G'\rangle = |g_0)_0 + \sum_{i,j} U_{ij}|g_j)_j e_j,$$

which leave the complex scalar product invariant, i.e.,

$$(G', F') = (G, F),$$

and hence belong to the gauge group U(4). However, these transformations are equivalent to transforming the imaginary units e_i as

$$\mathrm{SU}(3): e_i \to e_i' = \sum_j U_{ji}e_j = \sum_j e_j U_{ji}^*. \quad (4.12)$$

In the Appendix we show that these SU(3) transformations (written in terms of e_7) acting on the imaginary units e_i correspond to the SU(3) subgroup of the automorphism group G_2 of octonions that leave the imaginary unit e_7 invariant.

Therefore, under the above SU(3) transformations the algebraic relations over the Hilbert space H and the scalar products are left invariant. The complex conjugation operation * is an automorphism that lies outside the SU(3) subgroup of G_2 and induces a natural mapping from a representation of SU(3) to its conjugate representation.

The importance of the octonionic Hilbert space with complex scalar products lies in the fact one has to fix an imaginary octonion unit in order to implement the unitary representations of the Poincaré group over this Hilbert space as indicated in the next section. Another important property of the complex scalar product is that it satisfies the intermediate state decomposition property, i.e., if $|F\rangle$ and $|G\rangle$ are two states in the Hilbert space, one should be able to go from one to the other via intermediate states that form a complete set. In other words we want

$$\begin{aligned}(\langle F|, |G\rangle) &= \sum_N (\langle F|, |N\rangle)(\langle N|, |G\rangle) \\ &= \sum_N \{(\langle F|, |N\rangle)(\langle N|, \}|G\rangle) \\ &= \sum_N (\langle F|\{, |N\rangle)(\langle N|, |G\rangle)\} \end{aligned} \quad (4.13)$$

and this is satisfied by complex scalar products but not by quaternionic or octonionic products. Since we shall consider complex scalar products, we shall only allow multiplication by complex numbers.[20] Otherwise one can go from a state to another state orthogonal to it by scalar multiplication.

5. ONE PARTICLE UNITARY REPRESENTATIONS OF THE POINCARÉ GROUP OVER THE REAL OCTONIONIC HILBERT SPACE WITH COMPLEX SCALAR PRODUCTS

Below we shall construct the one particle unitary representations of the Poincaré Group over the Hilbert space $\mathcal{H}$ of the previous section.[21] For the space–time labels of our states we shall choose the Wigner basis, i.e., the Casimir operators

$$M^2 = P_\mu P^\mu \quad \text{and} \quad \mathbf{S}\cdot\mathbf{S}$$

corresponding to a definite mass and spin (m, s) and the operators

$$\mathbf{P}, S_3$$

corresponding to the 3-momentum and the third component of spin, respectively. The general octonionic eigenstate of these space–time operators will be a linear combination of states of the form

$$|m, s; \mathbf{p}, s_3, \lambda\rangle \quad |m, s; \mathbf{p}, s_3)^\lambda e_\lambda,$$
$$\lambda = 0, 1, 2, 3, \tag{5.1}$$

where complex component vectors $|m, s; \mathbf{p}, s_3)^\lambda$ are eigenstates of the Wigner basis. We shall drop the labels m, s from these states and label them simply as $|p, s_3, \lambda\rangle$. Then normalizing the complex component vectors in the conventional manner, i.e.,

$${}^\lambda(p', s_3' | p, s_3)^\lambda = \delta(\mathbf{p} - \mathbf{p}')\delta_{s_3, s_3'}, \quad \lambda = 0, 1, 2, 3,$$

and denoting the complex scalar product of two octonionic state vectors $|F\rangle$ and $|G\rangle$ as $(\langle F|G\rangle)$, we find that

$$(\langle p', s_3', \nu | p, s_3, \lambda\rangle) = \delta(\mathbf{p} - \mathbf{p}')\delta_{s_3, s_3'}\delta_{\nu\lambda},$$
$$\nu, \lambda = 0, 1, 2, 3. \tag{5.2}$$

Therefore, we see that the space–time labels do not form a complete set over the Hilbert space $\mathcal{H}$, and one needs additional labels. We shall show below that these additional labels can be taken from the automorphism group SU(3) of $\mathcal{H}$.

Let us now construct the unitary representations of the covering group $T_4 \circledS SL(2, C)$ of the Poincaré group over $\mathcal{H}$ following Wigner.[22] We will denote the 4-vectors by Hermitian 2×2 matrices:

$$p = p^0 + \sigma\cdot\mathbf{p} = \begin{pmatrix} p^0 + p^3 & p^1 - e_7 p^2 \\ p^1 + e_7 p^2 & p^0 - p^3 \end{pmatrix} = \sum_{\mu=0}^{3} p^\mu \sigma_\mu,$$

$$p\cdot p \equiv p_\mu p^\mu = \det p$$

and consider only the case of timelike 4-momenta. Then under SL(2, C)

$$SL(2, C)\colon p \to \Lambda p \Lambda^\dagger = p' = p^{0\prime} + \sigma\cdot\mathbf{p}'$$

where $\Lambda = \exp[e_7(\sigma/2)\cdot(\omega - e_7\nu)]$ is the 2×2 matrix characterizing the Lorentz transformation for timelike vectors. The scalar product of two 4-vectors x_μ, y_μ will be given by

$$x\cdot y = x_\mu y^\mu = \tfrac{1}{2}\mathrm{Tr}(x\dot{y}) = \tfrac{1}{2}\mathrm{Tr}(\dot{x}y),$$

where

$$\dot{y} \equiv y^0 - \sigma\cdot\mathbf{y} = \sum_{\mu=0}^{3} y_\mu \sigma_\mu,$$

i.e., dotted matrices correspond to covariant vectors.

Denoting the pure translations by a and elements of $SL(2, C)$ by Λ, the group equation for $T_4 \circledS SL(2, C)$ reads, in terms of 2×2 matrices, as

$$(a_1, \Lambda_1)(a_2, \Lambda_2) = (a_1 + \Lambda_1 a_2 \Lambda_1^\dagger, \Lambda_1\Lambda_2). \tag{5.3}$$

In terms of the unitary operators T_a and L_Λ acting in $\mathcal{H}$, the group equations will be

$$\begin{aligned} T_{a_1} T_{a_2} &= T_{a_1 + a_2}, \\ L_{\Lambda_1} L_{\Lambda_2} &= L_{\Lambda_1\Lambda_2}, \\ L_\Lambda T_a &= T_{\Lambda a \Lambda^\dagger} L_\Lambda, \end{aligned} \tag{5.4}$$

where T_a and L_Λ represent the translations and $SL(2, C)$ transformations, respectively.

Then under translations T_a the states $|p, s_3, \lambda\rangle$ will transform simply as[23]

$$T_a |p, s_3, \lambda\rangle = \exp(e_7 p\cdot a)|p, s_3, \lambda\rangle, \tag{5.5}$$

where

$$p\cdot a = \tfrac{1}{2}\mathrm{Tr}(p\dot{a}).$$

Under the $SL(2, C)$ transformations

$$L_\Lambda |p, s_3, \lambda\rangle = \left(\frac{\omega(\Lambda p \Lambda^\dagger)}{\omega(p)}\right)^{1/2} \sum_{s_3'=-s}^{s} U^{(s)}_{s_3' s_3}(\Lambda, p) \times |\Lambda p \Lambda^\dagger, s_3', \lambda\rangle, \tag{5.6}$$

where $U^s(\Lambda, p)$ is the $(2s+1)\otimes(2s+1)$ unitary matrix of Wigner and

$$\omega(p) = (\mathbf{p}\cdot\mathbf{p} + m^2)^{1/2}.$$

Unitarity of these group actions follows from the complex scalar product defined and is very easy to verify.

Since the states $|p, s_3, \lambda\rangle$ have delta function normalization, they do not strictly belong to the Hilbert space. To be rigorous, one ought to consider wavepacket states of the form

$$|m, s; f_\lambda, \lambda\rangle = \sum_{s_3=-s}^{s} \int \frac{d^3p}{\sqrt{2\omega(p)}} f_\lambda(p, s_3)|p, s_3, \lambda\rangle,$$

where the f_λ are square integrable complex functions.

So far we have designated the eigenstates of the Wigner basis by $|p, s_3, \lambda\rangle$, where states with different λ are orthogonal even though they may have the same space–time properties. To distinguish between such states, we need additional labels. As the additional labels we can take the two Casimir operators and the two generators of the Cartan subalgebra of the automorphism group SU(3) of the Hilbert space. This SU(3) group will be called the C-spin group[6] and its Casimir operators and Cartan subalgebra generators will be denoted by c_1, c_2 and I_3^c, Y^c in analogy with the unitary spin SU(3). Since the units e_1, e_2, e_3 transform like a triplet under the automorphism group, which we shall designate as $SU(3)_c$, the states $|p, s_3, n\rangle$ for $n = 1, 2, 3$ will be assigned the quantum numbers of an $SU(3)_c$ triplet and the complex states of the form $|p, s_3)^0 = |p, s_3, \lambda = 0\rangle$ will be $SU(3)_c$

singlets. Then the complex conjugated states $|p, s_3, n\rangle^* = |p, s_3)^{n^*} e_n$, $n=1,2,3$ will have the quantum numbers of an $SU(3)_c$ antitriplet.

6. TENSOR PRODUCT STATES OVER THE OCTONIONIC HILBERT SPACE AND THE C-SPIN GROUP

Above we saw that for states of the form

$$|f_0)^0 + \sum_{i=1}^{3} |f_i)^i e_i$$

the SU(3) subgroup of the gauge group under which $|f_i)^i$ $(i=1,2,3)$ transform as a triplet corresponds to the SU(3) subgroup of the automorphism group of octonions under which the units e_i transform as a triplet and vice versa. If all the states in the Hilbert space are of the form

$$\alpha |f_0)^0 + \beta \sum_{i=1}^{3} |f_i)^i e_i, \quad \alpha, \beta \in \mathbb{C} ,$$

then there is a complete correspondence between these two SU(3) groups and the Hilbert space has this SU(3) group as an algebraic automorphism group. However, if we allow, more generally, states of the form

$$\alpha_0 |f_0)^0 + \sum_{i=1}^{3} \alpha_i |f_i)^i e_i, \quad \alpha_\lambda \in \mathbb{C} ,$$

then only the Abelian subgroups of these two SU(3) groups overlap and the algebraic automorphism group of the Hilbert space becomes the Abelian group $U(1) \otimes U(1)$ generated by the Cartan subalgebra generators I_3^c and Y^c of $SU(3)_c$. Therefore, there will not be any change in the quantum number assignment scheme in this more general case.

Let us now consider the problem of tensor products over the octonionic Hilbert space in this general case. Because the underlying division algebra is the nonassociative octonion algebra, there will be superselection rules for the tensor product states. First we shall divide the Hilbert space into two parts,[6] longitudinal and transverse.

The states of the form $\sum_{i=1}^{3} |f)^i e_i$, where $|f)^i$ are complex component vectors, will be called transverse and the states of the form $|f)^0$ in the complex subspace will be called longitudinal. Longitudinal vectors in the Hilbert space will be C-spin singlets and the transverse vectors $|f)^i e_i$ $(i=1,2,3)$ will carry the quantum numbers of a C-spin triplet. Naturally the complex conjugate states $|f)^{i*} e_i$ will have the quantum numbers of an anti-triplet.

Now consider the product of the states $|f)^i e_i$ and $|f)^{i*} e_i$:

$$\{|f)^i e_i\}\{|f)^{i*} e_i\} = -|f)^i |f)^i. \qquad (6.1)$$

Since the right-hand side is a pure complex state, it must be a C-spin singlet. However, with respect to the C-spin indices it is a member of a C-spin sextet. Therefore, if we are to have a self-consistent scheme, we must not allow such product states over the Hilbert space.[24]

This example shows that the underlying octonion algebra introduces certain superselection rules for tensor product states. To understand this phenomenon better, recall that in a state of the form

$$|f)^\lambda e_\lambda$$

transformation properties of e_λ under the SU(3) automorphisms of octonions induce definite transformation properties on the complex vector components $|f)^\lambda$ and one can assign C-spin quantum numbers to e_λ and $|f)^\lambda$ interchangeably. Now under a tensor product of two copies of the Hilbert space we do not take two different copies of the octonion algebra. The octonion algebra is taken as the underlying division algebra of all the Hilbert spaces we are considering. In a product state of the form

$$\{|f)^i e_i\}\{|g)^j e_j\}$$

the units e_i, e_j satisfy the multiplication rule

$$e_i e_j = -\delta_{ij} + \epsilon_{ijk} e_k, \quad i,j = 1,2,3,$$

whereas the complex vector components $|f)^i$ and $|g)^i$ do not satisfy any such rule. Since the units e_i satisfy the above multiplication rule, the identity

$$\sum_m \epsilon_{ijm} U_{mk} = \sum_{m,l} \epsilon_{lmk} U^*_{li} U^*_{jm} \qquad (6.2)$$

derived for the 3×3 SU(3) matrices in the Appendix prevents the appearance of higher multiplets, whereas the complex vector components $|f)^i$ and $|g)^i$ do not satisfy such a multiplication rule and hence can combine to give higher multiplets. However, self-consistency requires that a tensor product state of the form $|f)^i |g)^i e_k$ agree in the C-spin quantum number assignment to the complex vector components $|f)^i |g)^j$ and to the units e_k. This implies that all the product states in our Hilbert space must be either C-spin singlets or members of a C-spin triplet. However, this does not imply that one cannot define higher C-spin tensor operators, such as an octet operator. What this implies is that these operators act nontrivially only on those states for which the resulting state is a C-spin singlet or a member of a C-spin triplet.[6]

7. OBSERVABLE STATES, C-SPIN GROUP AND QUARK STATISTICS

The state vectors in the octonionic Hilbert space decompose as follows

$$\begin{aligned} |F\rangle &= |F\rangle_L + |F\rangle_T, \quad |F\rangle \in H, \\ |F\rangle_L &= |F)^0 \in H_L, \\ |F\rangle_T &= |F)^1 e_1 + |F)^2 e_2 + |F)^3 e_3 \in H_T, \\ H &= H_L \oplus H_T, \end{aligned} \qquad (7.1)$$

where H_L stands for the "longitudinal" subspace of H spanned by pure complex state vectors and H_T denotes the "transverse" part of the Hilbert space spanned by states that have components along e_1, e_2, e_3 directions only. Now the states in H_L are commutative and associative, whereas the states in H_T are neither commutative nor associative in general. The states in H_T will be unobservable since the postulates for observable states as formulated by Birkhoff and von Neumann[25] cannot be satisfied by states having nonassociative components.[25–29] In fact for states having nonassociative and noncommutative components there is no satisfactory way of defining tensor product states.[6,30]

However, even though the transverse octonionic states do not correspond to observables, one can construct longitudinal product states from them that will correspond to observables.[6–8] In other words from states of the form $|A\rangle=\sum_{i=1}^{3}|A)^i e_i$, where $|A)^i$ transform as a C-spin triplet we want to construct product states that are C-spin singlets. Since these states are neither associative nor commutative, we consider two natural products that have $SU(3)_c$ as an automorphism group, namely,

$$|A\rangle * |B\rangle \equiv \tfrac{1}{2}[\,|A\rangle, |B\rangle, e_7] \equiv \tfrac{1}{2}[(|A\rangle|B\rangle)e_7 - |A\rangle(|B\rangle e_7)] \tag{7.2}$$

$$|A\rangle \circ |B\rangle \equiv \tfrac{1}{2}\{|A\rangle, |B\rangle, e_7\} \equiv \tfrac{1}{2}[(|A\rangle|B\rangle)e_7 + |A\rangle(|B\rangle e_7)] \tag{7.3}$$

where $[\ ,\ ,\ ]$ and $\{\ ,\ ,\ \}$ stand for the associator and the antiassociator, respectively. In the next section, we show how the $*$ product leads to a six-dimensional Malcev algebra that has $SU(3)_c$ as the automorphism group. For two states of the form $|A\rangle=\sum_i|A)^i e_i$, $|B\rangle=\sum_i|B)^i e_i$ the C-spin singlet product state will be

$$|A\rangle \circ |B\rangle = -\sum_{i=1}^{3}|A)^i|B)^{i*}e_7 \tag{7.4}$$

where * denotes complex conjugation ($e_7 \to -e_7$). Similarly the C-spin singlet product of three states $|A\rangle$, $|B\rangle$, and $|C\rangle$ will be given by

$$(|A\rangle * |B\rangle)\circ|C\rangle = -\epsilon_{ijk}|A)^{i*}|B)^{j*}|C)^{k*} \tag{7.5}$$

or

$$|A\rangle\circ(|B\rangle * |C\rangle) = -\epsilon_{ijk}|A)^i|B)^j|C)^k. \tag{7.6}$$

In Ref. 6, the C-spin group $SU(3)_c$ was identified with the exact SU(3) group of the three triplet quark models[2–5] and transverse split octonionic quark field operators were constructed. The observable product states that can be constructed from these unobservable quark fields correspond to the observed unitary spin multiplets. Same results can be obtained using the real octonionic quark field operators. For this consider the transverse real octonionic quark field operators Ψ_i, $i=1,2,3$, where the index i refers to the unitary spin SU(3) indices:

$$\Psi_i = \sum_{n=1}^{3} q_i^n(x)e_n, \tag{7.7}$$

where n refers to the C-spin SU(3) indices and $q_i^n(x)$ satisfy the usual anticommutation relations of a spin $\frac{1}{2}$ fermion field, i.e.,

$$\{q_i^{n\dagger}(\mathbf{x}), q_j^m(\mathbf{y})\}_{x^0=y^0} = \delta^{nm}\delta_{ij}\delta(\mathbf{x}-\mathbf{y}),\quad \{q_i^m(\mathbf{x}), q_j^n(\mathbf{y})\}=0. \tag{7.8}$$

Below we shall use the Majorana representation of γ matrices so that charge conjugation reduces to Hermitian conjugation:

$$q_i^{n*}(x) = (q_i^n(x))^C = Cq_i^n(x)C^{-1}. \tag{7.9}$$

The C-spin singlet (longitudinal) field operator that can be constructed from two copies of the transverse octonionic quark field operator is

$$V_{ij}(x) = \Psi_i \circ \Psi_j = -\sum_{n=1}^{3} q_i^n(x)q_j^{*n}(x)e_7. \tag{7.10}$$

Acting on the vacuum with this composite field operator $V_{ij}(x)|0\rangle = -e_7\Phi_{ij}(x)|0\rangle$, where $\Phi_{ij}(x)=\sum_{n=1}^{3}q_i^n(x)q_j^{*n}(x)$ are ordinary complex boson operators belonging to the singlet representation of $SU(3)_c$ and the (octet + singlet) representation of unitary spin SU(3). If we add also the suppressed quark spin indices, these states resolve into C-even spin-zero states and C-odd spin-one states in $\mathcal{H}_L$.

Similarly from three copies of the transverse octonionic quark fields we can construct the following longitudinal field operators:

$$\Psi_{ijk}^*(x) \equiv (\Psi_i(x) * \Psi_j(x))\circ\Psi_k(x) = -\epsilon_{mnp}q_i^{*m}(x)q_j^{*n}(x)q_k^{*p}(x), \tag{7.11}$$

$$\Psi_{ijk}(x) \equiv \Psi_i(x)\circ(\Psi_j(x) * \Psi_k(x)) = -\epsilon_{mnp}q_i^m(x)q_j^n(x)q_k^p(x). \tag{7.12}$$

These C-spin singlet fields acting on the vacuum create states which belong to the octet and decuplet representations of the unitary spin SU(3). With the suppressed spin indices included they belong to the correct 56-dimensional representation of SU(6). Any combination of Φ_{ij} and Ψ_{ijk}, Ψ_{ijk}^* will create states in $\mathcal{H}_L$. Thus, starting from transverse octonionic quark field operators as basic fields, we see that the only states which can be constructed that belong to the longitudinal subspace $\mathcal{H}_L$ correspond to the observed unitary spin multiplets with the correct spin and statistics. By choosing different charge assignment schemes one can show the equivalence of the above scheme to various three triplet quark models, such as the color quark scheme and the Han–Nambu scheme.[6]

8. THE ALGEBRA OF COLORS

Above we have seen that the representations of the Poincaré group and the scalar product select a certain direction in the space of octonions, which we chose as e_7. The remaining octonion units e_A ($A=1,2,\cdots 6$) correspond to the color (transverse) degrees of freedom.[6] Automorphisms that leave the imaginary unit e_7 invariant form the $SU(3)_c$ subgroup of G_2, under which the units e_A transform as the $(3+\bar{3})$ representation. The units do not close under the ordinary octonionic product or under the Lie product. The natural product under which they close is given by their associator with the fixed imaginary unit e_7, i.e., defining

$$e_A * e_B \equiv \tfrac{1}{2}[e_A, e_B, e_7] = \tfrac{1}{2}[(e_A e_B)e_7 - e_A(e_B e_7)]. \tag{8.1}$$

From the alternativity of the associator it follows that this product is antisymmetric

$$e_A * e_B = -e_B * e_A.$$

Designating the units e_i as k_i and e_{i+3} as j_i ($i=1,2,3$) we find that

$$j_i * j_j = \epsilon_{ijk}j_k,\quad k_i * k_j = -\epsilon_{ijk}j_k, \tag{8.2}$$

$$j_i * k_j = -\epsilon_{ijk} k_k.$$

These algebraic relations look extremely similar to the Lie algebra of SO(4),

$$\begin{aligned}[][L_i, L_j] &= -\epsilon_{ijk} L_k,\\ [K_i, K_j] &= -\epsilon_{ijk} L_k,\\ [L_i, K_j] &= -\epsilon_{ijk} K_k,\end{aligned} \tag{8.3}$$

or to the Lie algebra of SO(3, 1),

$$\begin{aligned}[][\mathcal{L}_i, \mathcal{L}_j] &= -\epsilon_{ijk} \mathcal{L}_k,\\ [\mathcal{K}_i, \mathcal{K}_j] &= \epsilon_{ijk} \mathcal{L}_k,\\ [\mathcal{L}_i, \mathcal{K}_j] &= -\epsilon_{ijk} \mathcal{K}_k.\end{aligned} \tag{8.4}$$

However, there is a fundamental difference between the algebra defined above and the Lie algebras of SO(4) or SO(3, 1), namely that the former is not a Lie algebra since the Jacobi identity is not satisfied. Yet the Jacobian

$$\begin{aligned} J(e_A, e_B, e_C) \equiv (e_A * e_B) * e_C + (e_C * e_A) * e_B \\ + (e_B * e_C) * e_A \end{aligned} \tag{8.5}$$

is an alternating function of its arguments and hence this algebra with an antisymmetric product and an alternating Jacobian is a six-dimensional Malcev algebra.[31] While the above Lie algebras have the automorphism groups SO(4) and SO(3, 1) respectively, this six-dimensional Malcev algebra has the automorphism group SU(3).

Since the elements of this six-dimensional Malcev algebra correspond to the color degrees of freedom of the color quark scheme,[6] we will call it the "algebra of colors." Its automorphism group will be the color (C-spin) group $SU(3)_c$. The units e_A themselves are not color eigenstates. The color eigenstates are the split octonion units $u_i = \frac{1}{2}(e_i + ie_{i+3})$ and $u_i^* = \frac{1}{2}(e_i - ie_{i+3})$ ($i = 1, 2, 3$), which transform as the 3 and $\bar{3}$ representation of $SU(3)_c$, respectively. Taking as the product of two split units, their associator with the unit ie_7, we find that the color algebra takes the form

$$\begin{aligned} u_i \vee u_j &= \epsilon_{ijk} u_k^*,\\ u_i^* \vee u_j^* &= -\epsilon_{ijk} u_k, \quad i, j, k = 1, 2, 3,\\ u_i \vee u_j^* &= 0,\end{aligned} \tag{8.6}$$

where

$$s \vee s' \equiv \tfrac{1}{2}[s, s', ie_7].$$

ACKNOWLEDGMENTS

The main results of this article were obtained as part of my thesis work at Yale University. I would like to express my deep gratitude to my adviser Professor F. Gürsey for his invaluable suggestions and encouragement. I thank Professor L.A. Radicati for the hospitality extended to me at Scuola Normale Superiore where part of this work was done. Finally I thank Professor K. Bleuler for his hospitality in Bonn and the Alexander von Humboldt Foundation for the granting of a fellowship.

APPENDIX: A NEW FORM OF THE SU(3) AUTOMORPHISMS OF THE REAL OCTONION ALGEBRA

In a previous publication,[16] it was shown that a basis of the split octonion algebra can be chosen as follows:

$$\begin{aligned} u_0 &= \tfrac{1}{2}(1 + ie_7), \quad u_i = \tfrac{1}{2}(e_i + ie_{i+3}) = \tfrac{1}{2}(1 + ie_7)e_i,\\ u_0^* &= \tfrac{1}{2}(1 - ie_7), \quad u_i^* = \tfrac{1}{2}(e_i - ie_{i+3}) = \tfrac{1}{2}(1 - ie_7)e_i,\\ i &= 1, 2, 3,\end{aligned} \tag{A1}$$

where the imaginary unit i commutes with all e_A ($A = 1, \ldots, 7$). From the multiplication rule for e_A it follows that

$$\begin{aligned} u_0^2 &= u_0, \quad u_i u_0 = 0, \quad u_0 u_i = u_i,\\ u_0^{*2} &= u_0^*, \quad u_i u_0^* = u_i, \quad u_0^* u_i = 0,\\ u_i u_j &= \epsilon_{ijk} u_k^*, \quad u_i u_j^* = -\delta_{ij} u_0,\end{aligned} \tag{A2}$$

together with the identities obtained by the conjugation ($i \to -i$).

Under the SU(3) subgroup of the automorphism group G_2 that leaves the imaginary unit e_7 (or equivalently the idempotents u_0 and u_0^*) invariant the units u_i and u_i^* ($i = 1, 2, 3$) transform like a triplet and an anti-triplet, respectively,[16]

$$\text{SU(3)}: \begin{aligned} u_0 \to u_0, \quad u_i \to u_i' = \sum_{j=1}^{3} U_{ij}(-i) u_j,\\ u_0^* \to u_0^*, \quad u_i^* \to u_i^{*\prime} = \sum_{j=1}^{3} U_{ij}(i) u_j^*,\end{aligned} \tag{A3}$$

where $U(-i)$ is the complex conjugate of the matrix $U(i)$. However, from the multiplication rule for e_A it follows that

$$e_7 u_k = -iu_k, \quad e_7 u_k^* = iu_k^*.$$

Therefore, the above equations can be written as

$$\begin{aligned} u_i' &= \tfrac{1}{2}(1 + ie_7)e_i' = \sum_{j=1}^{3} U_{ij}(e_7) u_j\\ &= \tfrac{1}{2}(1 + ie_7)\sum_{j=1}^{3} U_{ij}(e_7) e_j,\\ u_i^{*\prime} &= \tfrac{1}{2}(1 - ie_7)e_i' = \sum_{j=1}^{3} U_{ij}(e_7) u_j^*\\ &= \tfrac{1}{2}(1 - ie_7)\sum_{j=1}^{3} U_{ij}(e_7) e_j,\end{aligned} \tag{A4}$$

which imply

$$\text{SU(3)}: \quad e_i \to e_i' = \sum_{j=1}^{3} U_{ij}(e_7) e_j, \tag{A5}$$

where $U_{ij}(e_7)$ is the 3×3 special unitary matrix written in terms of the complex unit e_7. From here on we shall write $U_{ij}(e_7)$ simply as U_{ij}.

From this new form of SU(3) automorphisms of octonions leaving an imaginary unit (e_7 in this case) invariant, it follows that if the units e_7 and e_i generate a real octonion algebra, so do the units e_7 and $e_i' = \sum_{j=1}^{3} U_{ij} e_j$.

Using the multiplication rule

$$e_i' e_j' = -\delta_{ij} + \epsilon_{ijk} e_k'$$

and

$$e_i e_j = -\delta_{ij} + \epsilon_{ijk} e_k,$$

one finds the following interesting identity for 3×3 special unitary matrices U_{ij}:

$$\sum_m \epsilon_{ijm} U_{mk} = \sum_{m,l} \epsilon_{lmk} U^*_{il} U^*_{jm},$$

where U^* stands for complex conjugate of the matrix U.

*Fellow of the Alexander von Humboldt-Stiftung.

[1]H.D. Politzer, Phys. Rev. Lett. 30, 1346 (1973); D. Gross and F. Wilczek, Phys. Rev. Lett. 30, 1343 (1973); G. t'Hooft, unpublished.

[2]M.Y. Han and Y. Nambu, Phys. Rev. 139, B1006 (1965).

[3]M. Gell-Mann, Acta. Phys. Austr. Suppl. 9, 733 (1972); H. Fritzsch and M. Gell-Mann, in *Proceedings of the XVI International Conference on High Energy Physics, Chicago-Batavia, Ill.* 1972, edited by J.D. Jackson and A. Roberts (NAL, Batavia, Ill., 1973), Vol. 2, p. 135; M. Gell-Mann, *ibid.*, Vol. 4, p. 333.

[4]O.W. Greenberg, Phys. Rev. Lett. 13, 598 (1964).

[5]M.Y. Han and Y. Nambu, Phys. Rev. D 10, 674 (1974).

[6]M. Günaydin and F. Gürsey, Phys. Rev. D 9, 3387 (1974).

[7]F. Gürsey, in *Proceedings of Johns Hopkins University Workshop*, edited by G. Domokos and S.K. Domokos (Johns Hopkins University, Baltimore, 1974).

[8]M. Günaydin, Ph.D. thesis, Yale University (1973) (unpublished).

[9]M. Günaydin and F. Gürsey, Nuovo Cimento Lett. 6, 401 (1973).

[10]R. Casalbuoni, G. Domokos, and S. Kövesi-Domokos, The Johns Hopkins University preprint (1975).

[11]F. Gürsey, in Proceedings of the Kyoto Conference on Mathematical Problems in Theoretical Physics, Kyoto, Japan, 1975.

[12]F. Gürsey, P. Ramond, and P. Sikivie, Yale University Preprint (coo-3075-113), to be published.

[13]See W. Tanenbaum *et al.*, Phys. Rev. Lett. 35, 1323 (1975), and the references therein.

[14]See Benvenuti *et al.*, Phys. Rev. Lett. 35, 1203 (1975), and the references therein.

[15]For the references to the literature on octonions see, e.g., R.D. Schafer, *An Introduction to Nonassociative Algebras* (Academic, New York, 1966) and Ref. 16.

[16]M. Günaydin and F. Gürsey, J. Math. Phys. 14, 1651 (1973).

[17]This conjugation corresponds to an automorphism of the quaternion subalgebras generated by the imaginary units $(e_i, e_7, e_{i+3}=e_7 e_i)$ for $i=1,2,3$ and it corresponds to an anti-automorphism of the other possible quaternion subalgebras.

[18]By bilinearity here, we mean linearity of both of the arguments of (X, Y) with respect to multiplication by real numbers.

[19]However, this bilinear form has a property unique to octonions namely $\bar{X}$, Y and (X, Y) are in triality with each other. For definition of triality and details see Refs. 8 and 16 and M. Günaydin, Scuola Normale Superiore preprint (1975), to be published in Nuovo Cimento.

[20]This also guarantees the associativity of scalar multiplication which is part of the definition of a Hilbert space.

[21]An equivalent construction over the split octonionic Hilbert space was given in Ref. 9.

[22]See E.P. Wigner, "Unitary Representations of the Inhomogeneous Lorentz Group including Reflections," in *Group Theoretical Concepts and Methods in Elementary Particle Physics*, edited by F. Gürsey (Gordon and Breach, New York, 1964).

[23]Here we should note that the necessity of fixing an imaginary unit, in our case e_7, is not only dictated by the invariance of the complex scalar product under the automorphism group of the Hilbert space but also by the requirement that the translations be unitarily implementable. In fact the representation matrices of any group action on our octonionic Hilbert space can involve at most one imaginary unit since the group action must be associative.

[24]If one uses the split octonions as the underlying composition algebra, such inconsistent product states vanish automatically; see Refs. 6, 9, and 16.

[25]G. Birkhoff and J. von Neumann, Ann. Math. 37, 823 (1936).

[26]J.M. Jauch, *Foundations of Quantum Mechanics* (Addison-Wesley, Reading, Mass., 1968).

[27]P. Jordan, J. von Neumann, and E.P. Wigner, Ann. Math. 35, 29 (1934).

[28]G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley, New York, 1972).

[29]D. Fivel, "A New Approach to the Axiomatic Foundations of Quantum Mechanics," University of Maryland preprint (1973).

[30]For the problem of tensor products for states with quaternionic components, see D. Finkelstein, J.M. Jauch, S. Schiminovitch, and D. Speiser, J. Math. Phys. 3, 207 (1962).

[31]For a review of the properties of Malcev algebras see A.A. Sagle, Trans. Amer. Math. Soc. 101, 426 (1961). See also Ref. 8.

Expansion technique for crystal surface problems*

S. E. Trullinger†

Department of Physics, University of California, Irvine, California 92717
(Received 8 January 1976)

An expansion technique for representing functions localized at crystal surfaces is developed by making use of the Gottlieb polynomials to construct a basis set of orthonormal functions of a discrete variable (the integers from zero to infinity). Using these functions, it is shown that a large class of surface problems can be recast in the form of an eigenvalue problem involving a matrix whose elements can be obtained in closed form. The matrix size can be minimized by varying a parameter upon which the basis functions depend. As an example, the dispersion curve of a ferromagnetic spin wave localized at the (001) surface of an fcc crystal is calculated.

I. INTRODUCTION

In recent years there has been an enormous increase of interest on the part of physicists and chemists in the properties of crystal surfaces, e.g., the crystallographic structure, electronic structure, and elementary excitations of the surface region. In many investigations of these properties, one arrives at a stage in the calculation which requires the determination of some function (or set of functions) which depends on a layer index which specifies the distance of the layer from the surface. Examples are the amplitudes of various kinds of surface excitations such as surface phonons, spin waves, etc., atomic equilibrium positions, and quantities associated with the electronic structure, to mention a few. Usually one must resort to numerical methods for the determination of such quantities. In most cases the function of interest decays from some value at the surface layer to a constant value for layers deep in the bulk of the crystal, and numerical methods which utilize this fact are usually the most efficient.

In this paper, a method is described for representing functions defined for the integers from zero to infinity by an expansion in terms of orthonormal functions of a discrete variable. Since these functions decay exponentially, they provide a convenient basis for representing a function of interest which also decays, i.e., the expansion may be truncated after only a few terms with negligible error. The number of terms required can be minimized by varying a parameter upon which the basis functions depend.

In Sec. II below, a review is given of the properties of these orthogonal functions (the normalized Gottlieb functions), and in particular their use in solving a large class of surface problems is discussed. In Sec. III, these results are applied to a simple example, namely, the calculation of the dispersion curve of a ferromagnetic spin wave localized at the (001) surface of an fcc crystal. In Sec. IV, some general features of the method are discussed.

II. THE GOTTLIEB FUNCTIONS

In this section, the Gottlieb[1] polynomials are used to construct a set of orthonormal functions (Gottlieb functions) of a discrete variable[2] (the integers from zero to infinity). Several useful relations involving these functions are derived from the corresponding relations for the polynomials. By using these functions, it is then shown that systems of linear equations of the type encountered in many surface problems can be recast in the form of an eigenvalue problem involving a matrix whose elements can be obtained in closed form.

The Gottlieb polynomials $l_n(x;\lambda)$ are defined[1] by the following finite difference analog of Rodrigues' formula:

$$\exp(-\lambda x) l_n(x;\lambda) = \Delta^n \left\{ \exp(-\lambda x) \binom{x}{n} \right\}, \tag{2.1}$$

where $\lambda > 0$ is a parameter upon which the coefficients in the polynomials depend, Δ is the operator of forward differences, i.e.,

$$\Delta f(x) = f(x+1) - f(x), \tag{2.2}$$

$$\Delta^{n+1} f(x) = \Delta\{\Delta^n f(x)\}, \tag{2.3}$$

and $\binom{x}{n}$ denotes a binomial coefficient. The polynomial $l_n(x;\lambda)$ is of nth degree in the variable x. These polynomials satisfy the orthogonality and normalization conditions[1]

$$\sum_{m=0}^{\infty} \exp(-\lambda m) l_p(m;\lambda) l_q(m;\lambda) = \delta_{p,q} \exp(-p\lambda)(1-\exp(-\lambda))^{-1}$$

$$(p, q = 0, 1, 2, \cdots). \tag{2.4}$$

The summation in Eq. (2.4) is over all integers m from zero to infinity.

Gottlieb[1] has derived many useful properties of these polynomials. We simply list some of these here for easy reference. An explicit form for $l_n(x;\lambda)$ is given by

$$l_n(x;\lambda) = \exp(-n\lambda) \sum_{m=0}^{n} (1-e^{\lambda})^m \binom{n}{m} \binom{x}{m}. \tag{2.5}$$

For integer arguments, a symmetry relation is given by

$$\exp(n\lambda) l_n(m;\lambda) = \exp(m\lambda) l_m(n;\lambda). \tag{2.6}$$

A recurrence formula is

$$(n+1) l_{n+1}(x;\lambda) - [(n+1)e^{-\lambda} + n + (e^{-\lambda} - 1)x] l_n(x;\lambda) + n e^{-\lambda} l_{n-1}(x;\lambda) = 0. \tag{2.7}$$

The polynomials satisfy the following finite difference equation:

$$e^{-\lambda}(x+2)\Delta^2 l_n(x;\lambda) - [(1-e^{-\lambda})x+(n-2)e^{-\lambda}-(n-1)]\Delta l_n(x;\lambda) + n(1-e^{-\lambda})l_n(x;\lambda)=0. \tag{2.8}$$

A generating function is given by

$$G(x;\lambda;z)=\sum_{n=0}^{\infty} l_n(x;\lambda)z^n = (1-z)^x(1-e^{-\lambda}z)^{-x-1} \quad (|z|<1). \tag{2.9}$$

Gottlieb has also noticed a connection with the Laguerre polynomials in the limit $\lambda \to 0$:

$$\lim_{\lambda\to 0} l_n(x/\lambda;\lambda)=L_n(x). \tag{2.10}$$

For our purposes it is convenient to introduce orthonormal functions (Gottlieb functions) defined in terms of the Gottlieb polynomials $l_n(m;\lambda)$ by

$$\varphi_n(m;\lambda)\equiv \exp(\lambda n/2)(1-e^{-\lambda})^{1/2}\exp(-\lambda m/2)l_n(m;\lambda) \quad (\lambda>0), \tag{2.11}$$

where m is an integer $(0\le m<\infty)$. From Eq. (2.4), we see that these functions satisfy the orthonormality condition

$$\sum_{m=0}^{\infty} \varphi_p(m;\lambda)\varphi_q(m;\lambda)=\delta_{p,q}. \tag{2.12}$$

The relations involving $\varphi_n(m;\lambda)$ which are analogous to to Eqs. (2.5), (2.6), (2.7), (2.8), and (2.9) are easily found to be

$$\varphi_n(m;\lambda)=\exp[-\lambda(n+m)/2](1-e^{-\lambda})^{1/2} \times \sum_{p=0}^{n}(1-e^{\lambda})^p\binom{n}{p}\binom{m}{p}, \tag{2.13}$$

$$\varphi_n(m;\lambda)=\varphi_m(n;\lambda), \tag{2.14}$$

$$(n+1)\varphi_{n+1}(m;\lambda)-\exp(\lambda/2)[(n+1)e^{-\lambda}+n+(e^{-\lambda}-1)m] \times\varphi_n(m;\lambda)+n\varphi_{n-1}(m;\lambda)=0, \tag{2.15}$$

$$(m+2)\Delta^2\varphi_n(m;\lambda)+[(m+2)(2-\exp(-\lambda/2)) -(m+1)\exp(\lambda/2)+(\exp(\lambda/2)-\exp(-\lambda/2))n]\Delta\varphi_n(m;\lambda) +[(m+2)(1-\exp(-\lambda/2))+(m+1)\exp(\lambda/2) +(\exp(\lambda/2)-\exp(-\lambda/2))n]\varphi_n(m;\lambda)=0, \tag{2.16}$$

$$g(m;\lambda;z)=\sum_{n=0}^{\infty}\varphi_n(m;\lambda)z^n=\sum_{n=0}\varphi_m(n;\lambda)z^n =\exp(-\lambda m/2)(1-e^{-\lambda})^{1/2}(1-\exp(\lambda/2)z)^m \times(1-\exp(-\lambda/2)z)^{-m-1} \quad (|z|<\exp(-\lambda/2)). \tag{2.17}$$

We now discuss the use of the Gottlieb functions $\varphi_n(m;\lambda)$ in obtaining solutions to a large class of surface problems. This class of problems is such that the equations which determine some unknown function, $f_\alpha(m)$ can be written in the form

$$\Omega f_\alpha(m)=\sum_\beta \mathcal{L}_{\alpha\beta}(m)f_\beta(m), \tag{2.18}$$

where $\mathcal{L}_{\alpha\beta}(m)$ is a linear operator of the form

$$\mathcal{L}_{\alpha\beta}(m)=b_{\alpha\beta}+\sum_{n=1}^{M}c_{\alpha\beta}^{(n)}E^n+\sum_{n=1}^{N}d_{\alpha\beta}^{(n)}\theta(m-n)E^{-n} +\sum_{n=0}^{I}\sum_{l=0}^{J}e_{\alpha\beta}^{(nl)}\delta_{m,n}E^l+\sum_{n=1}^{R}\sum_{l=0}^{S}h_{\alpha\beta}^{(nl)} \times\delta_{m,n}\theta(m-l)E^{-l}+\sum_{n=1}^{T}g_{\alpha\beta}^{(n)}\theta(m-n). \tag{2.19}$$

The subscript α attached to the function f serves to label each member in the set of unknown functions to be determined. Equation (2.18) thus represents a set of coupled equations for the unknowns $f_\alpha(m)$ $(m=0,1,2,\cdots)$. The integer variable m serves as a layer index, where $m=0$ denotes the surface layer and large values of m serve to label layers deep in the interior of the semi-infinite crystal. In Eq. (2.19), the quantities $b_{\alpha\beta}$, $c_{\alpha\beta}^{(n)}$, $d_{\alpha\beta}^{(n)}$, $e_{\alpha\beta}^{(nl)}$, $h_{\alpha\beta}^{(nl)}$, and $g_{\alpha\beta}^{(n)}$ are assumed to be independent of m. The stepping operator E is defined by

$$E^n f(m)=f(m+n) \tag{2.20}$$

and its inverse by

$$E^{-n}f(m)=f(m-n) \quad (n\le m). \tag{2.21}$$

The Heaviside step function is defined by

$$\theta(l)\equiv\begin{cases}0, & l<0,\\ 1, & l\ge 0.\end{cases} \tag{2.22}$$

The quantity Ω appearing in Eq. (2.18) is an eigenvalue of the operator $\mathcal{L}$: for example, it might be the frequency of a surface excitation or the energy of a surface electronic state. The integers appearing as limits on the sums in Eq. (2.19) are such that $M,N,R,T\ge 1$ and $I,J,S\ge 0$. In most cases these integers will be finite and small. A simple example of such an $\mathcal{L}$ operator is given in Sec. III below.

We now expand the unknown functions $f_\alpha(m)$ in terms of the Gottlieb functions:

$$f_\alpha(m)=\sum_{q=0}\mathcal{A}_q^{(\alpha)}(\lambda)\varphi_q(m;\lambda) \quad (m=0,1,2,\cdots). \tag{2.23}$$

If the functions $f_\alpha(m)$ were known, then the expansion coefficients $\mathcal{A}_q^{(\alpha)}(\lambda)$ would be given by

$$\mathcal{A}_q^{(\alpha)}(\lambda)=\sum_{m=0}^{\infty}f_\alpha(m)\varphi_q(m;\lambda), \tag{2.24}$$

where use has been made of Eq. (2.12). The conditions which $f_\alpha(m)$ must satisfy in order for the expansion (2.23) to exist have been discussed by Gottlieb.[1] As long as $f_\alpha(m)$ decays in an over-all fashion to zero for $m\to\infty$, then the expansion is well defined. By substituting Eq. (2.23) into (2.18) and making use of Eq. (2.12), we obtain the following set of equations:

$$\Omega\mathcal{A}_p^{(\alpha)}(\lambda)=\sum_\beta\sum_{q=0}^{\infty}A_{pq}^{(\alpha\beta)}(\lambda)\mathcal{A}_q^{(\beta)}(\lambda) \quad (p=0,1,2,\cdots), \tag{2.25}$$

where the matrix elements $A_{pq}^{(\alpha\beta)}(\lambda)$ are given by

$$A_{pq}^{(\alpha\beta)}(\lambda)=\sum_{m=0}^{\infty}\varphi_p(m;\lambda)\mathcal{L}_{\alpha\beta}(m)\varphi_q(m;\lambda)\equiv\langle p|\mathcal{L}_{\alpha\beta}(m)|q\rangle. \tag{2.26}$$

Among the eigenvalues Ω of the matrix **A** will appear the surface solutions of interest. The corresponding eigenvectors, $\mathcal{A}$, when substituted into Eq. (2.23), yield the desired surface solutions for $f_\alpha(m)$. This same basic approach has been employed in the continuum limit $[f(m)\to f(x)]$ for several problems involving edges of crystals.[3]

In the strict sense, the matrix **A** is of infinite dimension. However, in most cases the parameter λ may be chosen so that one introduces negligible error in $f_\alpha(m)$ by truncating the series (2.23) after a few terms, so that

$$f_\alpha(m;\lambda)\approx\sum_{q=0}^{q_{\max}}\mathcal{A}_q^{(\alpha)}\varphi_q(m;\lambda). \tag{2.27}$$

If $q_{\max}$ is finite, then $f_\alpha(m)$ depends on λ as is indicated in Eq. (2.27). It is only in the limit $q_{\max}\to\infty$ that $f_\alpha(m;\lambda)$ is independent of λ. This fact is important to remember in practical applications of the method. From Eq. (2.11), we see that the parameter λ provides a rough measure of the rate of decay of $\varphi_n(m;\lambda)$ with increasing m. Thus, it also measures the rate of decay of $f_\alpha(m;\lambda)$, provided that $q_{\max}$ is fairly small. It seems reasonable then to choose a value of λ which will mimic the expected decay rate of $f_\alpha(m)$ since $q_{\max}$ will then be as small as possible. This in turn leads to the smallest possible dimensions of the matrix **A** which is to be diagonalized. We return to these considerations later in Sec. III in connection with a specific example.

By making use of Eqs. (2.26) and (2.19), the elements of the matrix **A** can be written in the form

$$\begin{aligned}A_{pq}^{(\alpha\beta)}(\lambda)=&\,b_{\alpha\beta}\delta_{p,q}+\sum_{n=1}^{M}c_{\alpha\beta}^{(n)}\langle p|E^n|q\rangle\\&+\sum_{n=1}^{N}d_{\alpha\beta}^{(n)}\langle p|\theta(m-n)E^{-n}|q\rangle\\&+\sum_{n=0}^{I}\sum_{l=0}^{J}e_{\alpha\beta}^{(nl)}\langle p|\delta_{m,n}E^l|q\rangle\\&+\sum_{n=1}^{R}\sum_{l=0}^{S}h_{\alpha\beta}^{(nl)}\langle p|\delta_{m,n}\theta(m-l)E^{-l}|q\rangle\\&+\sum_{n=1}^{T}g_{\alpha\beta}^{(n)}\langle p|\theta(m-n)|q\rangle,\end{aligned} \tag{2.28}$$

where

$$\langle p|E^n|q\rangle=\exp[-\lambda(n+q-p)/2]\sum_{j=0}^{n}(-1)^j\binom{n}{j}\binom{n+q-p-j-1}{n-1}\times\exp(\lambda j)\theta(q-p-j), \tag{2.29a}$$

$$\langle p|\theta(m-n)E^{-n}|q\rangle=\langle q|E^n|p\rangle, \tag{2.29b}$$

$$\langle p|\delta_{m,n}E^l|q\rangle=\varphi_p(n;\lambda)\varphi_q(n+l;\lambda), \tag{2.29c}$$

$$\langle p|\delta_{m,n}\theta(m-l)E^{-l}|q\rangle=\theta(n-l)\varphi_p(n;\lambda)\varphi_q(n-l;\lambda), \tag{2.29d}$$

$$\langle p|\theta(m-n)|q\rangle=\delta_{p,q}-\sum_{m=0}^{n-1}\varphi_p(m;\lambda)\varphi_q(m;\lambda). \tag{2.29e}$$

A derivation of Eq. (2.29a) is presented in the Appendix.

III. EXAMPLE

In this section, the method described in the previous section is used to determine the dispersion curve of a ferromagnetic spin wave localized at the (001) surface of a semi-infinite fcc crystal with nearest- and next-nearest-neighbor exchange interactions between spins. To the author's knowledge, an exact solution of this problem does not exist. The following analysis thus provides a nontrivial and yet simple example of the use of the method described in Sec. II. The equations governing the layer-dependent spin-wave amplitude, $S(m)$, are easily obtained and have the form

$$\Omega S(m)=\mathcal{L}(m)S(m), \tag{3.1}$$

where

$$\begin{aligned}\mathcal{L}(m)=&\,b+c^{(1)}E^1+c^{(2)}E^2+d^{(1)}\theta(m-1)E^{-1}\\&+d^{(2)}\theta(m-2)E^{-2}+g^{(1)}\theta(m-1)+g^{(2)}\theta(m-2)\end{aligned} \tag{3.2}$$

$(m=0,1,2,\cdots)$,

with

$$b\equiv 8-2\varphi(\mathbf{k}_\parallel)+5r-2r\chi(\mathbf{k}_\parallel) \tag{3.3a}$$

$$c^{(1)}=d^{(1)}\equiv-\psi(\mathbf{k}_\parallel), \tag{3.3b}$$

$$c^{(2)}=d^{(2)}\equiv-r, \tag{3.3c}$$

$$g^{(1)}\equiv 4, \tag{3.3d}$$

$$g^{(2)}\equiv r, \tag{3.3e}$$

and

$$r\equiv J_2/J_1, \tag{3.4a}$$

$$\varphi(\mathbf{k}_\parallel)\equiv\cos k_xa_0+\cos k_ya_0, \tag{3.4b}$$

$$\chi(\mathbf{k}_\parallel)\equiv\cos[(k_x+k_y)a_0]+\cos[(k_x-k_y)a_0] \tag{3.4c}$$

$$\psi(\mathbf{k}_\parallel)\equiv 1+\varphi(\mathbf{k}_\parallel)+\cos[(k_x+k_y)a_0], \tag{3.4d}$$

$$\Omega=\omega/SJ_1. \tag{3.4e}$$

In these expressions, a_0 is the periodicity length in two directions parallel to the surface (x–y plane), J_1 and J_2 denote nearest- and next-neighbor exchange integrals, ω is the frequency of oscillation, and $\mathbf{k}_\parallel=(k_x,k_y)$ is the wave vector of the spin wave. For simplicity it is assumed that J_1 and J_2 in the surface region have the same values as in the bulk. It is easy to include any changes in these quantities through the constants $e^{(nl)}$ and $h^{(nl)}$ appearing in Eq. (2.19).

Following the procedure outlined in Sec. II, we approximate $S(m)$ by

$$S(m)\approx\sum_{q=0}^{q_{\max}}\mathcal{A}_q(\lambda)\varphi_q(m;\lambda). \tag{3.5}$$

This leads to the equation

$$\Omega(\lambda;q_{\max})\mathcal{A}_p(\lambda;q_{\max})=\sum_{q=0}^{q_{\max}}A_{pq}(\lambda;q_{\max}) \tag{3.6}$$

with

$$A_{pq}(\lambda)=[8-2\varphi(\mathbf{k}_\parallel)+5r-2r\chi(\mathbf{k}_\parallel)]\delta_{p,q}+4\langle p|\theta(m-1)|q\rangle+r\langle p|\theta(m-2)|q\rangle$$

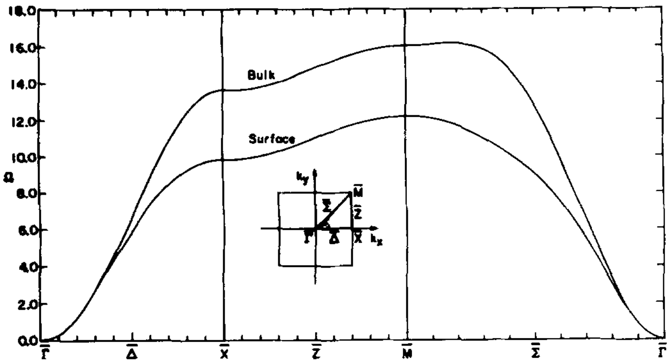


FIG. 1. Dispersion curve for the spin wave localized at the (001) surface of an fcc crystal ($J_2/J_1=0.2$). Also shown is the bulk dispersion curve with $\mathbf{k}_{\text{bulk}}=\mathbf{k}_{\parallel}$. The inset shows the surface Brillouin zone and its irreducible segment.

$$-\psi(\mathbf{k}_{\parallel})[\langle p|E|q\rangle+\langle q|E|p\rangle]$$
$$-r[\langle p|E^2|q\rangle+\langle q|E^2|p\rangle], \qquad (3.7)$$

where use has been made of Eq. (2.29b). The eigenvalues $\Omega(\lambda; q_{\max})$ are independent of λ and $q_{\max}$ only in the limit as $q_{\max}\to\infty$. However, we shall see that a good choice of λ will cause the lowest eigenvalue, $\Omega_1(\lambda; q_{\max})$, to converge rapidly as $q_{\max}$ increases. This lowest eigenvalue yields the frequency of oscillation of the surface spin wave [see Eq. (3.4e)] and the corresponding eigenvector yields the layer-dependent amplitude of the surface spin wave [see Eq. (3.5)].

With the help of Eqs. (2.29a) and (2.29e) the matrix elements $A_{pq}(\lambda)$ given by Eq. (3.7) can be obtained in closed form. This was done and a computer code was written which generated and diagonalized the nonsymmetric $(q_{\max}+1)\times(q_{\max}+1)$ real matrix $\mathbf{A}$ for arbitrary values of λ (>0) and $\mathbf{k}_{\parallel}$. The results for the lowest eigenvalue when $r\equiv J_2/J_1=0.2$ are plotted in Fig. 1, together with the bulk spin wave dispersion curve for $\mathbf{k}_{\text{bulk}}=\mathbf{k}_{\parallel}$. A general feature of surface spin waves[4] is that their amplitude decays very slowly into the bulk if $\mathbf{k}_{\parallel}$ is close to zero (the Brillouin zone center, $\bar{\Gamma}$). For $\mathbf{k}_{\parallel}$ near the zone boundary, however, the decay can be quite rapid. Thus, it was expected that the optimum value of λ giving the most rapid convergence of Ω as $q_{\max}$ is increased would depend on $\mathbf{k}_{\parallel}$, assuming small values near the zone center and large values near the zone boundary. This was found to be the case; for $\mathbf{k}_{\parallel}$ close to zero, $\exp(-\lambda_{\text{opt}})$ is very close to unity, while, for $\mathbf{k}_{\parallel}=(\pi/a_0, \pi/a_0)$, $\exp(-\lambda_{\text{opt}})\approx 0.02$. For the latter case, we have shown in Table I the dependence of Ω on $q_{\max}$ for $e^{-\lambda}=0.02$. The convergence is seen to be quite rapid.[5] Indeed, for all $\mathbf{k}_{\parallel}$ it was found that λ could be chosen so that Ω was accurate to at least three or four significant digits with $q_{\max}\lesssim 5$ which means that the matrix $\mathbf{A}$ need not be larger than 6×6 to obtain this kind of accuracy. It should be emphasized, however, that it is not necessary to determine λ_{opt} for each $\mathbf{k}_{\parallel}$. By fixing $e^{-\lambda}=0.5$, the values of Ω obtained are accurate to at least four significant digits with $q_{\max}\approx 25$ for all $\mathbf{k}_{\parallel}$ except those very close to zero. Simply stated, this means that the accuracy that is lost if $\lambda\neq\lambda_{\text{opt}}$ can be regained by increasing $q_{\max}$.

IV. DISCUSSION

From the discussion in Secs. II and III, we see that the use of Gottlieb functions leads to a very simple procedure for solving a large class of surface problems. To what extent is this procedure more efficient in a numerical sense than other methods which are available? The answer to this question depends on the particular nature of the problem under study. However, a few general remarks can be made. The most common method used to date for solving a variety of surface problems employs a slab of finite thickness (typically 20–30 layers) to serve as a model for the semi-finite crystal. The primary disadvantage with the use of such slabs is that they must be thick compared to the decay length of the function of interest. Otherwise the two surfaces "interfere" with each other. Since the method described here deals directly with semi-infinite crystals (one surface only), it suffers no such limitation. While in slab calculations the size of the matrices to be diagonalized increases (in some cases quite rapidly) with slab thickness, the necessary size of the matrices in the present method depends only on how many Gottlieb functions are needed to accurately represent the function of interest. As we have seen in Sec. III, the required matrix size in the present method can be minimized with a suitable choice(s) of the parameter λ.

Finally, it should be emphasized that the use of Gottlieb functions is not restricted to the class of problems discussed here. Since they provide a convenient set of basis functions for representing decaying functions of integer argument, they should prove useful in other types of problems involving crystal surfaces as well.[6]

ACKNOWLEDGMENT

The author wishes to thank A. A. Maradudin for helpful discussions.

APPENDIX

In this appendix, Eq. (2.29a) of the text is derived by making use of the generating function for the Gottlieb functions. The matrix element $\langle p|E^n|q\rangle$ is defined by

$$\langle p|E^n|q\rangle \equiv \sum_{m=0}^{\infty} \varphi_p(m;\lambda)E^n\varphi_q(m;\lambda). \qquad \text{(A1)}$$

By using Eq. (2.20), this becomes

TABLE I. Rate of convergence of Ω with increasing matrix size $(q_{\max}+1)\times(q_{\max}+1)$ for $\mathbf{k}_{\parallel}=\bar{M}$ (zone corner) and $e^{-\lambda}=0.02$.

$q_{\max}$	Ω ($k_x=k_y=\pi/a_0$)
1	12.208 305 500 10
2	12.192 409 326 18
3	12.190 717 676 43
4	12.190 505 406 43
5	12.190 479 749 08
6	12.190 476 623 82
7	12.190 476 243 20
8	12.190 476 196 77
9	12.190 476 191 06
10	12.190 476 190 60
11	12.190 476 190 22

$$\langle p | E^n | q\rangle = \sum_{m=0}^{\infty} \varphi_p(m;\lambda)\varphi_q(m+n;\lambda). \qquad \text{(A2)}$$

To evaluate the right-hand side of Eq. (A2), we first multiply both sides by $s^p t^q$, where s and t are independent quantities such that $|s|, |t| < \exp(-\lambda/2)$, and then sum over the integers p and q from zero to infinity:

$$\sum_{p=0}^{\infty}\sum_{q=0}^{\infty} \langle p | E^n | q\rangle s^p t^q = \sum_{m=0}^{\infty} \Big(\sum_{p=0}^{\infty} \varphi_p(m;\lambda)s^p\Big)\Big(\sum_{q=0}^{\infty} \varphi_q(m+n;\lambda)t^q\Big). \qquad \text{(A3)}$$

By making use of the generating function [Eq. (2.17)] the right-hand side (RHS) of Eq. (A3) becomes

$$\text{RHS} = \exp(-\lambda n/2)(1-e^{-\lambda})(1-\exp(-\lambda/2)s)^{-1} \times (1-\exp(\lambda/2)t)^n (1-\exp(-\lambda/2)t)^{-n-1} \times \sum_{m=0}^{\infty} \exp(-\lambda m)\left(\frac{(1-\exp(\lambda/2)s)(1-\exp(\lambda/2)t)}{(1-\exp(-\lambda/2)s)(1-\exp(-\lambda/2)t)}\right)^m \qquad \text{(A4)}$$

The geometric series is trivially summed to yield, after after some algebraic simplifications,

$$\sum_{p=0}^{\infty}\sum_{q=0}^{\infty} \langle p | E^n | q\rangle s^p t^q = \exp(-\lambda n/2)\frac{(1-\exp(\lambda/2)t)^n}{(1-\exp(-\lambda/2)t)^n(1-st)}. \qquad \text{(A5)}$$

By expanding the right-hand side of Eq. (A5) in powers of s and t and then equating coefficients of $s^p t^q$ on both sides of the resulting equation, Eq. (2.29a) is obtained.

*Research supported by AFOSR Grant No. 70-1936.

†Present address: Laboratory of Atomic and Solid State Physics, Clark Hall, Cornell University, Ithaca, New York 14853.

[1]M.J. Gottlieb, Am. J. Math. 60, 453 (1938). The Gottlieb polynomials are a special case of the more general Meixner polynomials, see Joseph Meixner, J. London Math. Soc. 9, 6 (1934).

[2]For a discussion of several types of orthogonal polynomials of a discrete variable, see G. Szegö, *Orthogonal Polynomials*, Am. Math. Soc. Colloq. Publications 23 (Am. Math. Soc., Providence, R.I., 1959), rev. ed.

[3]A.A. Maradudin, R.F. Wallis, D.L. Mills, and R.L. Ballard, Phys. Rev. B 6, 1106 (1972); S.L. Moss, A.A. Maradudin, and S.L. Cunningham, Phys. Rev. B 8, 2999 (1973); T.M. Sharon, A.A. Maradudin, and S.L. Cunningham, Phys. Rev. B 8, 6024 (1973); T.M. Sharon and A.A. Maradudin, Sol. State Comm. 13, 187 (1973).

[4]R.F. Wallis, A.A. Maradudin, I.P. Ipatova, and A.A. Klochikhin, Solid State Comm. 5, 89 (1967).

[5]For this particular wave vector $[\mathbf{k}_{\parallel} = (\pi/a_0, \pi/a_0)]$ an exact analytic solution is obtainable by recognizing the fact that layers 1, 3, 5, ··· decouple from layers 0, 2, 4, ···, with the consequence that the spins in layers 1, 3, 5, ··· are stationary. The spin wave amplitude for the even-numbered layers decays with depth as a pure exponential and has a frequency $\Omega = 12 + r - r^2/(4+r)$ which agrees (for $r = 0.2$) with the converged value in Table I. I am grateful to J. Dobson for pointing out the existence of this analytic solution.

[6]Investigations are currently underway concerning their use in low energy electron diffraction (LEED) intensity calculations and also in calculations of the surface band structure of transition metals.

Euclidean field theory. II. Remarks on embedding the relativistic Hilbert space

John L. Challifour

Departments of Mathematics and Physics, Indiana University, Bloomington, Indiana 47401
(Received 14 November 1975)

Starting from extensions of the Schwinger functions to a positive, symmetric linear functional on the Borchers algebra we give a theorem on the embedding of the physical Hilbert space as a closed subspace of a Euclidean Hilbert space. The case of locally L_1 strongly positive extensions is discussed as is the relation to Nelson's Markov property for Euclidean fields.

1. INTRODUCTION

In his abstract formulation of Euclidean quantum field theory, Nelson[1] introduced a Hilbert space for the Euclidean fields as $L_2(\mathcal{S}'_R, \mathfrak{A}, \mu)$ where $\mathcal{S}'_R$ denotes real tempered distributions, $\mathfrak{A}$ the Borel field over $\mathcal{S}'_R$ and μ a probability measure. The relativistic theory was then reconstructed by embedding the physical Hilbert space as the L_2 subspace corresponding to the sub-σ field generated by Euclidean fields localized at sharp time and then exploiting time reversal invariance of this subspace with a Markov property for the Euclidean field. Hegerfeldt[2] starting with a self-adjoint commutative representation of the Schwartz space $\mathcal{S}$ subsequently identified a more general property, "T-positivity," which sufficed for the positivity condition isolated by Osterwalder and Schrader[3] in their equivalence theorem for Euclidean and Wightman theories. Hegerfeldt found the physical Hilbert space to be the subspace on which T-positivity held and showed the relation between Nelson's Markov property for half-spaces and the requirement that the T-positive operator be a projection. The purpose of this note is to investigate this framework for the Euclidean theory which arises when the Schwinger functions (noncoincident arguments) have extensions as a positive, symmetric state on the Borchers algebra over $\mathcal{S}$.[4] To keep the notation to a minimum, we shall follow the conventions of Refs. 3 and 4.

Let $\underline{S}=\{1, S_1, S_2, \ldots, S_n, \cdots\}$ denote a Schwinger state on $\underline{\mathcal{S}}_0=\oplus_n\mathcal{S}(\mathcal{E}_n)$, where $\mathcal{S}(\mathcal{E}_n)$ consists of those Schwartz functions on R^{4n} vanishing with all derivatives unless $x_i \neq x_j$, $1 \leqslant i < j \leqslant n$, $x=(t, \mathbf{x}) \in R^4$. A positive, symmetric extension of $\underline{S}$, denoted ext$\underline{S}$, to the Borchers algebra $\underline{\mathcal{S}}=\oplus_n\mathcal{S}(R^{4n})$ may be assumed to be Euclidean and θ invariant without loss of generality (Sec. 2, Ref. 4). Here θ denotes the time inversion operator. The positive time-ordered subspace is $\underline{\mathcal{S}}_+=\oplus_n\mathcal{S}(\Omega^<_{n,+})$ for which $\mathcal{S}(\Omega^<_{n,+})$ consists of Schwartz functions on R^{4n} vanishing with all derivatives unless $0<t_1<t_2<\cdots<t_n<\infty$. Our Euclidean Hilbert space will be $\mathcal{H}_E:=\overline{\underline{\mathcal{S}}/N^E_E}$, where N_E is the Euclidean kernel and the factor space is closed relative to the topology derived from extS. For $\underline{f} \in \underline{\mathcal{S}}$, let $\psi(\underline{f})$ be the corresponding coset in $\underline{\mathcal{S}}/\overline{N}_E$ and define the reflection operator

$$K_\theta\psi(\underline{f}):=\psi(\theta\underline{f}). \tag{1.1}$$

Due to θ invariance of ext$\underline{S}$, K_θ extends by continuity to a unitary self-adjoint operator on $\mathcal{H}_E$. The notion of "T-positivity" is introduced by means of the subspace $\mathcal{H}_+:=\overline{(\underline{\mathcal{S}}_+ + N_E)/N_E}^E$ and its orthogonal projection P_+. Then defining

$$P:=P_+K_\theta P_+; \tag{1.2}$$

Osterwalder—Schrader positivity, $\underline{S}(\theta\underline{f}^*_-\times\underline{f}_-)\geqslant 0$ for $\underline{f}\in\underline{\mathcal{S}}_+$, is equivalent to $P\geqslant 0$ and the physical Hilbert space, $\mathcal{H}_0$, is unitarily embedded in $\mathcal{H}_E$ as the closure of the range of P. Further, P is a projection only when $\mathcal{H}_0$ consists of the invariant vectors for K_θ in $\mathcal{H}_+$. This is the appropriate abstraction of Nelson's Markov property for half-spaces.[1] Relation (1.2) generalizes Hegerfeldt's T-positivity in that $\mathcal{H}_+$ is generally a proper subspace of $\overline{sp}\{\exp[i\varphi(f)]\,|\,f\in\mathcal{S}(R^4_+)\}$ used in Ref. 2. Moreover since our requirements on ext$\underline{S}$ constitute a minimal setting for the Euclidean field theory appropriate to the Wightman axioms (we do not require the Euclidean field to be self-adjoint on $\mathcal{H}_E$), $P\geqslant 0$ is the Osterwalder—Schrader positivity condition for such. These results constitute the work of Sec. 2.

Section 3 specializes to the situation where each component, ext$\underline{S}_n$, is locally L_1 in the time variables for which $\mathcal{H}_+=\overline{\mathcal{P}(\mathcal{S}(R^4_+))}$; while Sec. 4 treats further the case when ext$\underline{S}$ is a strongly positive, symmetric, continuous extension of $\underline{S}$ leading to a maximal measure[4] on $\mathcal{S}'_R$. For this situation, $\mathcal{H}_+=L_2(\mathcal{S}'_R, \mathfrak{A}^+, \mu)$, where $\mathfrak{A}^+$ is the σ field generated by the Euclidean field $\varphi(f)$, $f\in\mathcal{S}(R^4_+)$, and P_+ is the conditional expectation. The Hopf—Chacon ergodic theorem allows recovery of Nelson's embedding, where $\mathcal{H}_0=L_2(\mathcal{S}'_R, \mathfrak{A}_0, \mu)$ with $\mathfrak{A}_0$ a σ field localized at $t=0$ but the sharp time fields do not necessarily exist. This requires also that P be a projection. Recent work[5] has shown the existence of ext$\underline{S}$ for a large class of Schwinger functions whose behavior near equal times is more singular than locally L_1. A study of P_+ in this case would shed more light on the abstract Markov property and when it can be expected to hold.

2. THE EMBEDDING THEOREM

The Osterwalder—Schrader positivity condition [E2 of Ref. 3] permits the introduction of a second Hilbert space, $\mathcal{H}:=\overline{\underline{\mathcal{S}}_+/N_{\theta+}}^R$, with null space $N_{\theta+}=\{\underline{f}\in\underline{\mathcal{S}}_+\,|\,\underline{S}(\theta\underline{f}^*\times\underline{f})=0\}$ and topology on the cosets $v(\underline{f})\in\underline{\mathcal{S}}/N_{\theta+}$ derived from the sesquilinear form $(v(\underline{f}), v(\underline{g}))_R=\underline{S}(\theta\underline{f}^*\times\underline{g})$. It is shown in Ref. 3 that $\mathcal{H}$ is unitarily equivalent to the physical Hilbert space of the corresponding Wightman theory. Now define a linear map

$E: \mathcal{H}_+ \to \mathcal{H}$ by

$$E\psi(\underline{f}) := v(\underline{f}), \quad \underline{f} \in \underline{S}_+. \tag{2.1}$$

The relations $\|E\psi(f)\|_R = \text{ext}S(\theta f^* \times f)^{1/2} \leq \|K_\theta \psi(f)\|_E^{1/2} \|\psi(f)\|_E^{1/2} = \|\psi(f)\|_E$ allow extension of E by continuity to a contraction with dense range. E^* is an invertible contraction and $\mathcal{H}_+ = N(E) \oplus \overline{R(E^*)}^E$, $N(E)$ is the null space of E and $R(E^*)$ the range of E^*. Several facts about the mappings in (1.1), (1.2), and (2.1) are conveniently summarized at this point.

Proposition 2.1:

(a) $E^*E = P_+ K_\theta P_+$,

(b) $\overline{R(E^*E)^E} = \overline{R(E^*)^E}$,

(c) $\mathcal{H}_+ \cap K_\theta \mathcal{H}_+ \subset R(E^*E)$,

(d) $K_\theta \mathcal{H}_+ \cap N(E) = \{0\}$.

Proof: For (a) notice that $(v(f_-), E\psi(g)) = S(\theta f_-^* \times g) = (K_\theta \psi(f), \psi(g))$ for $f, g \in \underline{S}_+$. Hence $E^*E = P_+ K_\theta P_+$ on a dense set. (b) is a consequence of $N(E)^\perp = N(E^*E)^\perp = \overline{R(E^*E)}^E$. In (c), suppose $u \in \mathcal{H}_+ \cap K_\theta \mathcal{H}_+$. Then $u = K_\theta^2 P_+ u = P_+ K_\theta P_+ K_\theta P_+ u = (E^*E)^2 u$, which also gives (d).

These observations allow the embedding of $\mathcal{H}$ in $\mathcal{H}_E$ in a manner canonical with respect to the coset spaces and generalizes Hegerfeldt's result.[2]

Theorem 2.2 : $\mathcal{H}$ may be identified with $\mathcal{H}_0 := \overline{R(P_+ K_\theta P_+)}^E$.

Proof: Let $E = U(E^*E)^{1/2}$ be the polar decomposition fo E. Then U is partially isometric with initial set $\overline{R[(E^*E)^{1/2}]}^E$ and final set $\overline{R(E)}^R$. Now notice $\overline{R[(E^*E)^{1/2}]}^E = N(E)^\perp = \overline{R(E^*)}^E$ and use (b) above.

To characterize $\mathcal{H}_0$ further, the following remarks will be useful in which we have denoted $\mathcal{H}_- := K_\theta \mathcal{H}_+$.

Lemma 2.3: The following conditions for $\psi \in \mathcal{H}_+$ are equivalent:

(a) $\|E^*E\psi\| = \|\psi\|$,

(b) $\|E\psi\| = \|\psi\|$,

(c) $\psi = E^*E\psi$,

(d) $\psi \in \mathcal{H}_+ \cap \mathcal{H}_-$,

(e) $K_\theta \psi = \psi$.

Proof: It is sufficient to take $\psi \neq 0$ in $\mathcal{H}_0$. Clearly, (a) and (b) are equivalent since E, E^* are contractions and (a) implies (c) since $\|\psi - E^*E\psi\|^2 = 2\|\psi\|^2 - 2\|E\psi\|^2 = 0$. Let $K_\theta \psi = \psi_1 + \psi_2$ with $\psi_1 \in \mathcal{H}_+$, $\psi_2 \in \mathcal{H}_+^\perp$. Then $\|\psi\|^2 = \|K_\theta \psi\|^2 = \|\psi_1\|^2 + \|\psi_2\|^2 = \|P_+ K_\theta P_+ \psi\|^2 + \|\psi_2\|^2 = \|\psi\|^2 + \|\psi_2\|^2$ and (c) implies (d). Finally, $\psi \in \mathcal{H}_+ \cap \mathcal{H}_-$ satisfies $\psi = (E^*E)^2\psi$ so (d) implies (a). (d) and (e) are obvious.

The special case for which P is a projection allows a sharper characterization of $\mathcal{H}_0$.

Proposition 2.4: The following are equivalent:

(a) $\mathcal{H}_0 = \mathcal{H}_+ \cap \mathcal{H}_-$,

(b) E^* is isometric,

(c) $P_+ K_\theta P_+$ is a projection onto $\mathcal{H}_0$.

Proof: (a) implies (b) as $\overline{R(E^*E)}^E = \mathcal{H}_+ \cap \mathcal{H}_-$ requires E isometric on $\mathcal{H}_0$ by Lemma 2.3(b). E^* isometric on the range of E shows for $\underline{f} \in \underline{S}_+$, $\|v(\underline{f})\|^2 = \|E^* v(\underline{f})\|^2 = (E^*E\psi(\underline{f}), \psi(\underline{f}))$. From this it follows that E^*E is a projection onto its range which is then closed, so (b) implies (c). Finally, (c) implies (a) as $R(E^*E) = R(E^*)$ for this case and both E^* and E have closed range. Let $v \in R(E)$ be nonzero. Then there exists $\psi \in R(E^*E)$ for which $v = E\psi$ and $\|\psi\|^2 = (\psi, E^*E\psi) = \|v\|^2$. By Lemma 2.3, $\psi \in \mathcal{H}_+ \cap \mathcal{H}_-$.

From these remarks, we learn that the unitary copy $\mathcal{H}_0$ of the physical Hilbert space may be decomposed into a direct sum $\mathcal{H}_{\text{inv}} \oplus \mathcal{N}$, where $\mathcal{H}_{\text{inv}} = \mathcal{H}_+ \cap \mathcal{H}_-$ contains the vacuum subspace and all vectors in $\mathcal{H}_+$ invariant under the reflexion operator. The subspace $\mathcal{N}$ is a defect space in the sense that E restricted to $\mathcal{N}$ is a strict contraction. Examples of generalized free fields show cases where $\mathcal{H}_{\text{inv}} = sp\{\Omega_E\}$, $\Omega_E = \psi(1)$, and $\mathcal{N} \neq \{0\}$;[6] while Nelson's $\mathcal{H}_{-1}$ Euclidean field theory with Markov property has $\mathcal{H}_0 = \mathcal{H}_{\text{inv}}$.

Remark: Let $U(t)$ be a unitary representation of the time translations induced by Euclidean invariance of ext$\underline{S}$. In this notation, the self-adjoint contraction semigroup T^t on $\mathcal{H}$ given by Osterwalder and Schrader [Eq. (4.6) in Ref. 3] is just $T^t E = EU(t)$, $t \geq 0$. Its representation in $\mathcal{H}_0$ is $U^* T^t U$ which becomes Nelson's original expression $P_0 U(|t|) P_0$, P_0 orthogonal projection on $\mathcal{H}_0$, for $P_+ K_\theta P_+$ a projection.

3. THE SUBSPACE $\mathcal{H}_+$

A concrete realization for $\mathcal{H}_0$ depends upon one for P_+ which may be difficult when the Schwinger functions are sufficiently singular at equal times. Due to the symmetry of ext$\underline{S}$ it is always the case that

$$\overline{(\underline{S}(\Omega_+^0) + N_E)/N_E}^E = \mathcal{H}_+ \subset \overline{P(\underline{S}(R_+^4))}^E, \tag{3.1}$$

where $\underline{S}(\Omega_+^0) = \oplus_n S(\Omega_{n,+}^0)$, $S(\Omega_{n,+}^0)$ those Schwartz functions which vanish with their derivatives unless $0 < t_i$; and no $t_i = t_j$ for $1 \leq i < j \leq n$. The right-hand inclusion in (3.1) may be strict, however. To obtain a result on $\mathcal{H}_+$ in what follows we are concerned only with the behavior of ext$\underline{S}$ in the time variables so we shall assume the space variables have been integrated out with test functions in $S(R^3)$. Given $\underline{f} \in \underline{S}(R^4)$ we try to find approximants $\{\underline{f}_N\} \subset \underline{S}_+$ such that $\lim_{N\to\infty} \|\psi(\underline{f}) - \psi(\underline{f}_N)\|_E = 0$. Due to the symmetry of ext$\underline{S}$ we need only

$$\lim_{N\to\infty} \text{ext}\underline{S}_{n+m}(\{f_n - f_{N,n}\} \times \{g_m - g_{N,m}\}) = 0 \tag{3.2}$$

for each pair of nonnegative integers n, m. Choose $\alpha(t) = 1$, $t \geq 1$; $\exp(-1/t)(\exp(-1/t) + \exp[-1/(1-t)])^{-1}$, $0 < t < 1$; 0 for $t \leq 0$; and put $\alpha_N(t) = \alpha(Nt)$. Then when $f_{N,n}(t_1, \ldots, t_n)$

$$= \sum_{\pi \in P_n} \alpha_N(t_{\pi(1)} - t_{\pi(2)}) \cdots \alpha_N(t_{\pi(n)} - t_{\pi(n-1)}) f_n(t_1, \ldots, t_n)$$

with a like expression for $g_{N,m}$, the approximants lie in $S(\Omega_{n,+}^0)$ and $S(\Omega_{m,+}^0)$, respectively. They may be chosen to have compact support. In the case that each ext$\underline{S}_n$ is locally L_1 in time, the limit (3.2) holds by dominated convergence. A slight improvement is to be had upon making a more detailed calculation of this limit. Symmetry and translation invariance for ext$\underline{S}$ permits rewriting (3.2) as a Lebesgue integral of the form

$$\int dt_1 d\xi_2 \cdots d\xi_n d\zeta d\eta_2 \cdots d\eta_m$$
$$\times F_{n+m-1}(\xi, \zeta, \eta) P\left(-\frac{\partial}{\partial \xi_1}, \cdots, -\frac{\partial}{\partial \eta_n}\right)$$
$$\times [f_n(t_1, t_1 + \xi_2, \ldots, t_1 + \xi_2 + \cdots + \xi_n) - \alpha_N(\xi_2) \cdots \alpha_N(\xi_n)$$
$$\times \sum_\pi (\pi^{-1} f_n)(t_1, t_1 + \eta_2, \ldots, t_1 + \xi_2 + \cdots + \xi_n)]$$
$$\times [g_m(s_1, s_1 + \eta_2, \ldots, s_1 + \eta_2 + \cdots + \eta_m) - \alpha_N(\eta_2) \cdots \alpha_N(\eta_m)$$
$$\times \sum_\pi (\pi^{-1} g_m)(s_1, s_1 + \eta_2, \ldots, s_1 + \eta_2 + \cdots + \eta_m)].$$

Here F_{n+m-1} is a continuous function, P a polynomial and $\eta_i = t_i - t_{i-1}$, $2 \le i \le n$; $\zeta = s_1 - t_n$; $\eta_j = s_j - s_{j-1}$, $2 \le j \le n$. Each derivative may be considered separately so we examine a typical expression of the form

$$(-1)^{k+1} \sum_{1 \le l \le k} \binom{k}{l} \int_0^\infty dt G(t) \alpha_N^{(l)}(t) f^{(k-l)}(t),$$

where G is continuous and $f \in \mathcal{S}$. For test functions for which $f^{(p)}(0) = 0$ when $p = 0, 1, \ldots, k-1$ a calculation gives the integral as $(-1)^k G(0) f^{(k-1)}(0)$. Consequently, (3.2) holds also when the derivatives in ξ_i, η_j are at most of order one and F_{n+m-1} vanishes when any $t_i = t_j$, $1 \le i < j \le n+m$. This means a singular term $\sin[(t_i - t_j)^{-1}]/(t_i - t_j)$ is allowed in ext$\underline{S}$ even though it is not locally L_1. In summary, we find:

Proposition 3.1: $\mathcal{H}_+ = \overline{\underline{P}(\mathcal{S}(R_+^4))}^E$ when each ext$\underline{S}_n$ is locally L_1 in the time variables or in the difference time variables is the derivative of a continuous function vanishing at equal times and no derivative need be of higher order than one.

As a final comment let us make contact with previous work[1,7] and suppose each ext$\underline{S}_n$ is continuous in the time variables. Then sharp time Euclidean fields are well defined. Choose $p \in C_c^\infty(R)$ with $p \ge 0$, supp$p \subset [-1, 1]$, and $\int p(t)\,dt = 1$. Now form $\delta_N(t) = Np(Nt - 1)$ which converges weakly to the δ_0 distribution so as $N \to \infty$

$$\text{ext}\underline{S}_n(\delta_N \otimes \mathbf{f}_1 \times \cdots \times \delta_N \otimes \mathbf{f}_n) - \text{ext}\underline{S}(\delta_0 \otimes \mathbf{f}_1 \times \cdots \times \delta_0 \otimes \mathbf{f}_n),$$
$$\mathbf{f}_i \in \mathcal{S}(R^3).$$

The coset mappings ψ, v extend by continuity as well as E in (2.1). Moreover $\psi(\delta_0 \otimes \mathbf{f}_1 \times \cdots \times \delta_0 \otimes \mathbf{f}_n)$ is clearly invariant under $P_+ K_\theta P_+$ and

$$\overline{\underline{P}(\delta_0 \otimes \mathcal{S}(R^3))}^E \subset \mathcal{H}_{\text{inv}} \subset \mathcal{H}_0. \tag{3.3}$$

4. THE SUBSPACE $\mathcal{H}_0$

In this section, we will study the subspace $\mathcal{H}_0$ for ext$\underline{S}$ satisfying the regularity conditions of Proposition 3.1. First, it is instructive to review the case when sharp time relativistic fields exist and define Wightman distributions continuous in the time with cyclicity of time zero fields. This is the setting of Nelson's $\mathcal{H}_{-1}$ theory and Simon's study of the connection with the relativistic theory.[1,7] From cyclicity, Parseval's relation leads to a sequence $\{\underline{h}^{(k)}\}$ such that

(i) $\underline{h}_n^{(k)} \in \text{sp}\{\delta_0 \otimes \mathbf{f}_1^{(k)} \times \cdots \times \delta_0 \otimes \mathbf{f}_n^{(k)} \mid \mathbf{f}_i^{(k)} \in \mathcal{S}(R^3)\}$,

(ii) $\underline{W}(\underline{h}^{(i)*} \times \underline{h}^{(j)}) = \delta_{ij}$,

(iii) $\underline{W}(\underline{f}^* \times \underline{f}) = \sum_k |\underline{W}(\underline{f}^* \times \underline{h}^{(k)})|^2$,

for the Wightman state $\underline{W}$ related to $\underline{S}$. The continuity arguments at the end of Sec. 3 and the Osterwalder–Schrader relations [Eqs. (2.1), (2.2) of Ref. 4] imply $\underline{W}(\underline{h}^{(i)*} \times \underline{h}^{(j)}) = \text{ext}\underline{S}(\underline{h}^{(i)*} \times \underline{h}^{(j)})$. Then in Theorem 2.2, $U\psi(\underline{h}^{(k)}) = E\psi(\underline{h}^{(k)}) = v(\underline{h}^{(k)})$ with consequence that $\{\psi(\underline{h}^{(k)})\}$ is a complete orthonormal system in $\mathcal{H}_0$. Hence by (3.3), $\overline{\underline{P}(\delta_0 \otimes \mathcal{S}(R^3))}^E = \mathcal{H}_0 = \mathcal{H}_{\text{inv}}$ with $P_+ K_\theta P_+$ the projection onto $\mathcal{H}_0$.

To continue this discussion a little further, suppose in addition to the regularity requirements of Proposition 3.1 that ext$\underline{S}$ is a strongly positive, continuous extension of $\underline{S}$ leading to a maximal measure μ on $\mathcal{S}_R'$. Theorem 4.3 of Ref. 4 gives Proposition 3.1 in the form

$$\mathcal{H}_E = L_2(\mathcal{S}_R', \mathfrak{A}, \mu), \quad \mathcal{H}_\pm = L_2(\mathcal{S}_R', \mathfrak{A}^\pm, \mu)$$

in which $\mathfrak{A}$ ($\mathfrak{A}^\pm$) are Borel fields generated by Euclidean fields $\varphi(f)$ as in Ref. 4 with $f \in \mathcal{S}(R^4)(\mathcal{S}(R_\pm^4))$, respectively. P_+ is now conditional expectation with respect to $\mathfrak{A}^+$. For functions u measurable with respect to $\mathfrak{A}$, Euclidean and θ invariance of ext$\underline{S}$ allow us to choose μ so that $K_\theta u(T) := u(\theta T)$ and $\eta_a u(T) := u(\eta_a^{-1} T)$ for $T \in \mathcal{S}_R'$ and $a \in R^4$ are measure preserving automorphisms of $\mathfrak{A}$. Then $P = P_+ K_\theta P_+$ is a doubly Markovian operator on L_2.

In addition to $\mathfrak{A}^\pm$, three other σ fields play a role in the description of $\mathcal{H}_0$. If $\sigma\{\cdots\}$ denotes the σ field generated by functions in $\{\cdots\}$ and χ_A the indicator of the set A, these are

$$\mathfrak{A}_{\text{inv}} = \{A \in \mathfrak{A}^+ \mid P\chi_A = \chi_A\},$$
$$\mathfrak{A}_0 = \bigcap_{n \ge 1} \mathfrak{A}_n,$$
$$\mathfrak{A}_n = \sigma\{\varphi(f) \mid f \in \mathcal{S}(R^4), \quad \text{supp} f \subset \{(t, \mathbf{x}) \mid |t| < 1/n\};$$

and, following McKean and Pitt,[8] the minimal splitting σ field of $\mathfrak{A}^+$ and $\mathfrak{A}^-$,

$$\mathcal{S}_0 = \sigma\{\mathfrak{A}^+ g \mid g \text{ bounded and } \mathfrak{A}^- \text{ measurable}\}.$$

Conditional expection with respect to a sub-σ field $\mathfrak{B}$ will be written variously as $\mathfrak{B}(\cdot) = E\{\cdot \mid \mathfrak{B}\}$. One easily shows that for $A \in \mathfrak{A}^+$, $A \in \mathfrak{A}^+ \cap \mathfrak{A}^-$ if and only if $P\chi_A = \chi_{\theta A}$. Consequently,

$$\mathfrak{A}_{\text{inv}} \subset \mathfrak{A}^+ \cap \mathfrak{A}^- \subset \mathcal{S}_0.$$

General relations between the various subspaces are readily obtained.

Proposition 4.1: $\mathcal{H}_{\text{inv}} = L_2(\mathfrak{A}_{\text{inv}}) \subset L_2(\mathfrak{A}^+ \cap \mathfrak{A}^-) \subset \mathcal{H}_0 \subset L_2(\mathcal{S}_0)$.

Proof: These inclusions are applications of almost everywhere convergence theorems. The Hopf–Chacon ergodic theory[9] shows the orthogonal projection onto $\mathcal{H}_{\text{inv}}$ to be

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k(\cdot) = E\{\cdot \mid \mathfrak{A}_{\text{inv}}\}, \text{ a.e.},$$

while if u is integrable, a theorem of Burkholder and

Chow[10] implies

$$E\{u \mid \mathfrak{A}^+ \cap \mathfrak{A}^-\} = \lim_{N \to \infty} (PK_\theta)^N u.$$

The projection onto $L_2(\mathfrak{A}^+ \cap \mathfrak{A}^-)$ therefore lies in the closure of the range of P. The remaining inclusions are just a matter of using the definitions.

Remark: The positivity condition of Osterwalder and Schrader now appears as $\mathcal{H}_0 \subset L_2(\mathcal{S}_0)$ (see Ref. 11, pages 104–105).

Corollary: $\mathcal{H}_0 = L_2(\mathcal{S}_0)$ if and only if P is a projection.

Proof: When P is a projection, Proposition 2.4 states $\mathcal{H}_0 = \mathcal{H}_{\text{inv}}$. Notice that for g bounded and $\mathfrak{A}^-$ measurable, $P_+ g = P(K_\theta g)$ so $\mathcal{S}_0 = \sigma\{Ph \mid h$ bounded and $\mathfrak{A}^+$ measurable$\}$. Then P a projection implies $\mathcal{S}_0 = \mathfrak{A}_{\text{inv}}$. Conversely, when $\mathcal{H}_0 = L_2(\mathcal{S}_0)$ we find by the splitting property $Pf = P_+ K_\theta f = \mathcal{S}_0 f = f$ for $f \in \mathcal{H}_0$. Thus, $\mathcal{H}_0 \subset \mathcal{H}_{\text{inv}}$ and P is a projection.

The relevance for these remarks to Nelson's Markov property can now be readily ascertained. First, for completeness, let us recall several notions from[1,8,11] appropriate to strongly positive extensions which are not necessarily locally L_1 in the time variables. Following Pitt,[8] introduce σ fields localized at sharp time but from the future and the past, respectively,

$$\mathfrak{A}_0^{\pm} = \bigcap_{n \geq 1} \mathfrak{A}_n^{\pm}, \quad \mathfrak{A}_n^{\pm} = \sigma\{\varphi(f) \mid \operatorname{supp} f \subset \{(t, \mathbf{x}) \mid 0 < \pm t < 1/n\}\}.$$

It is then true that

$$\mathfrak{A}_0^+ = \mathfrak{A}_0 = \mathfrak{A}_0^-.$$

For $A \in \mathfrak{A}$, as ext$\underline{S}$ solves a moment problem for a maximal measure (Theorem 4.3 of Ref. 4), for $\epsilon > 0$ there exists a polynomial $P(\varphi(f_1) \cdots \varphi(f_n))$ such that $\|\chi_A - P\|_2 < \epsilon$. Hence,

$$\lim_{t \to 0} \| \chi_A - \eta_t \chi_A \|_2 \leq 2\epsilon + \lim_{t \to 0} \|P - \eta_t P\|_2 = 2\epsilon$$

as η_t is measure preserving and when $t \to 0$, ext$\underline{S}(P^* \times \eta_t P) \to$ ext$\underline{S}(P^* \times P)$. If $A \in \mathfrak{A}_0$, then for each n we may find a polynomial $P_n \in L_2(\mathfrak{A}_n)$ with $\|\chi_A - P_n\|_2 < 1/2^n$. This means

$$\|\eta_{(1/n)} P_n - \chi_A\|_2 \leq \|\chi_A - \eta_{(1/n)} \chi_A\|_2 + 1/2^n,$$

the right-hand side tending to zero with $n \to \infty$. The result follows since $\eta_{(1/n)} P_n \in L_2(\mathfrak{A}_{2n}^+)$. The state ext$\underline{S}$ will be said to have a half-space Markov property in the event $\mathfrak{A}_0 = \mathcal{S}_0$ and to satisfy the reflection property when $K_\theta u = u$ for $u \in L_2(\mathfrak{A}_0)$, see Ref. 11. Suppose that ext$\underline{S}$ also has the regularity conditions in Proposition 3.1, then the Corollary to Proposition 4.1 and our discussion in Sec. 2 implies, easily, the equivalence of the following two statements:

(a) P is a projection and $\mathfrak{A}_{\text{inv}} \subset \mathfrak{A}_0$;

(b) ext$\underline{S}$ has both the reflection and half-space Markov properties.

For these cases the physical Hilbert space embeds in $L_2(\mathfrak{A})$ as $L_2(\mathfrak{A}_0)$, which was Nelson's original choice in Ref. 1. The requirement $\mathfrak{A}_{\text{inv}} \subset \mathfrak{A}_0$ is the analog of cyclicity for time zero fields mentioned at the beginning of this section. It allows $\mathfrak{A}_0$ to be large enough to split the past and future σ fields. Minimality as required by Pitt[8] is a consequence of P being a projection.

ACKNOWLEDGMENTS

It is a pleasure to record beneficial discussions with Victor Goodman and Steven Slinker on the subject of this paper and Euclidean field theory generally. I have also profited from a conversation with Loren Pitt on his work in Ref. 8.

[1]E. Nelson, J. Functional Analysis 12, 97 (1973).
[2]G. C. Hegerfeldt, Comm. Math. Phys. 35, 155 (1974).
[3]K. Osterwalder and R. Schrader, Comm. Math. Phys. 31, 83 (1973).
[4]J. L. Challifour and S. P. Slinker, Comm. Math. Phys. 43, 41 (1975).
[5]J. L. Challifour and S. P. Slinker (in preparation).
[6]V. Goodman (private communication).
[7]B. Simon, Helv. Phys. Acta 46, 686 (1973).
[8]L. D. Pitt, Arch. Ratl. Mech. Anal. 43, 367 (1971).
[9]A. M. Garsia, Topics in Almost Everywhere Convergence (Markham, Chicago, 1970).
[10]D. L. Burkholder and Y. S. Chow, Proc. Am. Math. Soc. 12, 490 (1961).
[11]E. Nelson, *Probability Theory and Euclidean Field Theory* in *Constructive Quantum Field Theory*, edited by G. Velo and A. Wightman (Springer-Verlag, Heidelberg, 1973), Vol. 25.

Space–time symmetries and linearization stability of the Einstein equations. II*

Vincent Moncrief†

Department of Physics, University of Utah, Salt Lake City, Utah 84112
(Received 4 June 1976)

In a previous paper we began a study of the Fischer–Marsden conditions for the linearization stability of vacuum space–times with compact, Cauchy hypersurfaces. We showed that a space–time of this class is linearization stable if and only if it admits no global Killing vector fields. In this paper we derive the general nonlinear constraints upon the perturbations which are necessary, whenever Killing symmetries occur, to exclude spurious perturbation solutions. We establish the hypersurface independence of these constraints by relating them to the conserved integrals of the perturbation equations associated with the Killing symmetries of the background. As a corollary of this result, we also establish the gauge invariance of the nonlinear constraints. We briefly discuss the noncompact case and mention a possible application of our results to the study of the Hawking press of quantum mechanical particle production by black holes.

1. INTRODUCTION

In a previous paper[1] (referred to here as paper I) we began a study of the Fischer and Marsden[2,3] criterion for the linearization stability of solutions to the vacuum Einstein equations. A solution is said to be linearization stable if and only if each of its linear perturbations is tangent to a smooth curve of exact solutions. In the unstable case there are always some spurious solutions to the associated perturbation equations which are not tangent to any curves of exact solutions. In paper I we concentrated on distinguishing the (linearization) stable space—times from the unstable ones. In this paper we consider perturbing an unstable space—time and attempt to distinguish the acceptable perturbations from the spurious ones.

The vacuum space—times considered in the Fischer—Marsden linearization stability theorem are those with compact, boundaryless Cauchy hypersurfaces. Fischer and Marsden consider the space of Cauchy data that can be defined over a compact, boundaryless three-manifold M and study the constraint subset of this (infinite dimensional) space. They apply the implicit function theorem (using Sobolev manifolds of Cauchy data) to show that the Einstein constraint equations are linearization stable at a given exact solution provided that the constraint map has surjective derivative at the given point. Using elliptic theory they show that the derivative of the constraint map (which defines the constraints of the corresponding linearized Einstein equations) is surjective if and only if an associated adjoint operator is injective. Thus, linearization stability of the constraint equations obtains at a given exact solution provided the associated Fischer—Marsden adjoint map has trivial kernel. The remainder of the Fisher—Marsden theorem consists of showing that linearization stability of a solution to the initial value equations extends to linearization stability of the full set of Einstein equations on a Cauchy development of the initial data. The corresponding result for C^∞ data is obtained by a regularity argument from that using Sobolev spaces.

In paper I we showed that if a vacuum space—time (with compact Cauchy surfaces) admits a Killing vector field ${}^{(4)}X$, then ${}^{(4)}X$ induces, upon any Cauchy hypersurface, a nontrivial element of the kernel of the Fischer—Marsden adjoint map associated with that hypersurface. Linearly independent Killing fields induce linearly independent elements of the kernel. We also showed that, when the adjoint map for some Cauchy surface has a kernel of dimension $k>0$, there is a Cauchy development of this surface admitting k linearly independent Killing vector fields. Thus, for vacuum space—times with compact Cauchy surfaces, the occurrence of a Killing vector field precludes linearization stability and so guarantees the occurrence of spurious solutions to the corresponding linearized Einstein equations. We wish to here characterize these spurious solutions of the perturbation equations.[††]

We shall first derive general formulas for the additional, nonlinear constraints upon the perturbations which are necessary, whenever Killing symmetries occur, to exclude spurious perturbation solutions. This derivation follows the same pattern set by Brill,[4] Brill and Deser,[5] and by Fischer and Marsden[3] for the special cases discussed below. Next we shall apply an argument due to Taub[6,7] to show that for each independent Killing vector field of the space—time there is a conservation law for the associated gravitational perturbations. The main result of this paper will be to show that the nonlinear constraints upon the perturbations, necessary to exclude spurious solutions, are equivalent to the requirement that each conserved quantity must be constrained to vanish.

In showing that the additional nonlinear constraints upon the perturbations are conserved quantities, we establish the hypersurface independence of these additional constraints. This result therefore will exclude the possibility that perturbed initial data which seems allowed on one hypersurface could propagate to induce a spurious perturbation on a different hypersurface. An important corollary of our result is that the nonlinear constraints are gauge invariant and so are identically satisfied for any pure gauge perturbation. The nonlinear constraints restrict only those nontrivial perturbations towards space—time distinct from the given one.

In the original study of linearization instabilities in general relativity Brill,[4] and later Brill and Deser,[5] treated the special case of a flat space—time with compact, flat Cauchy hypersurfaces. They derived several nonlinear restrictions upon the perturbations which were necessary to exclude spurious perturbations of the given, flat space—time. Fischer and Marsden then proved the general result described above relating linearization stability to an injective adjoint map.[2] They then derived two necessary conditions for an injective map. For a Cauchy hypersurface having first fundamental form g and second fundamental form k they require:

(i) if $k=0$, then g is not flat, and

(ii) there is no nonzero vector field X such that $\mathcal{L}_X g = \mathcal{L}_X k = 0$, where $\mathcal{L}_X$ signifies the Lie derivative with respect to X. When either of these necessary conditions fails to hold one must impose additional nonlinear constraints (derived by Marsden and Fischer)[3] upon the perturbations. The nonlinear constraints derived by Brill and Deser and by Fischer and Marsden are special cases of the general constraint derived in Sec. 2.

To obtain this generalized form of the nonlinear constraints we use the main result of paper I which states that the adjoint map for a compact Cauchy hypersurface has nontrivial kernel if and only if the initial data allows a (vacuum) Cauchy development with one or more independent Killing vector fields. The dimension of the kernel is equal to the number of independent Killing fields which occur. Section 3 applies the argument of Taub[6,7] to derive the conservation laws for the gravitational perturbations when Killing symmetries are present. Section 4 proves our main result that the nonlinear constraints of Sec. 2 are equivalent to the requirement that the conserved quantities of Sec. 3 must vanish in order to exclude spurious perturbations. Gauge invariance is discussed in Sec. 5.

In Sec. 6 we briefly discuss how our conclusions must be modified when noncompact, asymptotically flat, Cauchy hypersurfaces are considered. The occurrence of certain boundary integrals, absent in the compact case, alter the conclusions regarding linearization stability, even when Killing vector fields are present. The conserved quantities are no longer necessarily constrained to vanish but their values must be matched by boundary integrals which involve the second order perturbations. Depending upon the type of Killing field considered, the boundary integrals can represent second order corrections to the energy, momentum or angular momentum of the perturbed spacetime. Some possible applications of this result to the perturbations of black holes are briefly discussed. In particular, we mention its possible relevance to the study of the Hawking process of quantum mechanical particle production generated by gravitational collapse.[8] Our result suggests a method of computing the reaction effects of this particle production up to the second order of approximation.

Throughout this paper we use the same notation as that of paper I. Except for some sign conventions and other minor changes our notation is that of Fischer and Marsden.[3]

As in paper I we let M designate a fixed compact, oriented, C^∞ three-manifold without boundary and define the following spaces of C^∞ tensor fields over M:

$\mathcal{S}_2(\mathcal{S}_2^*)$ = space of symmetric, covariant, second rank tensor fields (tensor densities) over M,

$\mathcal{S}^2(\mathcal{S}_*^2)$ = space of symmetric, contravariant, second rank tensor fields (tensor densities) over M,

$\mathcal{M}$ = space of Riemannian metrics of M,

$C^\infty(C_*^\infty)$ = space of scalar functions (scalar densities) over M,

$\chi_1(\chi_1^*)$ = space of covariant vector fields (densities) over M.

$\chi^1(\chi_*^1)$ = space of contravariant vector fields (densities) over M. The gravitational phase space is $\mathcal{M}\times\mathcal{S}_*^2$ and its constraint subset $\mathcal{C}$ is defined by

$$\mathcal{C} = \{(g,\pi)\in\mathcal{M}\times\mathcal{S}_*^2 \mid \Phi(g,\pi)=0\}, \tag{1.1}$$

where

$$\Phi : \mathcal{M}\times\mathcal{S}_*^2 \to C_*^\infty\times\chi_*^1, \tag{1.2}$$
$$(g,\pi)\mapsto(\mathcal{H}(g,\pi),\,\delta(g,\pi))$$

where

$$\mathcal{H}(g,\pi) = (\det g)^{-1/2}[\pi^{ij}\pi_{ij} - \tfrac{1}{2}(\mathrm{tr}\pi)^2] - (\det g)^{1/2}R, \tag{1.3}$$

$$\delta^i(g,\pi) = 2\pi^{ij}{}_{|j}. \tag{1.4}$$

As before, a vertical bar signifies covariant differentiation with respect to g, $(\det g)$ is the determinant of g_{ij}, R is the curvature scalar of g, and $\mathrm{tr}\pi \equiv g_{ij}\pi^{ij}$ is the trace of π.

2. KILLING SYMMETRIES AND NONLINEAR CONSTRAINTS

Let (g_0,π_0) be an element of the constraint subset $\mathcal{C}$ and let $(g(\lambda),\pi(\lambda))$, with $\lambda\in(-\alpha,\alpha)$, be a smooth curve in $\mathcal{M}\times\mathcal{S}_*^2$ with $(g(0),\pi(0))=(g_0,\pi_0)$. If $\Phi(g(\lambda),\pi(\lambda))=0$, so that the entire curve lies in $\mathcal{C}$, we have

$$\frac{d\Phi}{d\lambda}(g(\lambda),\pi(\lambda))\Big|_{\lambda=0} = 0, \tag{2.1}$$

and

$$\frac{d^2\Phi}{d\lambda^2}(g(\lambda),\pi(\lambda))\Big|_{\lambda=0} = 0, \tag{2.2}$$

etc. Expressed in terms of the tangent vector $(h,p)\in T_{(g_0,\pi_0)}\mathcal{M}\times\mathcal{S}_*^2 \approx \mathcal{S}_2\times\mathcal{S}_*^2$ defined by

$$(h,p) = \left(\frac{dg(\lambda)}{d\lambda},\ \frac{d\pi(\lambda)}{d\lambda}\right)\Big|_{\lambda=0}, \tag{2.3}$$

Eq. (2.1) becomes

$$D\Phi(g_0,\pi_0)\cdot(h,p) = 0, \tag{2.4}$$

where $D\Phi(g_0,\pi_0)$ is the linear operator given explicitly by Eq. (3.2) of paper I. In terms of (h,p) and the second derivatives $(h',p')\in\mathcal{S}_2\times\mathcal{S}_*^2$ defined by

$$(h',p') = \left(\frac{d^2g(\lambda)}{d\lambda^2},\ \frac{d^2\pi(\lambda)}{d\lambda^2}\right)\Big|_{\lambda=0} = 0, \tag{2.5}$$

Eq. (2.2) becomes

$$D^2\Phi(g_0,\pi_0)\cdot((h,p),(h,p)) + D\Phi(g_0,\pi_0)\cdot(h',p')=0, \tag{2.6}$$

where $D^2\Phi(g_0,\pi_0)\cdot(\ ,\ )$ is a bilinear, symmetric map from $(\mathcal{S}_2\times\mathcal{S}^2_*)\times(\mathcal{S}_2\times\mathcal{S}^2_*)$ to $\mathcal{C}^\infty_*\times\chi^1_*$. The explicit formula $D^2\Phi(g_0,\pi_0)\cdot(\ ,\ )$ is given in the Appendix below.

If $(C,X)\in\mathcal{C}^\infty\times\chi_1$ and $(h',p')\in\mathcal{S}_2\times\mathcal{S}^2_*$, then

$$\int_M d^3x\langle(C,X);D\Phi(g_0,\pi_0)\cdot(h',p')\rangle$$
$$\equiv\int_M d^3x\{CD\mathcal{H}(g_0,\pi_0)\cdot(h',p') + X_iD\delta^i(g_0,\pi_0)\cdot(h',p')\}$$
$$=\int_M d^3x\langle\langle D\Phi(g_0,\pi_0)^*(C,X);(h',p')\rangle\rangle, \tag{2.7}$$

where $D\Phi(g_0,\pi_0)^*$ is the Fischer—Marsden adjoint map given explictly in our notation by Eq. (3.3) of paper I. Here $\langle\langle\ ;\ \rangle\rangle$ is defined by

$$\langle\langle(\omega,k);(h',p')\rangle\rangle = \omega^{ij}h'_{ij}+k_{ij}p'^{ij}\in\mathcal{C}^\infty_* \tag{2.8}$$

for any $(\omega,k)\in\mathcal{S}^2_*\times\mathcal{S}_2$. Combining Eq. (2.6) with Eq. (2.7) we see that

$$\int_M d^3x\langle(C,X);D^2\Phi(g_0,\pi_0)\cdot((h,p),(h,p))\rangle$$
$$\equiv\int_M d^3x\{CD^2\mathcal{H}(g_0,\pi_0)\cdot((h,p),(h,p)) + X_iD^2\delta^i(g_0,\pi_0)\cdot((h,p),(h,p))\} \tag{2.9}$$
$$=-\int_M d^3x\langle\langle D\Phi(g_0,\pi_0)^*(C,X);(h',p')\rangle\rangle$$

for any $(C,X)\in\mathcal{C}^\infty\times\chi_1$ provided (h,p) and (h',p') are the first and second derivatives [at (g_0,π_0)] of a smooth curve of solutions of the constraints. It follows that if $D\Phi(g_0,\pi_0)^*$ has nontrivial kernel and (C,X) is a nonvanishing element of this kernel, then the tangent vector (h,p) to any smooth curve of solutions of the constraint equations must obey

$$\int_M d^3x\langle(C,X);D^2\Phi(g_0,\pi_0)\cdot((h,p),(h,p))\rangle=0. \tag{2.10}$$

Each linearly independent element in the kernel of $D\Phi(g_0,\pi_0)^*$ gives rise to one such equation. Clearly no vector $(h,p)\in T_{(g_0,\pi_0)}\mathcal{M}\times\mathcal{S}^2_*$ can be tangent to a curve lying entirely in $\mathcal{C}$ unless (h,p) satisfies Eq. (2.10) for each independent $(C,X)\in\ker D\Phi(g_0,\pi_0)^*$.

Brill and Deser[4,5] and Marsden and Fischer[3] derived nonlinear constraints of this type for several cases in which they could present explicitly a nontrivial element in $\ker D\Phi(g_0,\pi_0)^*$. The Fischer—Marsden results (which include those of Brill and Deser) may be summarized as follows:

(i) if $\pi_0=0$ and g_0 is flat, put $(C,X)=(1,0)$, and

(ii) if these exists a nonzero vector field $Y\in\chi^1$ such that $\mathcal{L}_Yg_0=\mathcal{L}_Y\pi_0=0$, where $\mathcal{L}_Y$ is the Lie derivative with respect to Y, put $(C,X^i)=(C,g_0^{ij}X_j)=(0,Y^i)$.

For these cases, which are not exhaustive, Brill and Deser and Marsden and Fischer derived the nonlinear constraints explicitly. We can extend their results by recalling the main conclusion of paper I.

In paper I we showed that $D\Phi(g_0,\pi_0)^*$ has a nontrivial kernel of dimension k if and only if the initial data (g_0,π_0) admits a vacuum Cauchy development $({}^{(4)}g, M\times(-\epsilon,\epsilon))$ with k linearly independent Killing vector fields. Each Killing field ${}^{(4)}X$ induces a nonzero element of $\ker D\Phi(g_0,\pi_0)^*$ onto the hypersurface (M,g_0,π_0) as its normal and tangential projections. More specifically, if n^α is the unit (future directed) normal field of the hypersurface $\Sigma=(M,g_0,\pi_0)$ expressed in coordinates for which Σ occurs as an $x^0=t=$constant surface and if we express ${}^{(4)}X$ on Σ as

$${}^{(4)}X=-Cn^\alpha\frac{\partial}{\partial x^\alpha}+g_0^{ij}X_j\frac{\partial}{\partial x^i}, \tag{2.11}$$

then (C,X) satisfies $D\Phi(g_0,\pi_0)^*(C,X)=0$. Also, every element of $\ker D\Phi(g_0,\pi_0)^*$ extends to a Killing field on $({}^{(4)}g, M\times(-\epsilon,\epsilon))$. An immediate consequence of this result (Theorem 6.1 of paper I) and of Eq. (2.10) is:

Theorem 2.1: If a vacuum space—time with compact Cauchy slices admits a Killing vector field ${}^{(4)}X$ which has normal and tangential projections (C,X) at the hypersurface (M,g_0,π_0), then a necessary restriction upon the perturbations (h,p) of the initial data (g_0,π_0) is given by

$$\int_M d^3x\langle(C,X);D^2\Phi(g_0,\pi_0)\cdot((h,p),(h,p))\rangle=0.$$

A perturbation failing to satisfy any of these nonlinear constraints (one for each independent Killing field) cannot be tangent to a curve of exact solutions of the constraint equations.

It is not known whether the nonlinear constraints of Theorem (2.1) are sufficient to exclude all spurious first order perturbations. At present, however, it seems reasonable to conjecture that they are in fact sufficient.

The nonlinear constraints discussed above are, as we have shown, always associated with Killing vector fields of the background space—time. In Sec. 3, following an argument due to Taub,[6,7] we shall show how the presence of Killing fields implies conservation laws for the gravitational perturbation equations. In Sec. 4 we shall prove that the conserved quantity associated with each independent Killing field must be constrained to vanish as a consequence of Theorem (2.1). The importance of relating the nonlinear constraints to conserved quantities is that we thereby exclude the possibility of a perturbation satisfying the nonlinear constraints for one hypersurface but propagating to fail the corresponding constraints for another hypersurface.

3. CONSERVATION LAWS FOR GRAVITATIONAL PERTURBATIONS

In this section we discuss a method, due to Taub,[6,7] for constructing nontrivial conserved quantities in gravitational perturbation theory whenever the background space—time admits a Killing vector field. The idea is first to construct a symmetric tensor field $T^{\alpha\beta}({}^{(4)}g, {}^{(4)}h)$ from the background metric ${}^{(4)}g_{\alpha\beta}$ and the metric perturbation ${}^{(4)}h_{\alpha\beta}$, which has vanishing divergence with respect to ${}^{(4)}g$. As we shall show, this is always possible when ${}^{(4)}g$ obeys the Einstein equations and ${}^{(4)}h$ obeys the associated linearized Einstein equations.

Given such a tensor field one obtains a conservation law for each Killing vector field of the background space—time. If ${}^{(4)}X$ obeys Killing's equations, ${}^{(4)}X_{\alpha;\beta}$

$+{}^{(4)}X_{\beta;\alpha}=0$, and $T^{\alpha\beta}$ is symmetric and obeys $T^{\alpha\beta}{}_{;\beta}=0$ (where a semicolon signifies covariant differentiation with respect to ${}^{(4)}g$), then we have

$$[{}^{(4)}X_\alpha T^{\alpha\beta}]_{;\beta}=\tfrac{1}{2}({}^{(4)}X_{\alpha;\beta}+{}^{(4)}X_{\beta;\alpha})T^{\alpha\beta}+{}^{(4)}X_\alpha T^{\alpha\beta}_{;\beta}=0. \tag{3.1}$$

For the class of space–times considered here (with compact Cauchy surfaces) we then obtain, by a standard argument, the hypersurface independence (i.e., conservation) of

$$E_{{}^{(4)}X}({}^{(4)}h,\Sigma)\equiv\int_\Sigma {}^{(4)}X_\alpha T^{\alpha\beta}n_\beta\,d\sigma, \tag{3.2}$$

where Σ is a Cauchy hypersurface with unit (future directed) normal field n^α and induced volume element $d\sigma$. In coordinates x^α for which Σ is an $x^0=t=$constant hypersurface, the integral may be written

$$E_{{}^{(4)}X}({}^{(4)}h,\Sigma)=-\int_\Sigma[{}^{(4)}X_\alpha T^{\alpha 0}(-\det{}^{(4)}g)^{1/2}]\,d^3x, \tag{3.3}$$

where $(\det{}^{(4)}g)$ signifies the determinant of ${}^{(4)}g_{\alpha\beta}$ and $d^3x=dx^1\,dx^2\,dx^3$.

Taub derived an explicit expression for $T^{\alpha\beta}({}^{(4)}g,{}^{(4)}h)$ for the case in which ${}^{(4)}g$ is the Minkowski metric.[6] He has argued that one may always obtain a divergence-free $T^{\alpha\beta}({}^{(4)}g,{}^{(4)}h)$ by varying ${}^{(4)}g$ in a suitable variational principle for the perturbation equations.[7] Instead of using a variational principle, we shall here obtain a suitable divergence-free $T^{\alpha\beta}({}^{(4)}g,{}^{(4)}h)$ by studying perturbations of the contracted Bianchi identities. This approach is an extension of that used by Taub in Ref. 6, and shows how $T^{\alpha\beta}({}^{(4)}g,{}^{(4)}h)$ arises naturally in second order gravitational perturbation theory.

It will be useful to have the following lemma.

Lemma 3.1: Let ${}^{(4)}g$ be a Lorentzian metric and let ${}^{(4)}h$ be a covariant, symmetric, second rank tensor field on some four-dimensional manifold ${}^{(4)}V$. Then, for any point $p\in{}^{(4)}V$, there is a neighborhood $N_p\subset{}^{(4)}V$ of p and a constant $\alpha>0$ such that ${}^{(4)}g+\lambda{}^{(4)}h$ is Lorentzian on N_p for all $\lambda\in(-\alpha,\alpha)$.

Proof: Let M_p be a neighborhood of p which admits, relative to the metric ${}^{(4)}g$, a field of orthonormal frames ${}^{(4)}X_{(\mu)}$ $(\mu=0,1,2,3)$. Thus, the vector fields ${}^{(4)}X_{(\mu)}$ obey ${}^{(4)}g_{\alpha\beta}\,{}^{(4)}X^\alpha_{(\mu)}\,{}^{(4)}X^\beta_{(\nu)}=\eta_{(\mu)(\nu)}$, where $\eta_{(\mu)(\nu)}=\mathrm{diag}(-1,1,1,1)$ is the Minkowski metric. Let $K_p\subset M_p$ be a compact neighborhood of p (e.g., the inverse image, in a suitable coordinate chart, of a closed ball in R^4 which contains the image of p as an interior point). Given ${}^{(4)}h$ define the (continuous) functions

$$\gamma_{(\mu)(\nu)}={}^{(4)}X^\alpha_{(\mu)}\;{}^{(4)}X^\beta_{(\nu)}\;{}^{(4)}h_{\alpha\beta}$$

on M_p and set

$$\Gamma=\max_{(\mu,\nu)}\sup_{K_p}|\gamma_{(\mu)(\nu)}|.$$

Γ will be finite since the continuous functions $\gamma_{(\mu)(\nu)}$ are necessarily bounded on the compact set K_p. Now let N_p be some neighborhood of p contained in K_p and, if $\Gamma>0$, put $\alpha=1/4\Gamma$. If $\Gamma=0$ put $\alpha=1$, since the result is then trivial. One may now verify that ${}^{(4)}g_{\alpha\beta}(\lambda)\equiv{}^{(4)}g_{\alpha\beta}+\lambda\,{}^{(4)}h_{\alpha\beta}$ is Lorentzian on N_p for all $\lambda\in(-\alpha,\alpha)$.

To verify this explictly, consider vector fields of the form $V=V^{(i)}X_{(i)}$ $(i=1,2,3)$ defined on N_p. It is easy to show, as a consequence of the above inequalities, that ${}^{(4)}g_{\alpha\beta}(\lambda)V^\alpha V^\beta\geq 0$, with equality holding at some point only if V vanishes there. Thus the three-dimensional subspace spanned, at any point of N_p, by the $X_{(i)}$ remains spacelike for each of the metrics ${}^{(4)}g(\lambda)$ with $|\lambda|<\alpha$. Finally, verify that ${}^{(4)}X_{(0)}$ remains timelike on N_p for each of the metrics ${}^{(4)}g(\lambda)$ to complete the argument. ∎

Remark: Geroch[9] has shown that space–times admitting Cauchy hypersurfaces always admit global fields of orthonormal frames. Even though a global frame field exists, however, it is not in general true that ${}^{(4)}g+\lambda{}^{(4)}h$ is globally Lorentzian for any $|\lambda|>0$. The functions $\gamma_{(\mu)(\nu)}$ might diverge as one approached an "edge" of ${}^{(4)}V$ and thus prevent a single choice for λ working at every point of ${}^{(4)}V$.

Now, let $\mathrm{Ein}({}^{(4)}g)$ designate the Einstein tensor of a Lorentz metric ${}^{(4)}g$. $\mathrm{Ein}({}^{(4)}g)$ is defined, as usual, by

$$[\mathrm{Ein}({}^{(4)}g)]_{\alpha\beta}=R_{\alpha\beta}({}^{(4)}g)-\tfrac{1}{2}{}^{(4)}g_{\alpha\beta}\;{}^{(4)}g^{\mu\nu}R_{\mu\nu}({}^{(4)}g), \tag{3.4}$$

where

$$R_{\alpha\beta}({}^{(4)}g)\equiv[\mathrm{Ric}({}^{(4)}g)]_{\alpha\beta} \tag{3.5}$$

are the components of the Ricci tensor, $\mathrm{Ric}({}^{(4)}g)$, of ${}^{(4)}g$. Let ${}^{(4)}g(\lambda)$, with $\lambda\in(-\alpha,\alpha)$ for some $\alpha>0$, be a smooth curve of Lorentzian metrics on some fixed manifold ${}^{(4)}V$ and write ${}^{(4)}g_0$ for ${}^{(4)}g(0)$. Now differentiate $\mathrm{Ein}({}^{(4)}g(\lambda))$ with respect to λ, put $\lambda=0$, and express the result as

$$\frac{\partial\mathrm{Ein}}{\partial\lambda}({}^{(4)}g(\lambda))\Big|_{\lambda=0}=D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}h, \tag{3.6}$$

where

$${}^{(4)}h_{\alpha\beta}\equiv\frac{\partial}{\partial\lambda}\,{}^{(4)}g_{\alpha\beta}(\lambda)\Big|_{\lambda=0}. \tag{3.7}$$

The explicit expression for $D\,\mathrm{Ein}({}^{(4)}g)\cdot{}^{(4)}h$ may be readily derived from the corresponding, standard result for the Ricci tensor,

$$[D\,\mathrm{Ric}({}^{(4)}g)\cdot{}^{(4)}h]_{\alpha\beta}=\tfrac{1}{2}[{}^{(4)}h_{\alpha\mu;\beta}{}^{;\mu}+{}^{(4)}h_{\beta\mu;\alpha}{}^{;\mu}-{}^{(4)}h_{\alpha\beta;\mu}{}^{;\mu}-({}^{(4)}h_\mu{}^\mu)_{;\alpha\beta}]. \tag{3.8}$$

If ${}^{(4)}g_0$ obeys $\mathrm{Ein}({}^{(4)}g_0)=0$, then

$$D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}h=0 \tag{3.9}$$

are the (first order) gravitational perturbation equations.

If we differentiate $\mathrm{Ein}({}^{(4)}g(\lambda))$ twice with respect to λ and set $\lambda=0$ we may express the result as

$$\frac{\partial^2\mathrm{Ein}}{\partial\lambda^2}({}^{(4)}g(\lambda))\Big|_{\lambda=0}=D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}\tilde h+D^2\,\mathrm{Ein}({}^{(4)}g_0)\cdot({}^{(4)}h,{}^{(4)}h), \tag{3.10}$$

where

$${}^{(4)}\tilde h_{\alpha\beta}\equiv\frac{\partial^2\,{}^{(4)}g_{\alpha\beta}(\lambda)}{\partial\lambda^2}\Big|_{\lambda=0}. \tag{3.11}$$

The explicit expression for $D^2\,\mathrm{Ein}({}^{(4)}g_0)\cdot({}^{(4)}h, {}^{(4)}h)$ may be derived from Eqs. (3.4), (3.8) and from the corresponding, standard result for the Ricci tensor,

$$[D^2\,\mathrm{Ric}({}^{(4)}g)\cdot({}^{(4)}h, {}^{(4)}h)]_{\alpha\beta}$$
$$= + {}^{(4)}h^{\mu\lambda}[{}^{(4)}h_{\mu\lambda;\alpha\beta} + {}^{(4)}h_{\alpha\beta;\mu\lambda} - {}^{(4)}h_{\alpha\lambda;\beta\mu} - {}^{(4)}h_{\beta\lambda;\alpha\mu}]$$
$$+ (\tfrac{1}{2}{}^{(4)}h^{\mu}{}_{\mu}{}^{;\lambda} - {}^{(4)}h^{\mu\lambda}{}_{;\mu})[{}^{(4)}h_{\alpha\lambda;\beta} + {}^{(4)}h_{\beta\lambda;\alpha} - {}^{(4)}h_{\alpha\beta;\lambda}]$$
$$+ {}^{(4)}h_{\beta}{}^{\mu;\lambda}[{}^{(4)}h_{\mu\alpha;\lambda} - {}^{(4)}h_{\alpha\lambda;\mu}] + \tfrac{1}{2}{}^{(4)}h_{\mu\lambda;\alpha}\,{}^{(4)}h^{\mu\lambda}{}_{;\beta}. \tag{3.12}$$

If ${}^{(4)}g_0$ obeys the Einstein equations and ${}^{(4)}h$ the linearized Einstein equations (3.9), then

$$D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}\tilde{h} = -D^2\,\mathrm{Ein}({}^{(4)}g_0)\cdot({}^{(4)}h, {}^{(4)}h) \tag{3.13}$$

are the second order perturbation equations.

By perturbing the contracted Bianchi identities we shall show that $D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}h$ has identically vanishing divergence (with respect to ${}^{(4)}g_0$) whenever $\mathrm{Ein}({}^{(4)}g_0)=0$. We shall also show that $D^2\,\mathrm{Ein}({}^{(4)}g_0)\cdot({}^{(4)}h, {}^{(4)}h)$ has vanishing divergence whenever, in addition to $\mathrm{Ein}({}^{(4)}g_0)=0$, we have $D\,\mathrm{Ein}({}^{(4)}g_0)\circ{}^{(4)}h=0$. Given these results we shall identify $D^2\,\mathrm{Ein}({}^{(4)}g_0)\cdot({}^{(4)}h, {}^{(4)}h)$ with the divergence-free tensor $T^{\alpha\beta}({}^{(4)}g_0, {}^{(4)}h)$ that we are seeking. The first of these results is quite well known in linear perturbation theory. The second is clearly necessary for the consistency of the second order perturbation equations (3.13) since, from the first result, the left-hand side of (3.13) has vanishing divergence.

We shall write the contracted Bianchi identities as

$$\nabla_{{}^{(4)}g}\cdot\mathrm{Ein}({}^{(4)}g)\equiv 0, \tag{3.14}$$

where

$$[\nabla_{{}^{(4)}g}\cdot\mathrm{Ein}({}^{(4)}g)]^{\alpha}\equiv[\mathrm{Ein}({}^{(4)}g)]^{\alpha\beta}{}_{;\beta}. \tag{3.15}$$

We may evaluate the identity (3.14) on any smooth curve of Lorentz metrics ${}^{(4)}g(\lambda)$ and differentiate once with respect to λ and put $\lambda=0$. If, as we shall assume, ${}^{(4)}g_0={}^{(4)}g(0)$ obeys $\mathrm{Ein}({}^{(4)}g_0)=0$, the result simplifies to

$$[\nabla_{{}^{(4)}g_0}\cdot(D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}h)]^{\alpha}\equiv[D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}h]^{\alpha\beta}{}_{;\beta}=0, \tag{3.16}$$

where the semicolon signifies covariant differentiation with respect to ${}^{(4)}g_0$.

We have assumed that ${}^{(4)}g(\lambda)$ is a curve of Lorentz metrics but we really only need to assume this locally in order to obtain (3.16). Let ${}^{(4)}g_0$, with $\mathrm{Ein}({}^{(4)}g_0)=0$, and ${}^{(4)}h$ be given. At any point p we may, according to Lemma 3.1, find a neighborhood N_p of p and an $\alpha>0$ such that the curve ${}^{(4)}g_0+\lambda{}^{(4)}h$ is Lorentzian on N_p for all $\lambda\in(-\alpha,\alpha)$. Therefore, we can restrict attention to N_p and apply the foregoing argument to prove that Eq. (3.16) holds at p. Since the choice of p is arbitrary we may conclude that (3.16) holds globally.

Now, evaluate the identity (3.14) on any smooth curve of Lorentz metrics ${}^{(4)}g(\lambda)$ and differentiate twice with respect to λ and put $\lambda=0$. If, as we shall assume, $\mathrm{Ein}({}^{(4)}g_0)=0$ and $D\,\mathrm{Ein}({}^{(4)}g_0)\cdot{}^{(4)}h=0$, where

$${}^{(4)}g_0={}^{(4)}g(0)\quad\text{and}\quad{}^{(4)}h=\left.\frac{\partial\,{}^{(4)}g(\lambda)}{\partial\lambda}\right|_{\lambda=0},$$

then the result simplifies to

$$\nabla_{{}^{(4)}g_0}\cdot[D\,\mathrm{Ein}({}^{(4)}g_0)\circ{}^{(4)}\tilde{h}+D^2\,\mathrm{Ein}({}^{(4)}g_0)\circ({}^{(4)}h, {}^{(4)}h)]=0, \tag{3.17}$$

where

$${}^{(4)}\tilde{h}_{\alpha\beta}=\left.\frac{\partial^2\,{}^{(4)}g_{\alpha\beta}(\lambda)}{\partial\lambda^2}\right|_{\lambda=0}.$$

However, it follows from preceding arguments that $\nabla_{{}^{(4)}g_0}\circ(D\,\mathrm{Ein}({}^{(4)}g_0)\circ{}^{(4)}\tilde{h})\equiv 0$, so that Eq. (3.17) reduces to

$$\nabla_{{}^{(4)}g_0}\cdot(D^2\,\mathrm{Ein}({}^{(4)}g_0)\cdot({}^{(4)}h, {}^{(4)}h))=0. \tag{3.18}$$

In the above we have assumed a curve of Lorentz metrics ${}^{(4)}g(\lambda)$. However, if we are just given ${}^{(4)}g_0$ and ${}^{(4)}h$, obeying $\mathrm{Ein}({}^{(4)}g_0)=D\,\mathrm{Ein}({}^{(4)}g_0)\circ{}^{(4)}h=0$, we can still derive Eq. (3.18) by appealing to Lemma 3.1 to ensure the locally Lorentzian character of the curve ${}^{(4)}g_0+\lambda{}^{(4)}h$. For any point p, restrict to a neighborhood N_p and choose an $\alpha>0$ for which ${}^{(4)}g_0+\lambda{}^{(4)}h$ is Lorentzian on N_p for all $\lambda\in(-\alpha,\alpha)$. Then apply the above argument to prove that Eq. (3.18) holds at p. Since the choice of p is arbitrary, Eq. (3.18) holds everywhere.

Taub's result, rederived above, may thus be summarized as

Theorem 3.1: If ${}^{(4)}g$ obeys the vacuum Einstein equations and ${}^{(4)}h$ obeys the corresponding linearized Einstein equations,

$D\,\mathrm{Ein}({}^{(4)}g)\circ{}^{(4)}h=0$, then the symmetric tensor field $D^2\,\mathrm{Ein}({}^{4}g)\circ({}^{(4)}h, {}^{(4)}h)$ has vanishing divergence with respect to ${}^{(4)}g$.

If we perturb a vacuum space–time $({}^{(4)}g, {}^{(4)}V)$ with compact Cauchy surfaces and a Killing vector field ${}^{(4)}X$ we may define the conserved quantity

$$E_{{}^{(4)}X}({}^{(4)}h,\Sigma)=\int_{\Sigma}{}^{(4)}X_{\alpha}T^{\alpha\beta}({}^{(4)}g, {}^{(4)}h)n_{\beta}\,d\sigma \tag{3.19}$$

by setting

$$T^{\alpha\beta}({}^{(4)}g, {}^{(4)}h)=[D^2\,\mathrm{Ein}({}^{(4)}g)\cdot({}^{(4)}h, {}^{(4)}h)]^{\alpha\beta}. \tag{3.20}$$

There is one such conserved integral for each independent Killing vector field of the space–time. In the following section we shall show that the nonlinear constraints derived in Sec. 2, which restrict the perturbed Cauchy data on a hypersurface Σ, are equivalent to requiring that $E_{{}^{(4)}X}({}^{(4)}h,\Sigma)=0$ for each Killing vector field ${}^{(4)}X$ of the background space–time. Combining that with the hypersurface independence of the integrals $E_{{}^{(4)}X}({}^{(4)}h,\Sigma)$, we shall conclude that the nonlinear constraints of Sec. 2 are satisfied on every Cauchy hypersurface if and only if they are satisfied on any single one.

4. PROPAGATION OF THE NONLINEAR CONSTRAINTS

In this section we shall derive our main result which relates the nonlinear constraints of Sec. 2 to the conserved quantities of Sec. 3. The nonlinear constraints are expressed purely in terms of the Cauchy data specified on some hypersurface Σ whereas the conserved quantities are expressed in terms of the four-dimensional tensor fields ${}^{(4)}g$ and ${}^{(4)}h$. Our aim will be to show that, for any Killing field ${}^{(4)}X$, the conserved integral $E_{(4)_X}({}^{(4)}h, \Sigma)$ is expressible purely in terms of the Cauchy data induced on Σ and the normal and tangential projections of ${}^{(4)}X$ at this hypersurface. The connection between $E_{(4)_X}({}^{(4)}h, \Sigma)$ and the nonlinear constraint associated with ${}^{(4)}X$ will then be immediately evident.

First we shall recall some of the main points of the Cauchy development problem for the vacuum Einstein equations and for the associated linear perturbation equations. Given a pair $(g, \pi) \in \mathcal{M} \times \mathcal{S}^2_*$ which satisfies the constraints $\Phi(g,\pi)=0$, one specifies over M a time-dependent, positive definite function $N(x^k, t)$ (the lapse function) and a time-dependent vector field $N^i(x^k, t)$ (the shift vector field). One may then integrate the Einstein evolution equations to determine a Lorentzian metric ${}^{(4)}g$ on ${}^{(4)}V = (-\epsilon, \epsilon)\times M$ obeying $\mathrm{Ein}({}^{(4)}g)=0$. The resultant spacetime metric ${}^{(4)}g$ is expressible in the Arnowitt, Deser and Misner (ADM)[10] form,

$$ds^2 = -[N^2 - N_iN^i]dt\otimes dt + N_i[dt\otimes dx^i + dx^i\otimes dt] + g_{ij}dx^i\otimes dx^j, \qquad (4.1)$$

in which the $x^0=t=$constant surfaces are Cauchy surfaces for the space—time, $N_i = g_{ij}N^j$ and $g_{ij}(x^k, t)$ is the Riemannian metric induced on the $x^0=t=$constant surface. In these coordinates the momenta $\pi^{ij}(x^k, t)$ induced on the $x^0=t=$constant hypersurfaces are given by

$$\pi^{ij} = (\det g)^{1/2}(g^{ij}g^{mn} - g^{im}g^{jn})k_{mn}, \qquad (4.2)$$

where k_{mn} is the second fundamental form induced on these hypersurfaces,

$$k_{mn} = -N^{(4)}\Gamma^0_{mn}({}^{(4)}g). \qquad (4.3)$$

The Cauchy problem for the vacuum Einstein equations is treated extensively by Choquet-Bruhat in Ref. 11 and by Fischer and Marsden in Ref. 12 while the ADM formalism is discussed in detail in Ref. 10.

The Cauchy problem for the linearized Einstein equations is, of course, quite similar to that for the exact equations. On some Cauchy surface (Σ, g, π) for $({}^{(4)}g, {}^{(4)}V)$, one specifies perturbation Cauchy data $(h, p) \in \mathcal{S}_2 \times \mathcal{S}^2_*$ satisfying the linearized constraints, $D\Phi(g,\pi)\cdot(h,p)=0$. In addition one specifies perturbations $\delta N(x^k, t)$ and $\delta N^i(x^k, t)$ of the lapse function and shift vector field. Integration of the perturbed evolution equations determines a metric perturbation ${}^{(4)}h$ on $({}^{(4)}g, {}^{(4)}V)$ satisfying $D\,\mathrm{Ein}({}^{(4)}g)\cdot{}^{(4)}h=0$.

Now, suppose we have a Lorentzian metric ${}^{(4)}g_{\alpha\beta}$ expressed in the ADM form (4.1) on ${}^{(4)}V = (-\epsilon,\epsilon)\times M$ and let ${}^{(4)}h_{\alpha\beta}$ be an arbitrary, second rank, symmetric tensor field on ${}^{(4)}V$. We shall show that, for any point $p \in {}^{(4)}V$, there is a neighborhood N'_p of p and a number $\alpha' > 0$ such that the curve ${}^{(4)}g_{\alpha\beta}(\lambda) \equiv {}^{(4)}g_{\alpha\beta} + \lambda\,{}^{(4)}h_{\alpha\beta}$ is both Lorentzian and in the ADM form on N'_p for all $\lambda \in (-\alpha', \alpha')$. In other words, we shall show that ${}^{(4)}g_{\alpha\beta}(\lambda)$ is Lorentzian and that the $x^0=t=$constant surfaces remain spacelike on N'_p for each $\lambda\in(-\alpha',\alpha')$. This will guarantee that the ADM variables $N({}^{(4)}g(\lambda))$, etc., are defined on N'_p for each of the metrics ${}^{(4)}g_{\alpha\beta}(\lambda)$.

We already showed in Lemma (3.1) that, for any $p\in{}^{(4)}V$, there exists a neighborhood N_p and an $\alpha>0$ such that ${}^{(4)}g + \lambda{}^{(4)}h$ remains Lorentzian on N_p for all $\lambda\in(-\alpha,\alpha)$. Therefore we need only show that, by restricting to some $N'_p \subset N_p$ (with $p\in N'_p$) and some $\alpha' > 0$ (with $\alpha' \le \alpha$), we can maintain the spacelike character of the $x^0=t=$constant hypersurfaces within N'_p.

The surfaces labeled $x^0=t=$constant are spacelike with respect to ${}^{(4)}g_{\alpha\beta}(\lambda)$ if and only if ${}^{(4)}g^{00}(\lambda)<0$. Let $p\in{}^{(4)}V$ and let N_p and α be given such that ${}^{(4)}g(\lambda)$ is Lorentzian on N_p for all $\lambda\in(-\alpha,\alpha)$. Let K_p be a compact neighborhood of p contained in N_p and containing p as an interior point (e.g., K_p is the inverse image of some closed ball in $\mathbf{R}^4$ in some coordinate chart containing p). The function f defined by:

$$f: K_p\times[-\tfrac{1}{2}\alpha, \tfrac{1}{2}\alpha]\to\mathbf{R},$$

$$(q,\lambda)\mapsto f(q,\lambda) = [-{}^{(4)}g^{00}(q,\lambda)]$$

is continuous on $K_p\times[-\frac{1}{2}\alpha,\frac{1}{2}\alpha]$ and is positive definite on K_p for $\lambda=0$. We shall show that there exists an $\alpha'>0$ such that f is positive definite on $K_p\times[-\alpha',\alpha']$. Suppose that no such α' exists. Then, for each positive integer n, we can find a pair $(q_n,\lambda_n)\in K_p\times[-\alpha/2n, \alpha/2n]$ such that $f(q_n,\lambda_n)\le 0$. The sequence of points (q_n,λ_n) contained in the compact set $K_p\times[-\frac{1}{2}\alpha,\frac{1}{2}\alpha]$ contains a convergent subsequence which necessarily converges to a point with $\lambda=0$. Let (q_{n_i},λ_{n_i}), with $i=1,2,\ldots,$ be a subsequence which converges to some point $(q,0)$. Since f is continuous, the sequence $f(q_{n_i},\lambda_{n_i})$ converges to $f(q,0)$, which is necessarily greater than zero. However, this is impossible since, by construction, each of the numbers $f(q_{n_i},\lambda_{n_i})$ is less than or equal to zero. Thus there must exist an $\alpha'>0$ such that $f(q,\lambda)=[-{}^{(4)}g^{00}(q,\lambda)]$ is positive definite on $K_p\times[-\alpha',\alpha']$ and we need only choose some $N'_p\subset K_p$ and containing p to complete the argument. This last step is always possible since p is an interior point of K_p. Locally, therefore, we can always use the ADM variables for a curve of Lorentzian metrics of the form ${}^{(4)}g+\lambda{}^{(4)}h$.

A standard identity,[10] relating certain components of the Einstein tensor $\mathrm{Ein}({}^{(4)}g)$ to the hypersurface data (g,π) induced upon an $x^0=t=$constant, spacelike hypersurface, is given by

$$N\mathcal{H}(g,\pi) - N_i\delta^i(g,\pi) = 2(-\det{}^{(4)}g)^{1/2}[\mathrm{Ein}({}^{(4)}g)]^0_0,$$

$$-g_{ij}\delta^j(g,\pi) = 2(-\det{}^{(4)}g)^{1/2}[\mathrm{Ein}({}^{(4)}g)]^0_i. \qquad (4.4)$$

These equations provide the connections between the ADM constraint functions $\mathcal{H}(g,\pi)$ and $\delta^i(g,\pi)$, and the normal—normal and normal—tangential projections of the Einstein tensor $[\mathrm{Ein}({}^{(4)}g)]^\alpha_\beta$ onto the $t=$constant hypersurfaces.

Given a metric ${}^{(4)}g_{\alpha\beta}$ expressed in the ADM form (4.1) on ${}^{(4)}V=(-\epsilon,\epsilon)\times M$ and a perturbation ${}^{(4)}h_{\alpha\beta}$ defined on $({}^{(4)}g, {}^{(4)}V)$, we may apply the foregoing argument at any point $p\in {}^{(4)}V$ to ensure that ${}^{(4)}g_{\alpha\beta}(\lambda)\equiv {}^{(4)}g_{\alpha\beta}+\lambda\, {}^{(4)}h_{\alpha\beta}$ is in the ADM form on some neighborhood of p for all sufficiently small $|\lambda|$. Evaluating the identities (4.4) on this curve, we may differentiate twice with respect to λ and set $\lambda=0$. If we assume that ${}^{(4)}g(0) = {}^{(4)}g$ satisfies the exact Einstein equations, $\mathrm{Ein}({}^{(4)}g)=0$, and that ${}^{(4)}h$ satisfies the perturbed Einstein equations, $D\,\mathrm{Ein}({}^{(4)}g)\cdot {}^{(4)}h=0$, the resulting expressions are considerably simplified. In the notation of Secs. 2 and 3, we obtain

$$N[D^2\mathcal{H}(g,\pi)\cdot((h,p),(h,p))+D\mathcal{H}(g,\pi)\cdot(0,p')]$$
$$-N_i[D^2\delta^i(g,\pi)\cdot((h,p),(h,p))+D\delta^i(g,\pi)\cdot(0,p')]$$
$$=2(-\det{}^{(4)}g)^{1/2}[D^2\,\mathrm{Ein}({}^{(4)}g)\cdot({}^{(4)}h,{}^{(4)}h)]^0_0,$$
$$-g_{ij}[D^2\delta^j(g,\pi)\cdot((h,p),(h,p))+D\delta^j(g,\pi)\cdot(0,p')]$$
$$=2(-\det{}^{(4)}g)^{1/2}[D^2\,\mathrm{Ein}({}^{(4)}g)\cdot({}^{(4)}h,{}^{(4)}h)]^0_i \tag{4.5}$$

in which

$$N=N({}^{(4)}g),\quad N_i={}^{(4)}g_{0i},\quad g_{ij}={}^{(4)}g_{ij},\quad h_{ij}={}^{(4)}h_{ij},$$
$$p^{ij}=\frac{\partial\pi^{ij}}{\partial\lambda}({}^{(4)}g(\lambda))\Big|_{\lambda=0},\quad p'^{ij}=\frac{\partial^2\pi^{ij}}{\partial\lambda^2}({}^{(4)}g(\lambda))\Big|_{\lambda=0}. \tag{4.6}$$

Terms involving $\partial N/\partial\lambda$, $\partial N_i/\partial\lambda$, and $\partial^2N/\partial\lambda^2$ do not occur since these multiply factors which vanish by virtue of either the exact or the perturbed constraint equations. Terms involving $\partial^{2(4)}g(\lambda)/\partial\lambda^2$ vanish since ${}^{(4)}g(\lambda)$ is linear in λ. Notice that there is no necessity to distinguish ${}^{(4)}g^{\alpha\beta}[D^2\,\mathrm{Ein}({}^{(4)}g)\cdot({}^{(4)}h,{}^{(4)}h)]_{\beta\gamma}$ from $D^2[\mathrm{Ein}({}^{(4)}g)^{\alpha}_{\gamma}]\cdot({}^{(4)}h,{}^{(4)}h)$, since [for a λ linear curve ${}^{(4)}g(\lambda)$] the terms by which they would differ in general vanish by virtue of either $\mathrm{Ein}({}^{(4)}g)=0$ or $D\,\mathrm{Ein}({}^{(4)}g)\cdot{}^{(4)}h=0$.

Since Eqs. (4.5) hold at any point $p\in{}^{(4)}V$, they hold, in particular, at every point of the $x^0=t=\text{constant}$ hypersurface Σ on which the data (g,π) and (h,p) are induced by ${}^{(4)}g$ and ${}^{(4)}h$. Now, suppose that the space–time $({}^{(4)}g,{}^{(4)}V)$ admits a Killing vector field ${}^{(4)}X$ which, in ADM coordinates, is expressible as

$${}^{(4)}X=-Cn^{\alpha}\frac{\partial}{\partial x^{\alpha}}+X^i\partial\frac{\partial}{\partial x^i}, \tag{4.7}$$

where n^{α} is the unit, future directed, normal field to the $t=\text{constant}$ hypersurfaces. Evaluating ${}^{(4)}X$ on the $t=\text{constant}$ hypersurface (Σ,g,π) and contracting it with $[D^2\,\mathrm{Ein}({}^{(4)}g)\cdot({}^{(4)}h,{}^{(4)}h)]^0_{\alpha}$ we obtain, by means of Eqs. (4.5),

$$C[D^2\mathcal{H}(g,\pi)\cdot((h,p),(h,p))+D\mathcal{H}(g,\pi)\cdot(0,p')]$$
$$+g_{ij}X^j[D^2\delta^i(g,\pi)\cdot((h,p),(h,p))+D\delta^i(g,\pi)\cdot(0,p')]$$
$$=-2(-\det{}^{(4)}g)^{1/2}\,{}^{(4)}X^{\alpha}[D^2\,\mathrm{Ein}({}^{(4)}g)\cdot({}^{(4)}h,{}^{(4)}h)]^0_{\alpha}$$
$$=+2(\det g)^{1/2}n_{\beta}\,{}^{(4)}X^{\alpha}[D^2\,\mathrm{Ein}({}^{(4)}g)\cdot({}^{(4)}h,{}^{(4)}h)]^{\beta}_{\alpha}, \tag{4.8}$$

where we have used the formulas

$${}^{(4)}X^0=-C/N,\quad {}^{(4)}X^i=X^i+(C/N)N^i,$$
$$(-\det{}^{(4)}g)^{1/2}=N(\det g)^{1/2},\quad n_{\alpha}=-\delta^0_{\alpha}N. \tag{4.9}$$

We now integrate Eq. (4.8) over the hypersurface (Σ,g,π) and reexpress the terms involving $D\mathcal{H}(g,\pi)\cdot(0,p')$ and $D\delta^i(g,\pi)\cdot(0,p')$. Recalling Eq. (2.7) we obtain

$$\int_{\Sigma}d^3x\{CD\mathcal{H}(g,\pi)\cdot(0,p')+X_iD\delta^i(g,\pi)\cdot(0,p')\}$$
$$=\int_{\Sigma}d^3x\langle(C,X);D\Phi(g,\pi)\cdot(0,p')\rangle$$
$$=\int_{\Sigma}d^3x\langle\!\langle D\Phi(g,\pi)^*\cdot(C,X);(0,p')\rangle\!\rangle$$
$$=0. \tag{4.10}$$

The last equality follows from Lemma (4.1) of paper I. This lemma states that $D\Phi(g,\pi)^*\cdot(C,X)=0$ provided that C and X are the normal and tangential projections, onto a Cauchy surface (Σ,g,π), of a Killing vector field for the vacuum space–time determined by (g,π). Consequently we have, upon integrating Eq. (4.8) over Σ,

$$\int_{\Sigma}d^3x\{CD^2\mathcal{H}(g,\pi)\cdot((h,p),(h,p))$$
$$+X_iD^2\delta^i(g,\pi)\cdot((h,p),(h,p))\}$$
$$=+2\int_{\Sigma}d^3x\{(\det g)^{1/2}n_{\beta}\,{}^{(4)}X^{\alpha}[D^2\,\mathrm{Ein}({}^{(4)}g)$$
$$({}^{(4)}h,{}^{(4)}h)]^{\beta}_{\alpha}\} \tag{4.11}$$

or, in the notation of Secs. 2 and 3,

$$\int_{\Sigma}d^3x\langle(C,X);D^2\Phi(g,\pi)\cdot((h,p),(h,p))\rangle=+2E_{{}^{(4)}X}({}^{(4)}h,\Sigma). \tag{4.12}$$

This equation holds provided, as we have assumed, that ${}^{(4)}g$ satisfies the Einstein equations, ${}^{(4)}h$ satisfies the perturbed Einstein equations, (g,π) and (h,p) are the exact and perturbed Cauchy data induced on Σ by ${}^{(4)}g$ and ${}^{(4)}h$, and ${}^{(4)}X$ is a Killing field of $({}^{(4)}g,{}^{(4)}V)$ with projections (C,X) at Σ.

As shown by Theorem (2.1), the left side of Eq. (4.12) must be constrained to vanish as a necessary condition to exclude spurious perturbations. There is one such condition for each linearly independent Killing vector field of $({}^{(4)}g,{}^{(4)}V)$. It follows from Eq. (4.12) that $E_{{}^{(4)}X}({}^{(4)}h,\Sigma)$ must be constrained to vanish. As shown in Sec. 3, however, $E_{{}^{(4)}X}({}^{(4)}h,\Sigma)$ is a conserved quantity whose value is independent of the choice of Cauchy surface Σ. Thus $E_{{}^{(4)}X}({}^{(4)}h,\Sigma)$ must be constrained to vanish on at least one (and therefore on every) Cauchy surface of the space–time $({}^{(4)}g,{}^{(4)}V)$. It follows that, if the condition (2.10) is imposed on some initial surface Σ and if the perturbations are propagated to any other Cauchy surface (Σ',g',π'), then the perturbation data (h',p') induced on (Σ',g',π') will also satisfy condition (2.10). In summary, we have

Theorem (4.1): Let $({}^{(4)}g,{}^{(4)}V)$ be a vacuum space–time with compact Cauchy surfaces and a Killing vector field ${}^{(4)}X$, and impose the nonlinear constraint (2.10) upon perturbation Cauchy data (h,p) at some Cauchy surface (Σ, g, π). Then the corresponding constraint (2.10) will hold at every other Cauchy surface provided the perturbation data is propagated by the linearized Einstein equations, $D\,\mathrm{Ein}({}^{(4)}g)\cdot{}^{(4)}h=0$. This nonlinear constraint is equivalent to $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)=0$ which holds on every Cauchy surface if it holds on any single one.

This theorem assures us that a perturbation cannot appear acceptable on some initial Cauchy surface but spurious when propagated to some other Cauchy surface of the space—time.

It seems reasonable to conjecture that, in addition to being necessary, the nonlinear constraints discussed above are also sufficient conditions to exclude all spurious linear perturbations. However, no proof of sufficiency seems yet to be known. One can, of course, extend the nonlinear constraints of Sec. 2 to higher order perturbation theory by differentiating $\Phi(g(\lambda), \pi(\lambda))$ three or more times with respect to λ and proceeding as before. But the new constraints obtained in this way would always seem to involve the higher order perturbations $\partial^n g(\lambda)/\partial\lambda^n$ and $\partial^n\pi(\lambda)/\partial\lambda^n$, with $n\geq 2$. Thus they would provide no additional restrictions purely upon the first order perturbations (h, p) but would instead provide the corresponding restrictions upon the higher order perturbations. By perturbing the Bianchi identities to higher order one should also be able to derive conservation theorems for these higher order nonlinear constraints.

5. GAUGE INVARIANCE OF THE NONLINEAR CONSTRAINTS

It is well known that if ${}^{(4)}h$ is a solution of the linearized Einstein equations, defined over some vacuum space—time $({}^{(4)}g, {}^{(4)}V)$, then so is ${}^{(4)}h_{\alpha\beta}+{}^{(4)}Y_{\alpha;\beta}+{}^{(4)}Y_{\beta;\alpha}$ for any vector field ${}^{(4)}Y$. The special perturbations ${}^{(4)}Y_{\alpha;\beta}+{}^{(4)}Y_{\beta;\alpha}\equiv[\mathcal{L}_{{}^{(4)}Y}\,{}^{(4)}g]_{\alpha\beta}$ (called gauge transformations) represent infinitesimal coordinate transformations and so are physically trivial. One would therefore not expect such gauge transformations to affect the satisfaction of the nonlinear constraints. Indeed, we shall show here that the conserved quantities $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)$ are gauge invariant so that if $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)=0$ for some perturbation ${}^{(4)}h$, then $E_{{}^{(4)}X}({}^{(4)}h+\mathcal{L}_{{}^{(4)}Y}\,{}^{(4)}g, \Sigma)=0$ also holds for any vector field ${}^{(4)}Y$.

To obtain this result we note that the integral $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)$ depends upon ${}^{(4)}h$ only through its values and the values of its first and second derivatives at the hypersurface Σ. Thus if ${}^{(4)}h$ is held fixed on any tubular neighborhood (i.e., an open region bounded by two disjoint Cauchy surfaces) containing Σ, then $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)$ cannot change in value. Conversely, if we want to investigate the possible change in value of $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)$ under some gauge transformation of ${}^{(4)}h$, we need only know the transformed ${}^{(4)}h$ on an arbitrary tubular neighborhood of Σ.

Suppose we could alter the value of $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)$ by a gauge transformation generated by some vector field ${}^{(4)}Y$. This same change would be effected by any vector field ${}^{(4)}Z$ which agrees with ${}^{(4)}Y$ on an arbitrary tubular neighborhood of Σ. Choose a ${}^{(4)}Z$ which agrees with ${}^{(4)}Y$ on some tubular neighborhood of Σ but which vanishes on some tubular neighborhood of another Cauchy surface Σ' disjoint from Σ. We then have

$$\begin{aligned}E_{{}^{(4)}X}({}^{(4)}h+\mathcal{L}_{{}^{(4)}Y}\,{}^{(4)}g, \Sigma)&=E_{{}^{(4)}X}({}^{(4)}h+\mathcal{L}_{{}^{(4)}Z}\,{}^{(4)}g, \Sigma)\\&=E_{{}^{(4)}X}({}^{(4)}h+\mathcal{L}_{{}^{(4)}Z}\,{}^{(4)}g, \Sigma')=E_{{}^{(4)}X}({}^{(4)}h, \Sigma')\\&=E_{{}^{(4)}X}({}^{(4)}h, \Sigma),\end{aligned}\tag{5.1}$$

where the next to last equality follows from the vanishing of ${}^{(4)}Z$ (and thus $\mathcal{L}_{{}^{(4)}Z}\,{}^{(4)}g$) on a neighborhood of Σ' and where the last equality follows from the hypersurface independence of $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)$. We conclude that

$$E_{{}^{(4)}X}({}^{(4)}h+\mathcal{L}_{{}^{(4)}Y}\,{}^{(4)}g, \Sigma)=E_{{}^{(4)}X}({}^{(4)}h, \Sigma)\tag{5.2}$$

for any ${}^{(4)}Y$. An immediate corollary is that

$$E_{{}^{(4)}X}(\mathcal{L}_{{}^{(4)}Y}\,{}^{(4)}g, \Sigma)=0,$$

which means that pure gauge perturbations automatically satisfy the nonlinear constraints.

The above results are quite reasonable intuitively. The set of solutions of the perturbation equations may be divided into equivalence classes; two solutions belong to the same class if and only if they differ by a mere gauge transformation. All the classes except for that of pure gauge perturbations represent perturbations towards space—times distinct from the given one. It is quite natural that the nonlinear constraints which arise when Killing symmetries are present only impose restrictions upon complete classes of perturbations. They do not distinguish between different perturbations within the same class, since the latter are physically equivalent.

In paper I we discussed the gauge transformations of perturbation data (h, p) induced by an arbitrary vector field ${}^{(4)}Y$. In Ref. 13 we used this result to decompose the space of tangent vectors (h, p) satisfying $D\Phi(g, \pi)\cdot(h, p)=0$ into a direct sum of two subspaces. One of these subspaces contains all the pure gauge perturbations induced at the hypersurface (Σ, g, π). The other subspace is, with respect to a convenient inner product in the tangent space $T_{(g,\pi)}\mathcal{M}\times\mathcal{S}^2_*$, orthogonal to the gauge subspace. Thus every solution of $D\Phi(g, \pi)\cdot(h, p)=0$ can be split uniquely into two terms,

$$(h, p)=(h, p)_\perp+(h, p)_{\text{gauge}},\tag{5.3}$$

which both separately satisfy the perturbed constraints and for which one of the terms, $(h, p)_{\text{gauge}}$, is always a pure gauge perturbation. It follows from the foregoing results on the gauge invariance of the integrals $E_{{}^{(4)}X}({}^{(4)}h, \Sigma)$ and the relation of these integrals to the nonlinear constraints (2.10), that

$$\begin{aligned}&\int_\Sigma d^3x\langle(C, X); D^2\Phi(g, \pi)\cdot((h, p), (h, p))\rangle\\&\quad=\int_\Sigma d^3x\langle(C, X); D^2\Phi(g, \pi)\cdot((h, p)_\perp, (h, p)_\perp)\rangle\end{aligned}\tag{5.4}$$

with no dependence on $(h, p)_{\text{gauge}}$. This result generalizes that of Brill and Deser[5] who applied a standard transverse-traceless decomposition to the perturbations of a flat space—time with compact, flat Cauchy hypersurfaces.

6. DISCUSSION

To this point we have **only** considered the perturbation of vacuum space–times with compact Cauchy surfaces. If we consider perturbing space–times with noncompact Cauchy surfaces instead, then our conclusions must be considerably modified. For the noncompact case the arguments of Sec. 2 no longer imply nonlinear constraints upon the first order perturbations, but instead give formulas relating the quantities $E_{(4)_X}({}^{(4)}h, \Sigma)$ to certain boundary integrals involving the second order perturbations. Since these formulas have some potentially interesting applications to the perturbations of asymptotically flat space–times, we shall briefly discuss their derivation and significance here.

It follows from Lemma (4.1) of paper I that if (g, π) is Cauchy data for a vacuum space–time which admits a Killing vector field ${}^{(4)}X$, then the projections (C, X) of ${}^{(4)}X$ onto (Σ, g, π) satisfy $D\Phi(g,\pi)^* \cdot (C, X) = 0$. This is a purely local result which does not depend upon compactness of Σ. However, Eq. (2.7) relating $D\Phi(g,\pi)$ to $D\Phi(g,\pi)^*$ no longer holds, since the partial integrations give divergence terms or (equivalently) surface integrals over the boundary $\partial\Sigma$ of Σ. The general formula is

$$\begin{aligned}
&\int_\Sigma d^3x \langle (C,X); D\Phi(g,\pi)\circ(h,p)\rangle - \int_\Sigma d^3x \langle\langle D\Phi(g,\pi)^*\circ(C,X);(h,p)\rangle\rangle \\
&\quad = \int_\Sigma d^3x \{-[(\det g)^{1/2} C h^{ij}{}_{|i}]_{|j} + [(\det g)^{1/2} C (h^i{}_i)^{|j}]_{|j} \\
&\qquad + [(\det g)^{1/2} C_{|i} h^{ij}]_{|j} - [(\det g)^{1/2} C^{|j} h^i{}_i]_{|j} \\
&\qquad + 2(X_i p^{ij})_{|j} + [2X^i \pi^{jk} h_{ik} - X^j \pi^{ik} h_{ik}]_{|j}\},
\end{aligned} \qquad (6.1)$$

in which the right side is the integral of a pure divergence.

The argument given in Sec. 2, which led before to nonlinear constraints upon the first order perturbations, no longer obtains. If one assumes a curve of exact solutions of the constraints, differentiates $\Phi(g(\lambda), \pi(\lambda)) = 0$ twice with respect to λ, contracts with projections (C, X) of a Killing field ${}^{(4)}X$, and integrates over Σ, he is left, by virtue of Eq. (6.1), with boundary integrals involving the second order perturbations, $(\partial^2 g/\partial\lambda^2, \partial^2\pi/\partial\lambda^2)_{\lambda=0}$. Thus, instead of obtaining nonlinear constraints upon the first order perturbations, one obtains formulas relating the quantities $E_{(4)_X}({}^{(4)}h, \Sigma)$ to boundary integrals involving the second order perturbations. One can interpret the boundary integrals at spatial infinity as second order changes in the asymptotically defined energy (timelike ${}^{(4)}X$), momentum (spacelike, translational ${}^{(4)}X$) or angular momentum (spacelike, rotational ${}^{(4)}X$) of the perturbed space–time. Thus the integrals $E_{(4)_X}({}^{(4)}h, \Sigma)$ may be thought of as contributions to the total energy, momentum, or angular momentum which results from the first order perturbation ${}^{(4)}h$.

For perturbations of black holes such as those represented by the Schwarzschild or Kerr space–times, it may be of interest to consider a partial Cauchy surface with an inner boundary $\partial\Sigma_h$ at the event horizon of the unperturbed space–time in addition to the outer boundary $\partial\Sigma_\infty$ at spatial infinity. In this case, one expects to relate the surface integrals over $\partial\Sigma_h$ to properties of the perturbed (apparent) event horizon.

One motivation for studying the effect of first order perturbations upon second order boundary perturbations is the possible application of these results to the recently discovered Hawking process of spontaneous particle production by black holes. Hawking[8] has initiated a study of the properties of quantized fields on certain space–times which represent gravitational collapse and the formation of a black hole. He finds that a production of particles (field quanta) occurs for which the spectrum of particles escaping to infinity is that of a thermal distribution. It is as though the black hole were a body at some nonzero temperature which emits photons and other particles into the surrounding vacuum. In the initial investigations by Hawking and others, the reaction of this particle production upon the black hole was not taken into account. The perturbation results described above suggest a means of computing the reaction effects up to the second order of approximation. While we have only discussed the gravitational perturbations (which would become "gravitons" in a quantum mechanical treatment), it is straightforward to extend these results to include the electromagnetic and other standard perturbations.

The idea would be to treat the perturbations (h_{ij}, p^{ij}, etc.) quantum mechanically and to construct operators modeled on the quantities $E_{(4)_X}({}^{(4)}h, \Sigma)$ which represent the energy or angular momentum of the quantized perturbations. One could then define the (second order) boundary integral operators through the use of the formulas obtained from Eq. (6.1). Even on the classical level the integrals $E_{(4)_X}({}^{(4)}h, \Sigma)$ are conserved only in the absence of energy or angular momentum flow across the boundaries. The Hawking process should lead to particles crossing the event horizon and modifying the properties of the black hole. The formalism mentioned above suggests a natural (perturbative) method for studying precisely these reaction effects upon the black hole. We propose to develop this treatment and its physical interpretation more carefully in subsequent work.

APPENDIX

The explicit formulas for $D^2\mathcal{H}(g,\pi)\cdot((h,p),(h,p))$ and $D^2\delta^i(g,\pi)\circ((h,p),(h,p))$ are given by

$$\begin{aligned}
&D^2\mathcal{H}(g,\pi)\circ((h,p),(h,p)) \\
&\quad = (\det g)^{-1/2}\{[\tfrac{1}{4}(h^k{}_k)^2 + \tfrac{1}{2}h^{mn}h_{mn}][\pi^{ij}\pi_{ij} - \tfrac{1}{2}(\pi^i{}_i)^2] \\
&\qquad - 2h^m{}_m(h_{ik}g_{jl} - \tfrac{1}{2}h_{ij}g_{kl})\pi^{ij}\pi^{kl} - 2h^k{}_k(\pi_{ij}p^{ij} \\
&\qquad - \tfrac{1}{2}\pi^i{}_i p^j{}_j) + 2(h_{ik}h_{jl} - \tfrac{1}{2}h_{ij}h_{kl})\pi^{ij}\pi^{kl} \\
&\qquad + 2(h_{ik}g_{jl} + g_{ik}h_{jl} - \tfrac{1}{2}h_{ij}g_{kl} - \tfrac{1}{2}g_{ij}h_{kl})\cdot(p^{ij}\pi^{kl} \\
&\qquad + \pi^{ij}p^{kl}) + 2(g_{ik}g_{jl} - \tfrac{1}{2}g_{ij}g_{kl})\circ(p^{ij}p^{kl})\} \\
&\qquad - (\det g)^{1/2}\{[\tfrac{1}{4}(h^i{}_i)^2 - \tfrac{1}{2}h^{ij}h_{ij}]R - h^k{}_k h^{ij}R_{ij} \\
&\qquad + 2h^{im}h_m{}^j R_{ij} - h^{ij}(h_{ik|j}{}^{|k} + h_{kj|i}{}^{|k}) + h^k{}_k(h_{ij}{}^{|ij} \\
&\qquad - (h^i{}_i)_{|j}{}^{|j}) + 2h^{kl}(h_{kl|i}{}^{|i} + (h^i{}_i)_{|kl} - h_{ik}{}^{|i}{}_{|l})
\end{aligned}$$

$$+\left[\tfrac{1}{2}(h^j{}_j)^{|k} - h^{jk}{}_{|j}\right]\cdot(2h_{ik}{}^{|i} - (h^l{}_l)_{|k})$$

$$+\tfrac{1}{2}h_{kl|i}h^{kl|i} + h_i{}^{k|l}(h_k{}^i{}_{|l} - h^i{}_{l|k})\}$$

and

$$D^2\delta^i(g,\pi)\cdot((h,p),(h,p))$$

$$=2p^{jk}g^{il}(h_{jl|k} + h_{kl|j} - h_{jk|l})$$

$$-2\pi^{jk}h^{il}(h_{jl|k} + h_{kl|j} - h_{jk|l}).$$

These formulas are also useful for the construction of a variational principle for the linearized evolution equations. For an application to the special case of perturbing the Reissner–Nordstrom family of black holes, see Ref. 14.

ACKNOWLEDGMENTS

I am grateful to A. H. Taub, Jerrold Marsden, and Arthur Fischer for many helpful discussions.

*Supported in part by NSF Grant GP-43718X.

†Address after August 1, 1976: Dept. of Physics, Yale University, New Haven, Connecticut 06520.

[1]V. Moncrief, J. Math. Phys. **16**, 493 (1975).

[2]A. Fischer and J. Marsden, Bull. Am. Math. Soc. **79**, 997 (1973).

[3]A. Fischer and J. Marsden, Proc. Symp. Pure Math. **27**, 219 (1975).

[4]D. Brill, "Isolated solutions in General Relativity," University of Maryland Technical Report No. 71-076 (1971).

[5]D. Brill and S. Deser, Commun. Math. Phys. **32**, 291 (1973).

[6]A. Taub, J. Math. Phys. **2**, 787 (1961).

[7]Taub's general argument is based upon the use of a Lagrangian variational principle for the gravitational perturbation equations. He derives a suitable variational integral in: A. Taub, Commun. Math. Phys. **15**, 235 (1969). The remainder of the argument, which leads to a divergence-free stress tensor for the perturbations, is given in: A. Taub, *Relativistic Fluid Dynamics*, edited by C. Cattaneo (Lectures at the Centro Internazionale Matematico Estivo, Bressanone, 1970), Edizioni Cremonese, Rome (1971). See also: S. Hawking and G. Ellis, *The Large Scale Structure of Space—Time* (Cambridge, London, 1973), Sec. (3.3).

[8]S. Hawking, Commun. Math. Phys. **43**, 199 (1975).

[9]R. Geroch, J. Math. Phys. **11**, 343 (1970).

[10]The ADM formalism is discussed in detail in Chapter 21 of C. Misner, K. Thorne, and J. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[11]Y. Choquet-Bruhat, on the Cauchy problem in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).

[12]A. Fischer and J. Marsden, Commun. Math. Phys. **28**, 1 (1972).

[13]V. Moncrief, J. Math. Phys. **16**, 1556 (1975).

[14]V. Moncrief, Phys. Rev. D **12**, 1526 (1975).

††*Note added in proof*: A proof that the Fischer—Marsden adjoint map must have trivial kernel to ensure linearization stability has not yet appeared in the literature. One can establish this necessity by showing that the second order constraints discussed here are always nontrivial. This issue together with the sufficiency of the second order constraints to exclude all spurious perturbations will be discussed in a forthcoming paper by A. Fischer, J. Marsden, and V. Moncrief.

Tensor spherical harmonics and tensor multipoles. I. Euclidean space

M. Daumens and P. Minnaert

Laboratoire de Physique Théorique, Université de Bordeaux I, 33170 Gradignan, France*
(Received 29 October 1975)

Two bases in the Hilbert space of tensor fields on the unit sphere are discussed: the tensor spherical harmonics and the tensor multipoles. For vector fields these two bases are related by an orthonormal transformation whose coefficients are shown to be Clebsch–Gordan coefficients. This remark suggests a method of building multipole bases for higher order tensor fields. The second order tensor multipoles are studied in detail as well as their relations with the symmetric ones defined by several authors for application to the gravitational radiation.

1. INTRODUCTION

Many problems of mathematical physics require a knowledge of bases in the Hilbert spaces of complex tensor fields on the unit sphere $\mathcal{S}^2$, embedded in the three-dimensional Euclidean space $\mathcal{E}^3$. In this paper, we describe two such bases, the *tensor spherical harmonics* and the *tensor multipoles*, and their relationship.

The rth-order tensor spherical harmonics are built by coupling the scalar spherical harmonics with the irreducible tensor basis of $(\mathcal{E}^3)^{\otimes r}$, through Clebsch-Gordan coefficients, so that they transform according to a given representation of the rotation group. Of course for the lowest-order tensor fields this basis is already known: for $r=0$ it is the scalar spherical harmonics themselves; for $r=1$ it is the vector spherical harmonics of Blatt and Weisskopff;[1] for $r=2$ it is the tensor spherical harmonics defined by Mathews[2] and Zerilli.[3]

Without speaking of the scalar fields for which it is exactly the scalar spherical harmonics, the basis that we call tensor multipole basis has been already considered for the first- and second-order tensor fields. Indeed in the study of electromagnetic radiation[1,4,5] one has been led to consider the basis of vector fields obtained by action of the vector operators $\mathbf{r}/r$, $r\nabla$, and $-i\mathbf{r}\times\nabla$ on scalar spherical harmonics. Quite in the same spirit, and motivated by the work of Regge and Wheeler[6] on gravitational radiation, Zerilli[3] has built a basis for second-order symmetric tensors by action of the tensor products of the previous operators on scalar spherical harmonics. For both cases it has been noticed that the derived vectors and tensors are orthogonal linear combinations of the corresponding vector and tensor spherical harmonics. However the structure of this relationship has not been clearly exhibited and studied in great detail. We shall emphasize that the coefficients of the linear transformation for vector and second-order tensor fields are actually Clebsch—Gordan coefficients. This remark provides a powerful tool to build multipole bases in the Hilbert spaces of higher-order tensor fields which may be useful for the description of strong interactions since there exist massive bosonic states with higher spins. Moreover, many properties of the tensor multipoles can be easily deduced from those of the tensor spherical harmonics. In particular it will be easily shown that, besides being orthonormal functions for the scalar product in their Hilbert space, the tensor multipoles are orthogonal for the scalar product in $(\mathcal{E}^3)^{\otimes r}$ whereas the tensor spherical harmonics were not.

This paper studies successively vector fields (Sec. 2), second-order tensor fields (Sec. 3) and arbitrary rth-order tensor fields (Sec. 4). For each tensorial order we build the irreducible tensor basis of $(\mathcal{E}^3)^{\otimes r}$, the tensor spherical harmonics and the tensor multipoles.

Throughout the paper Euclidean tensors are denoted by bold face letters. We use the summation convention for repeated cartesian and magnetic quantum number indices but we always write explicit summations over angular momentum indices. The scalar products in $\mathcal{E}^3$, $\mathcal{E}^3\otimes\mathcal{E}^3$ and $(\mathcal{E}^3)^{\otimes r}$ are denoted by a single dot (.), a double dot(:) and $(\overset{r}{.})$, respectively, e.g., $\mathbf{u}\cdot\mathbf{v}=u_iv_i$, $\mathbf{t}:\delta=t_{ij}\delta_{ij}$, $\mathbf{t}(\overset{r}{.})\mathbf{T}=t_{i_1\ldots i_r}T_{i_1\ldots i_r}$. Concerning rotation matrices, Clebsch—Gordan coefficients (CG coefficients) and $6j$-symbols, the reader is referred to the books of Wigner,[7] Rose,[8] Edmonds,[9] and Brink and Satchler.[10] Finally, we use the short notation $\hat{X}=(2X+1)^{1/2}$.

2. VECTOR FIELDS

A. Spherical basis

Consider three vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ which form a right-handed orthonormal basis, denoted by $\{\mathbf{e}\}$, of the real Euclidean space $\mathcal{E}^3$. In the complexification $\mathcal{E}^3_c$ of this space, the spherical basis associated to $\{\mathbf{e}\}$ is constituted by the three complex vectors $\mathbf{e}_\pm=\mp(1/\sqrt{2})(\mathbf{e}_1\pm i\mathbf{e}_2)$, $\mathbf{e}_0=\mathbf{e}_3$. These vectors satisfy the identity $(\mathbf{e}_n)^*=(-1)^n\mathbf{e}_{-n}$, where (*) means complex conjugation. They also verify the following orthonormality, orientation, and closure relations

$$\mathbf{e}^*_m\cdot\mathbf{e}_n=\delta_{mn}, \tag{1}$$

$$(\mathbf{e}_m,\mathbf{e}_n,\mathbf{e}_r)\equiv\mathbf{e}_m\cdot\mathbf{e}_n\times\mathbf{e}_r=-i\epsilon_{mnr}, \tag{2}$$

$$\mathbf{e}^*_n\otimes\mathbf{e}_n=\delta, \tag{3}$$

where ϵ_{mnr} is the Levi-Civita tensor for the spherical basis (with $\epsilon_{+0-}=1$) and δ the identity tensor. Note that the tensor ϵ_{mnr} can be represented by a CG coefficient

$$\epsilon_{mnr}=(-1)^{r+1}\sqrt{2}\langle 1m1n|1-r\rangle. \tag{4}$$

Under a rotation R of the basis $\{\mathbf{e}\}$, the spherical basis vectors transform according to the representation $D^1(R)$

of the rotation group

$$R\mathbf{e}_n = \mathbf{e}_m D^1(R)^m_n. \tag{5}$$

In the basis $\{\mathbf{e}\}$, the spherical components v_n of a vector $\mathbf{v}$ are

$$v_n = \mathbf{v}\cdot\mathbf{e}_n, \quad \mathbf{v} = v_n \mathbf{e}_n^*, \tag{6}$$

and, in the rotated basis $R\{\mathbf{e}\}$, the components v_n^R are

$$v_n^R \equiv \mathbf{v}\cdot R\mathbf{e}_n = v_m D^1(R)^m_n. \tag{7}$$

The second-order dual tensor of an arbitrary vector is denoted by $\mathbf{v}^\times$ and defined by

$$\mathbf{v}^\times = (\mathbf{e}_m, \mathbf{v}, \mathbf{e}_n)\mathbf{e}_m^* \otimes \mathbf{e}_n^*. \tag{8}$$

In particular, the dual tensors of the spherical basis vectors are

$$\mathbf{e}_m{}^\times = i\epsilon_{mnr}\mathbf{e}_n^* \otimes \mathbf{e}_r^* = i\sqrt{2}\langle 1n1r|1m\rangle \mathbf{e}_n \otimes \mathbf{e}_r. \tag{9}$$

The three tensors $S_n = i\mathbf{e}_n{}^\times$ can be considered as the spherical components of a vector operator $\mathbf{S}$. They act on $\mathcal{E}_c^3$ by $S_n\mathbf{v} = i\mathbf{e}_n \times \mathbf{v}$. They form a spin-one realization of the angular momentum operator

$$\mathbf{S}\times\mathbf{S} = i\mathbf{S}, \quad \mathbf{S}^2 = 2\delta, \tag{10}$$

and the three vectors $\mathbf{e}_n$ are a realization of the spin-one standard basis, in particular,

$$S_3\mathbf{e}_n = n\mathbf{e}_n. \tag{11}$$

B. Scalar spherical harmonics

To any vector $\mathbf{r}$ of $\mathcal{E}^3$ with spherical coordinates (r, θ, φ) in the basis $\{\mathbf{e}\}$, we associate a unit vector $\mathbf{u} = \mathbf{r}/r$, whose extremity lies on the unit sphere $\mathcal{S}^2$. Let us consider the space $\mathcal{L}_0^2(\mathcal{S}^2)$ of scalar complex functions $f(\mathbf{u})$ on $\mathcal{S}^2$, square integrable ($\int |f(\mathbf{u})|^2 d\mathbf{u} < \infty$) with respect to the usual measure $d\mathbf{u} = d(\cos\theta)\, d\varphi$ on the unit sphere. It is a Hilbert space for the scalar product

$$\langle f_1, f_2\rangle = \int_{\mathcal{S}^2} f_1(\mathbf{u})^* f_2(\mathbf{u})\, d\mathbf{u}. \tag{12}$$

Let $\mathbf{L} = -i\mathbf{r}\times\nabla$ be the orbital angular momentum operator. An orthonormal basis of the space $\mathcal{L}_0^2(\mathcal{S}^2)$ is the set of the spherical harmonics $Y^l_m(\theta,\varphi)$ which are eigenfunctions of $\mathbf{L}^2$ and L_3 with the eigenvalues $l(l+1)$ and m. In the following we shall denote the spherical harmonics by $Y^l_m(\mathbf{u},\{\mathbf{e}\})$ or simply $Y^l_m(\mathbf{u})$ when no confusion is possible. Using CG coefficients, they are built from the spherical components of $\mathbf{u}$ by an iterative process with $Y^0_0 = (4\pi)^{-1/2}$

$$Y^l_m(\mathbf{u}) = \hat{l}/\sqrt{l}\,\langle l-1m'1n|lm\rangle Y^{l-1}_{m'}(\mathbf{u})u_n. \tag{13}$$

In particular, one has $Y^1_n(\mathbf{u}) = (3/4\pi)^{1/2}u_n$. They satisfy the identities

$$Y^l_m(\mathbf{u})^* = (-1)^m Y^l_{-m}(\mathbf{u}), \quad Y^l_m(-\mathbf{u}) = (-1)^l Y^l_m(\mathbf{u}), \tag{14}$$

and the orthonormality relation

$$\langle Y^l_m, Y^{l'}_{m'}\rangle = \delta_{ll'}\delta_{mm'}. \tag{15}$$

Under a rotation of the basis, the transformation law of the scalar spherical harmonics is

$$Y^l_m(\mathbf{u}, R\{\mathbf{e}\}) = Y^l_{m'}(\mathbf{u},\{\mathbf{e}\})D^l(R)^{m'}_m, \tag{16}$$

i.e., at fixed l the $2l+1$ spherical harmonics transform according to the lth-order representation of the rotation group. For higher-order tensor fields we shall look for bases which have similar transformation properties under a rotation of the basis.

The product of two spherical harmonics is a scalar function on $\mathcal{S}^2$, which can be expanded on the basis of spherical harmonics. The corresponding reduction formula is given in Appendix A, Eq. (A1).

C. Vector spherical harmonics (Refs. 1 and 4)

The space $\mathcal{L}_1^2(\mathcal{S}^2)$ of complex vector fields on the sphere $\mathcal{S}^2$, with integrable modulus squared, is a Hilbert space for the scalar product

$$\langle \mathbf{f}_1, \mathbf{f}_2\rangle = \int_{\mathcal{S}^2} \mathbf{f}_1(\mathbf{u})^*\cdot\mathbf{f}_2(\mathbf{u})\, d\mathbf{u}. \tag{17}$$

The vector functions $Y^l_m(\mathbf{u})\mathbf{e}_n$, which are eigenvectors of the set of operators $\{\mathbf{L}^2, L_3, \mathbf{S}^2, S_3\}$, form an orthonormal basis of this space

$$\langle Y^l_m\mathbf{e}_n, Y^{l'}_{m'}\mathbf{e}_{n'}\rangle = \langle Y^l_m, Y^{l'}_{m'}\rangle \mathbf{e}_n^*\cdot\mathbf{e}_{n'} = \delta_{ll'}\delta_{mm'}\delta_{nn'}, \tag{18}$$

but in a rotation of the basis $\{e\}$ they transform according to the tensor product of representations $D^l\otimes D^1$. Then by coupling the spherical harmonics $Y^l_m(\mathbf{u})$ and the basis vectors $\mathbf{e}_n$ with the CG coefficients which reduce this product of representations, one gets the vector spherical harmonics (VSH)

$$\mathbf{Y}^{lJ}_M(\mathbf{u}) = \langle lm\,1n|JM\rangle Y^l_m(\mathbf{u})\mathbf{e}_n \tag{19}$$

which transform according to the representation D^J,

$$\mathbf{Y}^{lJ}_M(\mathbf{u}, R\{\mathbf{e}\}) = \mathbf{Y}^{lJ}_{M'}(\mathbf{u},\{\mathbf{e}\})D^J(R)^{M'}{}_M. \tag{20}$$

Let us call $\mathbf{J} = \mathbf{L}+\mathbf{S}$ the total angular momentum operators, then the VSH are eigenfunctions of the set of operators $\{\mathbf{J}^2, J_3, \mathbf{L}^2, \mathbf{S}^2\}$ with the eigenvalues $J(J+1)$, M, $l(l+1)$, 2, respectively. The VSH satisfy the orthonormality relation

$$\langle \mathbf{Y}^{lJ}_M, \mathbf{Y}^{l'J'}_{M'}\rangle = \delta_{ll'}\delta_{JJ'}\delta_{MM'}, \tag{21}$$

and they also verify the identities,

$$\mathbf{Y}^{lJ}_M(\mathbf{u})^* = (-1)^{l+J+M+1}\mathbf{Y}^{lJ}_{-M}(\mathbf{u}), \tag{22}$$

$$\mathbf{Y}^{lJ}_M(-\mathbf{u}) = (-1)^l\mathbf{Y}^{lJ}_M(\mathbf{u}), \tag{23}$$

where the sign $(-1)^l$ is the intrinsic parity of the VSH. The scalar product in $\mathcal{E}_c^3$ of two VSH $\mathbf{Y}^{lJ}_M\cdot\mathbf{Y}^{l'J'}_{M'}$ is a scalar function whose expansion in terms of the spherical harmonics is given in Appendix A, Eq. (A2).

For $J=0$ one has only one VSH, $\mathbf{Y}^{10}_0(\mathbf{u}) = -(4\pi)^{-1/2}\mathbf{u}$, while for fixed $J\geq 1$ and M, one has three VSH, $\mathbf{Y}^{JJ}_M(\mathbf{u})$ with parity $(-1)^J$ and $\mathbf{Y}^{J\pm1\,J}_M(\mathbf{u})$ with parity $(-1)^{J+1}$. The equation (A2) for $J=J'$ and $M=M'$ exhibits the geometrical properties in $\mathcal{E}_c^3$ of the VSH. The symmetry of the CG coefficients implies that the product $\mathbf{Y}^{lJ}_M\cdot\mathbf{Y}^{l'J}_M$ vanishes for $(-1)^l\neq(-1)^{l'}$. Hence the vector $\mathbf{Y}^{JJ}_M$ is perpendicular to both vectors $\mathbf{Y}^{J\pm1\,J}_M$ but these latter are not mutually perpendicular. This drawback will be eliminated in the following basis that we shall build.

D. Vector multipoles (Refs. 1, 4, 5, and 10)

The study of electromagnetic radiation leads to the basis of the vector multipole (VM) defined by action of the operators u, $r\nabla$, and L on the scalar harmonics, and called "longitudinal electric, transverse electric, and transverse magnetic" respectively,[11]

$$\mathcal{E}_L{}^J_M(\mathbf{u}) = \mathbf{u}Y^J_M(\mathbf{u}), \tag{24a}$$

$$\mathcal{E}_T{}^J_M(\mathbf{u}) = [J(J+1)]^{-1/2} r\nabla Y^J_M(\mathbf{u}), \tag{24b}$$

$$\mathcal{M}_T{}^J_M(\mathbf{u}) = [J(J+1)]^{-1/2} \mathbf{L} Y^J_M(\mathbf{u}). \tag{24c}$$

By construction, the VM are orthonormalized vector functions for the scalar product (21) of the space $\mathcal{L}^2_1(\mathcal{S}^2)$, they are pairwise orthogonal vectors in $\mathcal{E}^3_c$, and they have a well-defined parity. Their names follow from the geometrical properties of the VM in $\mathcal{E}^3_c$ and from their parities. A multipole is "electric" (resp. "magnetic") if its intrinsic parity is $(-1)^{L+1}$[resp. $(-1)^L$] and it is "longitudinal" (resp. "transverse") if it is proportional (resp. perpendicular) to the vector u. In other words, the longitudinal VM is orthogonal to the sphere $\mathcal{S}^2$ while the transverse VM are tangent to this sphere.

An interesting remark is that the 3 VM are orthonormal combinations of VSH with the same parity, the coefficients being CG coefficients, cf., Appendix Eqs. (B1), (B2), (B3). This suggests the new notation $\mathbf{X}_\mu{}^J_M$ for the VM, and the new definition[12]

$$\mathbf{X}_{0M}^J(\mathbf{u}) \equiv \mathcal{E}_L{}^J_M(\mathbf{u}) = \sum_l (\hat{l}/\hat{J})\langle 10l0|J0\rangle \mathbf{Y}^{lJ}_M(\mathbf{u}), \tag{25a}$$

$$\mathbf{X}_{-1M}^J(\mathbf{u}) \equiv \mathcal{E}_T{}^J_M(\mathbf{u}) = \sum_l \{[1-(-1)^{l+J}]/\sqrt{2}\}(\hat{l}/\hat{J}) \times\langle 11l0|J1\rangle \mathbf{Y}^{lJ}_M(\mathbf{u}), \tag{25b}$$

$$\mathbf{X}_{+1M}^J(\mathbf{u}) \equiv \mathcal{M}_T{}^J_M(\mathbf{u}) = \sum_l \{[1+(-1)^{l+J}]/\sqrt{2}\}(\hat{l}/\hat{J}) \times\langle 11l0|J1\rangle \mathbf{Y}^{lJ}_M(\mathbf{u}). \tag{25c}$$

With this notation, the properties of the VM are easily written down, e.g., the orthogonality in $\mathcal{E}^3_c$ reads

$$\mathbf{X}_\mu{}^J_M \cdot \mathbf{X}_{\mu'}{}^J_M = \delta_{\mu\mu'}\epsilon_\mu (4\pi)^{-1/2}\sum_k (\hat{J}^2/\hat{k}) \times\langle J\mu J-\mu|k0\rangle\langle JMJM|ku\rangle Y^k_n(\mathbf{u}), \tag{26}$$

where $\epsilon_+ = \epsilon_0 = +1$ and $\epsilon_- = -1$. By using analytic expressions of the CG coefficients, the change of functions from the VSH to the VM can be written as the rotation

$$\begin{pmatrix}\mathbf{X}_{+1M}^J\\ \mathbf{X}_{0M}^J\\ \mathbf{X}_{-1M}^J\end{pmatrix} = \begin{pmatrix}0 & 1 & 0\\ -\sqrt{J+1}/\hat{J} & 0 & \sqrt{J}/\hat{J}\\ \sqrt{J}/\hat{J} & 0 & \sqrt{J+1}/\hat{J}\end{pmatrix}\begin{pmatrix}\mathbf{Y}^{J+1J}_M\\ \mathbf{Y}^{JJ}_M\\ \mathbf{Y}^{J-1J}_M\end{pmatrix}. \tag{27}$$

Under a rotation, the VM transform according to a representation D^J, i.e., they are eigenvectors of the operators $\mathbf{J}^2$ and J_3 with the eigenvalues $J(J+1)$ and M. They are also eigenfunctions of $\mathbf{S}^2$ but not of $\mathbf{L}^2$. Let $S_u \equiv i\mathbf{u}\times$ the component of the spin one, operator S along u. Then the VM satisfy

$$S_u \mathbf{X}_{0M}^J(\mathbf{u}) = 0, \tag{28a}$$

$$S_u[\mathbf{X}_{+1M}^J(\mathbf{u}) \pm \mathbf{X}_{-1M}^J(\mathbf{u})] = \pm[\mathbf{X}_{+1M}^J(\mathbf{u}) \pm \mathbf{X}_{-1M}^J(\mathbf{u})], \tag{28b}$$

i.e., $\mathbf{X}_0$ is eigenvector of S_u with the eigenvalue 0 while $X_\pm$ are orthogonal combinations of the two eigenvectors of S_u associated to the eigenvalues +1 and −1.

3. SECOND-ORDER TENSOR FIELDS

A. Second-order tensor spherical basis (Refs. 3 and 10)

The spherical basis of the space $\mathcal{E}^3_c\otimes\mathcal{E}^3_c$ is composed of the nine tensors $\mathbf{t}^j_n (j=0,1,2;\ n=-j,\ldots,+j)$ built by reducing the tensor product of the basis vectors with CG coefficients

$$\mathbf{t}^j_n = \langle 1m\,1m'|jn\rangle \mathbf{e}_m \otimes \mathbf{e}_{m'}. \tag{29}$$

For given j, they transform according to a representation D^j in a rotation R of the basis $\{\mathbf{e}\}$

$$R\otimes R\mathbf{t}^j_n = \mathbf{t}^j_{n'} D^j(R)^{n'}_n. \tag{30}$$

This transformation law implies that the tensors $\mathbf{t}^j_n$ describe the states of a spin j system. By using explicit values of CG coefficients, Eq. (29) gives

$$\begin{aligned}
\mathbf{t}^0_0 &= (1/\sqrt{3})(\mathbf{e}_+\otimes\mathbf{e}_- + \mathbf{e}_-\otimes\mathbf{e}_+ - \mathbf{e}_0\otimes\mathbf{e}_0),\\
\mathbf{t}^1_0 &= (1/\sqrt{2})(\mathbf{e}_+\otimes\mathbf{e}_- - \mathbf{e}_-\otimes\mathbf{e}_+),\\
\mathbf{t}^1_{\pm1} &= (\pm 1/\sqrt{2})(\mathbf{e}_\pm\otimes\mathbf{e}_0 - \mathbf{e}_0\otimes\mathbf{e}_\pm),\\
\mathbf{t}^2_0 &= (1/\sqrt{6})(\mathbf{e}_+\otimes\mathbf{e}_- + \mathbf{e}_-\otimes\mathbf{e}_+ + 2\mathbf{e}_0\otimes\mathbf{e}_0),\\
\mathbf{t}^2_{\pm1} &= (1/\sqrt{2})(\mathbf{e}_\pm\otimes\mathbf{e}_0 + \mathbf{e}_0\otimes\mathbf{e}_\pm),\\
\mathbf{t}^2_{\pm2} &= \mathbf{e}_\pm\otimes\mathbf{e}_\pm.
\end{aligned} \tag{31}$$

Let us study some properties of these tensors. They satisfy the identity

$$(\mathbf{t}^j_n)^* = (-1)^{j+n}\mathbf{t}^j_{-n} \tag{32}$$

and the orthonormality relation

$$\mathbf{t}^{j*}_n : \mathbf{t}^{j'}_{n'} = \delta_{jj'}\delta_{nn'}. \tag{33}$$

The tensors $\mathbf{t}^0_0$ and $\mathbf{t}^2_n$ are symmetric, while the tensors $\mathbf{t}^1_n$ are antisymmetric and dual of the vectors of the spherical basis, see Eq. (9), $\mathbf{t}^1_n = (1/i\sqrt{2})\mathbf{e}_n\times$. All tensors but $\mathbf{t}^0_0$ have a vanishing trace

$$\mathrm{tr}(\mathbf{t}^j_n) \equiv \delta : \mathbf{t}^j_n = -\sqrt{3}\delta_{j0}\delta_{n0}, \tag{34}$$

and the closure relation (3) implies that $\mathbf{t}^0_0 = -(1/\sqrt{3})\delta$.

Note that the saturation of a tensor $\mathbf{t}^j_n$ by $\mathbf{u}\otimes\mathbf{u}$ vanishes for $j=1$ and is proportional to the spherical harmonics $Y^j_n(\mathbf{u})$ for $j=0, 2$

$$\mathbf{t}^j_n : (\mathbf{u}\otimes\mathbf{u}) = (\sqrt{4\pi}/\hat{j})\langle 1010|j0\rangle Y^j_n(\mathbf{u}). \tag{35}$$

B. Second-order tensor spherical harmonics (Refs. 2 and 3)

The space $\mathcal{L}^2_2(\mathcal{S}^2)$ of complex second-order tensor fields on the sphere $\mathcal{S}^2$, with integrable modulus squared, is a Hilbert space for the scalar product

$$\langle \mathbf{f}_1, \mathbf{f}_2\rangle = \int_{\mathcal{S}^2} \mathbf{f}_1(\mathbf{u})^* : \mathbf{f}_2(\mathbf{u})\,d\mathbf{u}. \tag{36}$$

As for the vector fields, the second-order tensor spherical harmonics (TSH) are built by coupling the tensors of the spherical basis and the spherical harmonics through CG coefficients

$$\mathbf{Y}^{ljJ}_M(\mathbf{u}) = \langle lmjn | JM\rangle Y^l_m(\mathbf{u})\mathbf{t}^j_n. \tag{37}$$

By construction, the TSH transform according to a representation D^J in a rotation, and they form an orthonormal basis of $\mathcal{L}^2_2(\mathcal{S}^2)$

$$\langle \mathbf{Y}^{ljJ}_M, \mathbf{Y}^{l'j'J'}_{M'}\rangle = \delta_{ll'}\delta_{jj'}\delta_{JJ'}\delta_{MM'} . \tag{38}$$

The TSH satisfy the identity

$$\mathbf{Y}^{ljJ}_M(\mathbf{u})^* = (-1)^{l+j+J+M}\mathbf{Y}^{ljJ}_{-M}(\mathbf{u}) \tag{39}$$

and they have the parity

$$\mathbf{Y}^{ljJ}_M(-\mathbf{u}) = (-1)^l\mathbf{Y}^{ljJ}_M(\mathbf{u}). \tag{40}$$

For fixed values of $J \geq 2$ and M, there are nine tensorial harmonics which can be divided in three classes according to the value of j.

(i) $j=0$. There is one TSH for $l=J$ which is proportional to the identity tensor

$$\mathbf{Y}^{J0J}_M(\mathbf{u}) = -(1/\sqrt{3})\,Y^L_M(\mathbf{u})\,\delta. \tag{41}$$

(ii) $j=1$. There are three TSH for $l=J$, $J\pm 1$ which are antisymmetric and dual tensors of the VSH

$$\mathbf{Y}^{l1J}_M(\mathbf{u}) = (1/i\sqrt{2})\mathbf{Y}^{lJ}_M(\mathbf{u})\times. \tag{42}$$

(iii) $j=2$. There are five TSH for $l=J$, $J\pm 1$, $J\pm 2$ which are symmetric and have a vanishing trace.

By action of the operators $\mathbf{u}\otimes$, $r\nabla\otimes$, $\mathbf{L}\otimes$ on the VSH, we get linear combinations of TSH involving CG coefficients and $6j$ symbols. These expressions are given in Appendix B, Eqs. (B4), (B5), (B6).

To study the geometric properties of these nine TSH consider their scalar product in the space $\mathcal{E}^3_c\otimes\mathcal{E}^3_c$, given in Appendix A, Eq. (A3). This equation shows that the TSH with different j or opposite parity are orthogonal, but the TSH with the same j and parity are not orthogonal. Therefore in the following subsection we define the basis of tensor multipoles which form an orthonormal set of $\mathcal{L}^2_2(\mathcal{S}^2)$ and also an orthogonal set in $\mathcal{E}^3_c\otimes\mathcal{E}^3_c$.

C. Second-order tensor multipoles

To define a set of tensor multipole (TM) which form an orthogonal set in the space $\mathcal{E}^3_c\otimes\mathcal{E}^3_c$, Zerilli[2] follows the same procedure as for the vector fields. He applies tensor products of the operators $\mathbf{u}$, ∇, $\mathbf{L}$ on the spherical harmonics, and he takes appropriate combinations to obtain an orthogonal set. We shall follow another way; the generalization of Eqs. (25) for second-order tensors provides us with a method to build TM as orthonormal combinations of the TSH with the same j and parity. Our TM are denoted by $\mathbf{X}^{jJ}_{\mu M}(\mathbf{u})$, ($j=0,1,2$; $\mu = -j,\ldots,+j$) and they are defined by

$$\mathbf{X}^{jJ}_{0M}(\mathbf{u}) = \sum_l (\hat{l}/\hat{j})\langle j0l0|J0\rangle\mathbf{Y}^{ljJ}_M(\mathbf{u}), \tag{43a}$$

$$\mathbf{X}^{jJ}_{\pm\mu M}(\mathbf{u}) = \sum_l \{[1\pm(-1)^{l+J}]/\sqrt{2}\}\langle j\mu l0|J\mu\rangle\mathbf{Y}^{ljJ}_M(\mathbf{u}), \quad \mu > 0. \tag{43b}$$

The orthogonality of the CG coefficients allows us to show that the change from the TSH to the TM is an orthogonal transformation. This transformation is made explicit with analytic values of the CG coefficients in Table I.

By construction the TM transform according to a representation D^J under a rotation, and they form an orthonormal basis of the space $\mathcal{L}^2_2(\mathcal{S}^2)$

$$\langle \mathbf{X}^{jJ}_{\mu M}, \mathbf{X}^{j'J'}_{\mu'M'}\rangle = \delta_{jj'}\delta_{\mu\mu'}\delta_{JJ'}\delta_{MM'}, \tag{44a}$$

TABLE I. Relation between the tensor multipoles and the tensor spherical harmonics. (The indices J and M and the $\mathbf{u}$ dependence are omitted.)

$$\mathbf{X}^0_0 = \mathbf{Y}^{J0}, \qquad \mathbf{X}^1_{+1} = \mathbf{Y}^{J1}$$

$$\begin{pmatrix} \mathbf{X}^1_0 \\ \mathbf{X}^1_{-1} \end{pmatrix} = \begin{pmatrix} -\left[\frac{J+1}{2J+1}\right]^{1/2} & \left[\frac{J}{2J+1}\right]^{1/2} \\ \left[\frac{J}{2J+1}\right]^{1/2} & \left[\frac{J+1}{2J+1}\right]^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{Y}^{J+11} \\ \mathbf{Y}^{J-11} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{X}^2_{-1} \\ \mathbf{X}^2_{-2} \end{pmatrix} = \begin{pmatrix} -\left[\frac{J+2}{2J+2}\right]^{1/2} & \left[\frac{J-1}{2J+1}\right]^{1/2} \\ \left[\frac{J-1}{2J+1}\right]^{1/2} & \left[\frac{J+2}{2J+1}\right]^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{Y}^{J+12} \\ \mathbf{Y}^{J-12} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{X}^2_{+2} \\ \mathbf{X}^2_{+1} \\ \mathbf{X}^2_0 \end{pmatrix} = \begin{pmatrix} \left[\frac{J(J-1)}{2(2J+1)(2J+3)}\right]^{1/2} & \left[\frac{3(J-1)(J+2)}{(2J-1)(2J+3)}\right]^{1/2} & \left[\frac{(J+1)(J+2)}{2(2J-1)(2J+3)}\right]^{1/2} \\ -\left[\frac{2J(J+2)}{(2J+1)(2J+3)}\right]^{1/2} & -\left[\frac{3}{(2J-1)(2J+3)}\right]^{1/2} & \left[\frac{2(J+1)(J-1)}{(2J-1)(2J+3)}\right]^{1/2} \\ \left[\frac{3(J+1)(J+2)}{2(2J+1)(2J+3)}\right]^{1/2} & -\left[\frac{J(J+1)}{(2J-1)(2J+3)}\right]^{1/2} & \left[\frac{3J(J-1)}{2(2J-1)(2J+3)}\right]^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{Y}^{J+22} \\ \mathbf{Y}^{J2} \\ \mathbf{Y}^{J-22} \end{pmatrix}$$

TABLE II. Construction of the tensor multipoles by action of the operators $\mathbf{u}\otimes$, $r\nabla\otimes$, and $\mathbf{L}\otimes$ on the vector multipoles. (The indices J and M and the $\mathbf{u}$ dependence are omitted.)

$\mathbf{X}_0^0=\frac{1}{\sqrt{3}}\{\mathbf{u}\otimes\mathbf{X}_0-[J(J+1)]^{-1/2}(\mathbf{L}\otimes\mathbf{X}_{+1}-r\nabla\otimes\mathbf{X}_{-1})\}$
$\mathbf{X}_0^2=-\sqrt{\tfrac{2}{3}}\{\mathbf{u}\otimes\mathbf{X}_0+\tfrac{1}{2}[J(J+1)]^{-1/2}(\mathbf{L}\otimes\mathbf{X}_{+1}-r\nabla\otimes\mathbf{X}_{-1}\}$
$\mathbf{X}_0^1=\sqrt{2}\,[J(J+1)]^{-1/2}[(r\nabla-\mathbf{u})\otimes\mathbf{X}_{+1}]^a$
$\mathbf{X}_{-2}^2=\sqrt{2}\,[(J-1)(J+2)]^{-1/2}[(r\nabla+\mathbf{u})\otimes\mathbf{X}_{-1}]^s$
$\mathbf{X}_{+2}^2=[2(J-1)(J+2)]^{-1/2}[\mathbf{L}\otimes\mathbf{X}_{+1}+(r\nabla+2\mathbf{u})\otimes\mathbf{X}_{-1}]^s$
$\mathbf{X}_{-1}^2=\sqrt{2}\,[\mathbf{u}\otimes\mathbf{X}_{+1}]^s$, $\mathbf{X}_{-1}^1=\sqrt{2}\,[\mathbf{u}\otimes\mathbf{X}_{+1}]^a$
$\mathbf{X}_{+1}^2=\sqrt{2}\,[\mathbf{u}\otimes\mathbf{X}_{-1}]^s$, $\mathbf{X}_{+1}^1=\sqrt{2}\,[\mathbf{u}\otimes\mathbf{X}_{-1}]^a$
$[\ ]^s$ and $[\ ]^a$ mean symmetric or antisymmetric part.

while for fixed values of J and M, they are orthogonal in the space $\mathcal{E}_c^3\otimes\mathcal{E}_c^3$

$$\mathbf{X}_{\mu M}^{jJ}(\mathbf{u}):\mathbf{X}_{\mu' M}^{j' J}(\mathbf{u})$$

$$=\delta_{jj'}\delta_{\mu\mu'}\epsilon_\mu^j(4\pi)^{-1/2}\sum_k(\hat{J}^2/\hat{k})\langle J\mu J-\mu|k0\rangle \times\langle JMJM|kn\rangle Y_n^k(\mathbf{u}), \tag{44b}$$

where $\epsilon_\mu^j=(-1)^j$ for $\mu=0$, $(-1)^\mu$ for $\mu>0$, and $(-1)^{\mu+1}$ for $\mu<0$. Like the TSH, the TM have a well defined parity

$$\mathbf{X}_0^j(-\mathbf{u})_M^J=(-1)^{J+j}\mathbf{X}_0^j(\mathbf{u})_M^J, \tag{45a}$$

$$\mathbf{X}_{\pm\mu}^j(-\mathbf{u})_M^J=\pm(-1)^J\mathbf{X}_{\pm\mu}^j(\mathbf{u})_M^J. \tag{45b}$$

We note that the product of a TM by the vector $\mathbf{u}$ is either vanishing or proportional to one VM

$$\mathbf{X}_0^j(\mathbf{u})_M^J\cdot\mathbf{u}=\langle 1010|j0\rangle\mathbf{X}_0(\mathbf{u})_M^J, \tag{46a}$$

$$\mathbf{X}_{\pm\mu}^j(\mathbf{u})_M^J\cdot\mathbf{u}=\pm\langle 1\mu 10|j\mu\rangle\mathbf{X}_{\mp}(\mathbf{u})_M^J. \tag{46b}$$

The TM can be built by action of the operators $\mathbf{u}\otimes$, $r\nabla\otimes$, and $\mathbf{L}\otimes$ on the VM, by using the identities (B4), (B5), (B6) and the properties of the $6j$ symbols. The result is given in Table II. Then, the definitions of the VM and the identity

$$\{\mathbf{u}\otimes\mathbf{u}+[J(J+1)]^{-1}(\mathbf{L}\otimes\mathbf{L}-r\nabla\otimes r\nabla)\}Y_M^J(\mathbf{u})=\boldsymbol{\delta}Y_M^J(\mathbf{u}) \tag{47}$$

allow us to deduce the TM from the spherical harmonics by action of tensor products of the vector operators or by action of the identity tensor, see Table III. With this we can relate our TM to those defined by other authors. Zerilli[2] defines the following symmetric TM

$$\mathbf{a}_{JM}=-(1/\sqrt{3})\mathbf{X}_{0M}^{0J}+\sqrt{(2/3)}\,\mathbf{X}_{0M}^{2J}, \tag{48a}$$

$$\mathbf{b}_{JM}=\mathbf{X}_{1M}^{2J},\qquad \mathbf{c}_{JM}=\mathbf{X}_{-1M}^{2J}, \tag{48b}$$

$$\mathbf{d}_{JM}=\mathbf{X}_{-2M}^{2J},\qquad \mathbf{f}_{JM}=\mathbf{X}_{+2M}^{2J}, \tag{48c}$$

$$\mathbf{h}_{JM}=\sqrt{(2/3)}\mathbf{X}_{0M}^{0L}+(1/\sqrt{3})\mathbf{X}_{0M}^{2J}, \tag{48d}$$

while Regge and Wheeler[6] also consider $\mathbf{a}_{JM}$, $\mathbf{b}_{JM}$, $\mathbf{c}_{JM}$, $\mathbf{d}_{JM}$ and two others TM $\mathbf{e}_{JM}$ and $\mathbf{g}_{JM}$ which are linear combinations of $\mathbf{f}_{JM}$ and $\mathbf{h}_{JM}$:

$$\mathbf{e}_{JM}=[J(J+1)/2]^{1/2}\{[(J-1)(J+2)]^{1/2}\mathbf{f}_{JM} -[J(J+1)]^{1/2}\mathbf{h}_{JM}\}, \tag{49a}$$

$$\mathbf{g}_{JM}=[J(J+1)/2]^{1/2}\{[(J-1)(J+2)]^{1/2}\mathbf{f}_{JM} +[J(J+1)]^{1/2}\mathbf{h}_{JM}\}. \tag{49b}$$

The TM set of Regge and Wheeler does not form an orthonormal basis of the space $\mathcal{L}_2^2(\mathcal{S}^2)$. The set of Zerilli does, but for fixed values of J and M it is not orthogonal in the space $\mathcal{E}_c^3\otimes\mathcal{E}_c^3$. Furthermore the TM $\mathbf{a}_{JM}$, $\mathbf{h}_{JM}$ (and $\mathbf{e}_{JM}$, $\mathbf{g}_{JM}$) have a nonvanishing trace whereas, only one of our TM, $\mathbf{X}_{0M}^{0J}$ has trace.

By analogy with the electromagnetic appellation of the VM, we can characterize each TM by its parity and its geometrical properties in $\mathcal{E}_c^3\otimes\mathcal{E}_c^3$. For this we use the convention adopted by Thorne and Campolattaro[13] and by Zerilli.[14] The TM are denoted by a symbol $\mathcal{M}$ (magnetic) or $\mathcal{E}$ (electric) for parity $(-1)^{J+1}$ or $(-1)^J$, respectively,[15] with an upper index s (symmetric) or a (antisymmetric), and a lower index $\mathcal{S}$ (scalar) or a couple of $\mathcal{T}$ (transverse) or $\mathcal{L}$ (longitudinal) in the following geometrical configurations

$\mathbf{X}_{\mathcal{S}}$ for $\mathbf{X}:\delta\neq 0$,

$\mathbf{X}_{\mathcal{L}}$ for $\mathbf{X}:(\mathbf{u}\otimes\mathbf{u})\neq 0$ and $\mathbf{X}:\delta=0$,

$\mathbf{X}_{\mathcal{LT}}$ for $\mathbf{X}\cdot\mathbf{u}\neq 0$ and $\mathbf{X}:(\mathbf{u}\otimes\mathbf{u})=0$,

$\mathbf{X}_{\mathcal{TT}}$ for $\mathbf{X}\cdot\mathbf{u}=0$.

With these conventions, we get the following new notations (the fixed values of J and M are omitted):

(i) $j=0$. There is one TM, with parity $(-1)^J$, proportional to δ

$$\mathcal{E}_{\mathcal{S}}\equiv\mathbf{X}_0^0=-(1/\sqrt{3})\delta Y. \tag{50}$$

(ii) $j=1$. There are three antisymmetric TM, dual tensors of the VM

$$\mathcal{E}^a{}_{\mathcal{LT}}\equiv\mathbf{X}_{+1}^1=(1/i\sqrt{2})\mathcal{M}_{\mathcal{T}}\times, \tag{51a}$$

$$\mathcal{M}^a{}_{\mathcal{TT}}\equiv\mathbf{X}_0^1=(1/i\sqrt{2})\mathcal{E}_{\mathcal{L}}\times, \tag{51b}$$

$$\mathcal{M}^a{}_{\mathcal{LT}}\equiv\mathbf{X}_{-1}^1=(1/i\sqrt{2})\mathcal{E}_{\mathcal{T}}\times. \tag{51c}$$

(iii) $j=2$. There are five symmetric traceless TM

$$\mathcal{E}^s{}_{\mathcal{TT}}\equiv\mathbf{X}_{+2}^2,\quad \mathcal{E}^s{}_{\mathcal{LT}}\equiv\mathbf{X}_{+1}^2,\quad \mathcal{E}^s{}_{\mathcal{LL}}\equiv\mathbf{X}_0^2,$$

$$\mathcal{M}^s{}_{\mathcal{LT}}\equiv\mathbf{X}_{-1}^2,\quad \mathcal{M}^s{}_{\mathcal{TT}}\equiv\mathbf{X}_{-2}^2. \tag{52}$$

TABLE III. Construction of the tensor multipoles by action of the tensor products of operators $\mathbf{u}$, $r\nabla$, $\mathbf{L}$ and by action of the identity tensor $\boldsymbol{\delta}$ on the spherical harmonics. (The indices J and M and the $\mathbf{u}$ dependence are omitted.)

$\mathbf{X}_0^0=-(1/\sqrt{3})\boldsymbol{\delta}Y$, $\mathbf{X}_0^2=\sqrt{\tfrac{3}{2}}\,(\mathbf{u}\otimes\mathbf{u}-\boldsymbol{\delta})Y$
$\mathbf{X}_0^1=\sqrt{2}\,[J(J+1)]^{-1}[(r\nabla-\mathbf{u})\otimes\mathbf{L}]^sY$
$\mathbf{X}_{-2}^2=\sqrt{2}\,[(J-1)J(J+1)(J+2)]^{-1/2}[(r\nabla+\mathbf{u})\otimes\mathbf{L}]^sY$
$\mathbf{X}_{+2}^2=[2(J-1)J(J+1)(J+2)]^{-1/2}[\mathbf{L}\otimes\mathbf{L}+(r\nabla+2\mathbf{u})\otimes r\nabla]^sY$
$\mathbf{X}_{-1}^2=\sqrt{2}\,[J(J+1)]^{-1/2}[\mathbf{u}\otimes\mathbf{L}]^sY$, $\mathbf{X}_{-1}^1=\sqrt{2}\,[J(J+1)]^{-1/2}[\mathbf{u}\otimes\mathbf{L}]^aY$
$\mathbf{X}_{+1}^2=\sqrt{2}\,[J(J+1)]^{-1/2}[\mathbf{u}\otimes r\nabla]^sY$, $\mathbf{X}_{+1}^1=\sqrt{2}\,[J(J+1)]^{-1/2}[\mathbf{u}\otimes r\nabla]^aY$
$[\ ]^s$ and $[\ ]^a$ mean symmetric or antisymmetric part.

4. ARBITRARY ORDER TENSOR FIELDS

A. Tensor spherical basis

The tensors of the spherical basis of the space $(\mathcal{E}_c^3)^{\otimes r}$ are built by an iterating process, from the vector $\mathbf{e}_n$. We have already seen the second-order tensors. The tensors of order $r > 2$ are obtained from the $(r-1)$th-order tensors and the vectors $\mathbf{e}_n$ by means of CG coefficients

$$\mathbf{t}_m^{j_r \cdots j_2} = \langle j_{r-1} m' 1 n | j_r m \rangle \mathbf{t}_{m'}^{j_{r-1} \cdots j_2} \otimes \mathbf{e}_n. \tag{53}$$

These tensors satisfy the identity

$$\mathbf{t}_m^{j_r \cdots *} = (-1)^{r+j_r+m} \mathbf{t}_{-m}^{j_r} \tag{54}$$

and the orthonormality relations

$$\mathbf{t}_m^{j_r \cdots *} (\overset{r}{.}) \mathbf{t}_{m'}^{j_r' \cdots} = \left(\prod_{i=2}^{r} \delta_{j_i j_i'} \right) \delta_{mm'}. \tag{55}$$

The contraction of the tensor product $\otimes^r \mathbf{u}$ with a basis tensor is proportional to a spherical harmonics

$$\overset{r}{\otimes} \mathbf{u} (\overset{r}{.}) \mathbf{t}_m^{j_r \cdots} = (\sqrt{4\pi} / \hat{j}_r) \left(\prod_{k=2}^{r} \langle j_{k-1} 010 | j_k 0 \rangle \right) Y_m^{j_r}(\mathbf{u}). \tag{56}$$

The tensors obtained for the maximal couplings (i.e., $j_2 = 2, \ldots, j_k = k, \ldots, j_r = r$) are simply denoted by $\mathbf{t}_m^r$. They are completely symmetric and have vanishing trace

$$(\mathbf{t}_m^r)^{i_1 \cdots i_k \cdots i_l \cdots i_r} = (\mathbf{t}_m^r)^{i_1 \cdots i_l \cdots i_k \cdots i_r}, \tag{57}$$

$$(\mathbf{t}_m^r)^{i_1 \cdots i_k \cdots i_l \cdots i_r} \delta_{i_k i_l} = 0. \tag{58}$$

Furthermore, for these tensors, the CG coefficients product in Eq. (56) can be calculated and one gets the identity

$$Y_m^l(\mathbf{u}) = \left[\frac{(2l+1)!!}{4\pi l!} \right]^{1/2} \overset{l}{\otimes} \mathbf{u} (\overset{l}{.}) \mathbf{t}_m^l. \tag{59}$$

B. Tensor spherical harmonics

The space $\mathcal{L}_r^2(\mathcal{S}^2)$ of complex rth-order tensor fields on the sphere $\mathcal{S}^2$, with integrable modulus squared, is a Hilbert space for the scalar product

$$\langle \mathbf{f}_1, \mathbf{f}_2 \rangle = \int_{\mathcal{S}_2} \mathbf{f}_1(\mathbf{u}) (\overset{r}{.}) \mathbf{f}_2(\mathbf{u})\, d\mathbf{u}. \tag{60}$$

As for lower-order fields, the rth-order TSH are built by coupling basis tensors and spherical harmonics through CG coefficients

$$\mathbf{Y}^{l j_r \cdots J}_{M}(\mathbf{u}) = \langle l m j_r n | JM \rangle Y_m^l(\mathbf{u}) \mathbf{t}_n^{j_r \cdots}, \tag{61}$$

so that they transform according to a representation D^J in a rotation, and they form an orthonormal basis of $\mathcal{L}_r^2(\mathcal{S}^2)$

$$\langle \mathbf{Y}^{l j_r \cdots J}_{M}, \mathbf{Y}^{l' j_r' \cdots J'}_{M'} \rangle = \delta_{ll'} \left(\prod_{i=2}^{r} \delta_{j_i j_i'} \right) \delta_{JJ'} \delta_{MM'}. \tag{62}$$

Besides they satisfy the identities

$$\mathbf{Y}^{l j_r \cdots J}_{M}(\mathbf{u})^* = (-1)^{r+l+j_r+L+M} \mathbf{Y}^{l j_r \cdots J}_{-M}(\mathbf{u}), \tag{63}$$

$$\mathbf{Y}^{l j_r \cdots J}_{M}(-\mathbf{u}) = (-1)^l \mathbf{Y}^{l j_r \cdots J}_{M}(\mathbf{u}). \tag{64}$$

The contraction formula of two TM is quite similar to that of the second-order TM, see Ref. 12, Eq. (3–65). It shows that for fixed values of JM, the TSH do not form an orthogonal set in $(\mathcal{E}_c^3)^{\otimes r}$. To remedy this we shall now define the rth-order TM.

C. Tensor multipoles

The rth-order TM are built in the same way as the second-order TM, by the orthogonal transformation

$$\mathbf{X}_\mu^{j_r \cdots J}{}_M = \sum_l M(j_r, J)_{\mu l} \mathbf{Y}^{l j_r \cdots J}_{M}, \tag{65}$$

where the matrix elements $M(j_r, J)_{\mu l}$ are defined by

$$M(j_r, J)_{0l} = (\hat{l} / \hat{J}) \langle j_r 0 l 0 | J 0 \rangle, \tag{66a}$$

$$M(j_r, J)_{\pm\mu l} = \{[1 \pm (-1)^{l+J}] / \sqrt{2}\} (\hat{l} / \hat{J}) \langle j_r \mu l 0 | J \mu \rangle, \quad \mu > 0. \tag{66b}$$

The orthogonality properties of the CG coefficients imply that the matrices $M(j_r, J)$ are orthogonal. Furthermore, each matrix can be split into a direct sum of two orthogonal submatrices

$$M(j_r, J) = M_+(j_r, J) \oplus M_-(j_r, J) \tag{67}$$

defined according to the parity of j_r as follows.

$$j_r \text{ even} \begin{cases} M_+(j_r, J) : \mu = 0, 1, \ldots, j_r; \\ l = |J - j_r|, |J - j_r| + 2, \ldots, J + j_r, \\ M_-(j_r, J) : \mu = -1, -2, \ldots, -j_r; \\ l = |J - j_r| + 1, |J - j_r| + 3, \ldots, J + j_r - 1, \end{cases} \tag{68a}$$

$$j_r \text{ odd} \begin{cases} M_+(j_r, J) : \mu = 1, 2, \ldots, j_r; \\ l = |J - j_r| + 1, |J - j_r| + 3, \ldots, J + j_r - 1, \\ M_-(j_r, J) : \mu = 0, -1, \ldots, -j_r; \\ l = |J - j_r|, |J - j_r| + 2, \ldots, J + j_r. \end{cases} \tag{68b}$$

By construction, the TM transform according to one irreducible representation under a rotation, they form an orthonormal basis of $\mathcal{L}_r^2(\mathcal{S}^2)$

$$\langle \mathbf{X}_\mu^{j_r \cdots J}{}_M, \mathbf{X}_{\mu'}^{j_r' \cdots J'}{}_{M'} \rangle = \delta_{\mu\mu'} \left(\prod_{i=2}^{r} \delta_{j_i j_i'} \right) \delta_{JJ'} \delta_{MM'} \tag{69}$$

and for fixed J and M, they are orthogonal in $(\mathcal{E}_c^3)^{\otimes r}$

$$\mathbf{X}_\mu^{j_r \cdots J}{}_M \cdot \mathbf{X}_{\mu'}^{j_r' \cdots J}{}_M = \left(\prod_{j=2}^{r} \delta_{j_i j_i'} \right) \delta_{\mu\mu'} \epsilon_\mu^{j_r} (4\pi)^{-1/2} \sum_k (\hat{J}^2 / \hat{k}) \times \langle J \mu J - \mu | k 0 \rangle \langle JMJM | kn \rangle Y_n^{k*}(\mathbf{u}), \tag{70}$$

where $\epsilon_\mu^{j_r} = (-1)^{j_r}$ for $\mu = 0$, $(-1)^\mu$ for $\mu > 0$ and $(-1)^{\mu+1}$ for $\mu < 0$. Another interesting geometrical property of the TM is that the scalar product by the vector $\mathbf{u}$ of a rth-order TM is either vanishing or proportional to a $(r-1)$th-order multipole

$$\mathbf{X}_0^{j_r \cdots J}{}_M \cdot \mathbf{u} = \langle 1 0 j_{r-1} 0 | j_r 0 \rangle \mathbf{X}_0^{j_{r-1} \cdots J}{}_M(\mathbf{u}), \tag{71a}$$

$$\mathbf{X}_{\pm\mu}^{j_r \cdots J}{}_M \cdot \mathbf{u} = \pm \langle 1 0 j_{r-1} \mu | j_r \mu \rangle \mathbf{X}_{\mp\mu}^{j_{r-1} \cdots J}{}_M(\mathbf{u}), \quad \mu > 0. \tag{71b}$$

The TM have the following intrinsic parities

$$\mathbf{X}_0^{j_r \cdots J}{}_M(-\mathbf{u}) = (-1)^{J+j_r} \mathbf{X}_0^{j_r \cdots J}{}_M(\mathbf{u}), \tag{72a}$$

$$\mathbf{X}_{\pm\mu}^{j_r \cdots J}{}_M(-\mathbf{u}) = \pm(-1)^J \mathbf{X}_{\pm\mu}^{j_r \cdots J}{}_M(\mathbf{u}). \tag{72b}$$

By analogy with electromagnetism, we can call magnetic, the TM with intrinsic parity $(-1)^{J+r+1}$, and electric those with intrinsic parity $(-1)^{J+r}$.

5. CONCLUSION

In this article we have considered bases in the space of Euclidean tensor fields on the unit sphere S^2. In the following paper we shall build bases in the space of Minkowski tensor fields on S^2. We shall generalize the concept of tensor spherical harmonics and tensor multipoles and study their transformation properties under rotations and Lorentz transformations.

APPENDICES

In these Appendices, we have gathered useful formulas concerning: (in Appendix A) the product of spherical harmonics and the contraction of vector and tensor spherical harmonics; (in Appendix B) the action of the vectorial operators $\mathbf{u}$, $r\nabla$, $\mathbf{L} = -i\mathbf{r}\times\nabla$ on the scalar and vector spherical harmonics. For the demonstration of these formulas and of some formulas of the main text, the reader is referred to the "Thèse de Doctorat" of one of the authors.[12]

APPENDIX A

$$Y_m^l Y_{m'}^{l'} = (4\pi)^{-1/2} \sum_k (\hat{l}\hat{l}'/\hat{k}) \langle l0l'0 | k0\rangle \langle lml'm' | kn\rangle Y_n^k, \tag{A1}$$

$$\mathbf{Y}^{lJ}_M \cdot \mathbf{Y}^{l'J'}_{M'} = (-1)^{L'+l'} (4\pi)^{-1/2} \sum_k (\hat{J}\hat{l}\hat{J}'\hat{l}'/\hat{k}) \times \langle l0l'0 | k0\rangle \begin{Bmatrix} k & l & l' \\ 1 & J' & J \end{Bmatrix} \langle JMJ'M' | kn\rangle Y_n^k, \tag{A2}$$

$$\mathbf{Y}^{ljJ}_M : \mathbf{Y}^{l'j'J'}_{M'} = \delta_{jj'} (-1)^{L'+l'+j'} (4\pi)^{-1/2} \sum_k (\hat{J}\hat{l}\hat{J}'\hat{l}'/\hat{k}) \times \langle l0l'0 | k0\rangle \begin{Bmatrix} k & l & l' \\ 1 & J' & J \end{Bmatrix} \langle JMJ'M' | kn\rangle Y_n^k. \tag{A3}$$

APPENDIX B

$$\mathbf{u}Y_M^J(\mathbf{u}) = \sum_l (\hat{J}/\hat{l}) \langle 1010 | J0\rangle \mathbf{Y}^{lJ}_M(\mathbf{u}), \tag{B1}$$

$$r\nabla Y_M^J(\mathbf{u}) = [J(J+1)]^{1/2} \sum_l \{[1-(-1)^{J+l}]/\sqrt{2}\} \times \langle 11l0 | J1\rangle \mathbf{Y}^{lJ}_M(\mathbf{u}), \tag{B2}$$

$$\mathbf{L}Y_M^J(\mathbf{u}) = [J(J+1)]^{1/2} \mathbf{Y}^{JJ}_M(\mathbf{u}) = [J(J+1)]^{1/2} \sum_l \{[1+(-1)^{J+l}]/\sqrt{2}\} \times \langle 11l0 | J1\rangle \mathbf{Y}^{lJ}_M(\mathbf{u}), \tag{B3}$$

$$\mathbf{u}\otimes \mathbf{Y}^{lJ}_M(\mathbf{u}) = (-1)^{l+J} \sum_{k,j} \hat{j}\hat{l} \langle 10l0 | k0\rangle \begin{Bmatrix} l & 1 & J \\ j & k & 1 \end{Bmatrix} \mathbf{Y}^{kjJ}_M(\mathbf{u}), \tag{B4}$$

$$r\nabla\otimes \mathbf{Y}^{lJ}_M(\mathbf{u}) = (-1)^{l+J} [l(l+1)]^{1/2} \sum_{k,j} \frac{(-1)^{l+k}-1}{\sqrt{2}} \hat{j}\hat{l} \times \langle 11l-1 | k0\rangle \begin{Bmatrix} l & 1 & J \\ j & k & 1 \end{Bmatrix} \mathbf{Y}^{kjJ}_M(\mathbf{u}), \tag{B5}$$

$$\mathbf{L}\otimes \mathbf{Y}^{lJ}_M(\mathbf{u}) = (-1)^{l+J} [l(l+1)]^{1/2} \sum_j \hat{j}\hat{l} \begin{Bmatrix} l & 1 & J \\ j & l & 1 \end{Bmatrix} \mathbf{Y}^{ljJ}_M(\mathbf{u}) = (-1)^{l+J} [l(l+1)]^{1/2} \sum_{k,j} \frac{(-1)^{l+k}+1}{\sqrt{2}} \hat{j}\hat{l} \times \langle 11l-1 | k0\rangle \begin{Bmatrix} l & 1 & J \\ j & k & 1 \end{Bmatrix} \mathbf{Y}^{kjJ}_M(\mathbf{u}). \tag{B6}$$

*Equipe de Recherche associée au C.N.R.S.

[1]J.M. Blatt and V.F. Weisskopf, *Theoretical Nuclear Physics* (Wiley, New York, 1952), Appendix.
[2]J. Mathews, J. Soc. Ind. Appl. Math. **10**, 768 (1962).
[3]F.J. Zerilli, J. Math. Phys. **11**, 2203 (1970).
[4]M.E. Rose, *Multipole Fields* (Wiley, New York, 1955).
[5]L. Landau and F. Lifschitz, *Relativistic Quantum Theory, Part I* (Pergamon Press, Oxford, 1971).
[6]T. Regge and J.A. Wheeler, Phys. Rev. **108**, 1063 (1967).
[7]E.P. Wigner, *Group Theory* (Academic Press, New York, 1959).
[8]M.E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957).
[9]A.R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton University Press, Princeton, New Jersey, 1957).
[10]D.M. Brink and G.R. Satchler, *Angular Momentum* (Clarendon Press, Oxford, 1962).
[11]This nomenclature should not be confused with that of electromagnetic radiation modes. For a precise justification, see the resolution of vectorial Helmholtz equation in Appendix C of Ref. 12.
[12]M. Daumens, Thèse de doctorat, Université de Bordeaux I, no 441 (1974). Note that the multipoles X_μ^j defined in the thesis differ from that of this paper by a factor $(-1)^{j+\mu}$.
[13]K. Thorne and A. Campolattaro, Astrophys. J. **149**, 591 (1967).
[14]F.J. Zerilli, Phys. Rev. D **2**, 2141 (1970); **9**, 860 (1974).
[15]This convention opposite to that of Mathews, Ref. 2, is more natural for gravitational radiation field since the "electric" gravitational multipoles couple to "electric" electromagnetic multipoles (and similarly for magnetic multipoles) via the Einstein-Maxwell equations, see Ref. 14.

Resistance inequalities for the isotropic Heisenberg ferromagnet*

Robert T. Powers†

Department of Mathematics, University of California, Berkeley, California 94720
(Received 21 November 1975)

Inequalities concerning the spin correlations of states of finite energy of the isotropic ferromagnetic Heisenberg model are proved. These inequalities estimate the spin correlation between two lattice sites in terms of the total energy and the electrical resistance between the lattice sites as calculated using the inverse of the coefficients occurring in the Heisenberg Hamiltonian. These resistance inequalities are combined with the resistance properties of a regular three-dimensional lattice to yield the result that all states of finite energy for the isotropic Heisenberg model in three dimensions have long range order.

INTRODUCTION

In this paper we prove inequalities concerning the spin correlations of states of finite energy of the isotropic ferromagnetic Heisenberg model. These inequalities estimate the spin correlation between two lattice sites in terms of the total energy and the electrical resistance between the lattice sites as calculated using the inverse of the coefficients occurring in the Heisenberg Hamiltonian as the resistances between neighboring lattice sites. These resistance inequalities are stated and proved in Sec. 3, Theorem 3.3.

One application of these inequalities is to the isotropic Heisenberg model in three dimensions. For a regular three-dimensional lattice the resistance between any two lattice sites is bounded by a constant R_∞, independent of the distance between the sites. The resistance properties of a regular three-dimensional lattice when combined with the resistance inequalities yield in Theorem 3.5 the result that all states of finite energy for the isotropic Heisenberg model in three dimensions have long range order.

In Sec. 1 of this paper we introduce notation for describing the isotropic Heisenberg model in a C^*-algebraic setting. In Sec. 2 we discuss the ways of calculating and estimating the resistance of electrical networks composed solely of resistors. In Sec. 3 we prove the resistance inequalities.

I. C^*-ALGEBRAIC FORMULATION OF THE HEISENBERG MODEL

We refer to Ruelle's book[1] for a general reference to quantum lattice systems, to Sakai's book[2] for a general reference on C^*-algebras, and to Wigner's book[3] for a general reference to the description of quantum spin.

To describe the spin of a single particle of spin $j=\frac{1}{2}, 1, 1\frac{1}{2}, 2, 2\frac{1}{2}, \cdots$, one uses a Hilbert space of dimension $d=2j+1$. On this Hilbert space there act three Hermitian linear operators $\mathbf{S}=(S_x, S_y, S_z)$ satisfying the relations

$$[S_x, S_y]=iS_z, \quad [S_y, S_z]=iS_x,$$
$$[S_z, S_x]=iS_y, \qquad (1.1)$$
$$\mathbf{S}^2=\mathbf{S}\cdot\mathbf{S}=S_x^2+S_y^2+S_z^2=j(j+1)I,$$

where I is the identity operator on the Hilbert space $\mathcal{H}$ and $[A,B]=AB-BA$.

As is well known, the operators $S=(S_x, S_y, S_z)$ set irreducibly on $\mathcal{H}$. One can choose an orthonormal basis $\{f_m: m=-j, 1-j, 2-j, \cdots, j-1, j\}$ so that

$$S_z f_m = m f_m,$$
$$S_+ f_m = (S_x+iS_y)f_m=\sqrt{(j-m)(j+m+1)}\,f_{m+1}, \qquad (1.2)$$
$$S_- f_m = (S_x-iS_y)f_m=\sqrt{(j+m)(j-m+1)}\,f_{m-1}.$$

We also introduce the operators $\mathbf{s}=(s_x, s_y, s_z)$ where $\mathbf{s}=\mathbf{S}/j$. The advantage of these operators is that they are normalized so that for $\mathbf{n}\in R^3$

$$\|\mathbf{n}\cdot\mathbf{s}\|=\|n_x s_x+n_y s_y+n_z s_z\|=|\mathbf{n}|=(n_x^2+n_y^2+n_z^2)^{1/2},$$

where $\|\cdot\|$ indicates the operator norm. For $j=\frac{1}{2}$ the $\mathbf{s}$ are the Pauli spin matrices,

$$s_x=\begin{pmatrix}0&1\\1&0\end{pmatrix}, \quad s_y=\begin{pmatrix}0&-1\\1&0\end{pmatrix}, \quad s_z=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

The Heisenberg model describes a system of particles on a lattice $\mathcal{L}$ where the interaction between particles is only through their spin. We define the C^*-algebra associated with the Heisenberg model.[4,5] Let $\mathcal{L}$ be a finite or countably infinite set. We will usually denote the points of $\mathcal{L}$ by i,j,k,l, etc. We suppose that at each point $k\in\mathcal{L}$ there is a particle of spin j_k. In more mathematical terms we assume that to each point $k\in\mathcal{L}$ there is a C^*-algebra $\mathfrak{A}_k$ generated by all polynomials in the three operators $\mathbf{S}_k=(S_{kx}, S_{ky}, S_{kz})$, where these elements satisfy equations (1.1) with $j=j_k$. As is well known, this algebra $\mathfrak{A}_k$ has one and only one irreducible $*$-representation (up to unitary equivalence) on a Hilbert space and that representation is on a Hilbert space $\mathcal{H}_k$ of dimension $d_k=2j_k+1$. One may choose a basis for $\mathcal{H}_k$ so that the operators $\mathbf{S}_k$ act on these basis vectors as given in Eq. (1.2).

Since $\mathfrak{A}_k$ is $*$-isomorphic to $\mathcal{B}(\mathcal{H}_k)$ the $*$-algebra of all bounded operators on $\mathcal{H}_k$, $\mathfrak{A}_k$ is an $(n\times n)$-matrix algebra with $n=2j_k+1=$ dimension of $\mathcal{H}_k$.

If $\Lambda\subset\mathcal{L}$ is a finite subset of $\mathcal{L}$ we denote by $\mathfrak{A}_\Lambda$ the tensor product of the algebras $\mathfrak{A}_k$ with $k\in\Lambda$, i.e.,

$$\mathfrak{A}_{\{k_1,k_2,\cdots,k_n\}}=\mathfrak{A}_{k_1}\otimes\mathfrak{A}_{k_2}\otimes\cdots\otimes\mathfrak{A}_{k_n}.$$

Since the tensor product of an $(n\times n)$-matrix algebra with an $(m\times m)$-matrix algebra is an $(nm\times nm)$-matrix algebra, $\mathfrak{A}_\Lambda$ is an $(r\times r)$-matrix algebra with $r=\prod_{k\in\Lambda}(2j_k+1)$.

Let Q be the collection of all finite subsets Λ of $\mathcal{L}$. The matrix algebras $\mathfrak{A}_\Lambda$ satisfy the following relations. If $\Lambda_1 \supset \Lambda_2$ then $\mathfrak{A}_{\Lambda_1} \supset \mathfrak{A}_{\Lambda_2}$. If $\Lambda_1, \Lambda_2 \in Q$ then $\mathfrak{A}_{\Lambda_1 \cup \Lambda_2}$ is generated as a $*$-algebra by $\mathfrak{A}_{\Lambda_1}$ and $\mathfrak{A}_{\Lambda_2}$ and $\mathfrak{A}_{\Lambda_1 \cap \Lambda_2} = \mathfrak{A}_{\Lambda_1} \cap \mathfrak{A}_{\Lambda_2}$. If Λ_1 and Λ_2 are disjoint, then each element of $\mathfrak{A}_{\Lambda_1}$ commutes with each element of $\mathfrak{A}_{\Lambda_2}$.

If Λ is an infinite subset of $\mathcal{L}$, we define $\mathfrak{A}_\Lambda = \overline{\cup_{\Lambda'} \mathfrak{A}_{\Lambda'}}$, where the union is taken over all $\Lambda' \in Q$ with $\Lambda' \subset \Lambda$. The bar denotes the completion of the algebra with respect to its unique norm. A more detailed account of this construction can be found in Refs. 4 and 5. For Λ infinite $\mathfrak{A}_\Lambda$ is a uniformly hyperfinite algebra or UHF-algebra because $\mathfrak{A}_\Lambda$ contains a increasing sequence of matrix algebra whose union is norm dense in $\mathfrak{A}_\Lambda$. UHF-algebra were defined and studied by Glimm.[6] We will refer to $\mathfrak{A}_\Lambda$ as the Heisenberg spin algebra over Λ.

The interaction in the isotropic Heisenberg model is given in terms of a Hamiltonian,

$$H = \tfrac{1}{2} \sum_{i,j \in \mathcal{L}} J(i,j)(I - \mathbf{s}_i \cdot \mathbf{s}_j),$$

where $\mathbf{s}_i \cdot \mathbf{s}_j = s_{ix}s_{jx} + s_{iy}s_{jy} + s_{iz}s_{jz}$ and $\mathbf{s}_i = \mathbf{S}_i / j_i$. The $J(i,j)$ are real numbers so that $J(i,i) = 0$ and $J(i,j) = J(j,i)$ for all $i,j \in \mathcal{L}$. In this paper we will consider only the ferromagnetic case with $J(i,j) \geq 0$ for all $i,j \in \mathcal{L}$.

A difficulty with the above expression for H is that for cases of physical interest the sum is not convergent. If Λ is a finite subset of $\mathcal{L}$, we define

$$H_\Lambda = \tfrac{1}{2} \sum_{i,j \in \Lambda} J(i,j)(I - \mathbf{s}_i \cdot \mathbf{s}_j).$$

This expression defines an element of the algebra. The dynamics of the Heisenberg model is given by a strongly continuous one parameter group of $*$-automorphisms $\{t \to \alpha_t\}$ of $\mathfrak{A}$. Heuristically, there automorphisms are given by

$$\alpha_t(A) = \exp(itH) A \exp(-itH).$$

Since H is not an element of the algebra, the above expression is not well defined. One can define the above automorphisms by replacing H by H_Λ and letting Λ increase to $\mathcal{L}$. In fact, it is known that if

$$\sup\Big\{ \sum_{j \in \mathcal{L}} |J(i,j)| \, ; i \in \mathcal{L} \Big\} < \infty \tag{1.3}$$

and if Λ_n is an increasing sequence of finite sets whose union is $\mathcal{L}$, then there is a strongly continuous one-parameter group of $*$-automorphisms $\{\alpha_t\}$ of $\mathfrak{A}_{\mathcal{L}}$ and

$$\alpha_t(A) = \lim_{n \to \infty} \exp(itH_{\Lambda_n}) A \exp(-itH_{\Lambda_n})$$

for all $A \in \mathfrak{A}(\mathcal{L})$, where the limit converges in norm. A discussion of the proof of this result can be found in Refs. 4 and 5. Automorphism groups which can be approximated in the above fashion by inner automorphism groups are called approximately inner automorphism groups.[7]

We conclude this section with some remarks concerning the spectrum of $I - \mathbf{s}_i \cdot \mathbf{s}_j$. Suppose $r, s \in \mathcal{L}$ and $\mathbf{S} = \mathbf{S}_r + \mathbf{S}_s$. We have

$$[S_x, S_y] = iS_z, \quad [S_y, S_z] = iS_x, \quad [S_z, S_x] = iS_y.$$

Let $\mathcal{B}$ be the $*$-algebra generated by the three elements S_x, S_y, and S_z. Since $\mathcal{B}$ is a finite-dimensional $*$-algebra, $\mathcal{B}$ has a central decomposition as a direct sum of matrix algebras. This decomposition of $\mathcal{B}$ is directly related to the problem of decomposing the tensor product of two irreducible representation of SU(2) into irreducible representations. The solution to these problems is well known. Let $n = 2\min(j_r, j_s) + 1$. There are n orthogonal minimal central projections $\{E_k, k = 1, \ldots, n\}$ (i.e., $E_k \in \mathcal{B}$ and $AE_k = E_kA$ for all $A \in \mathcal{B}$ and $k = 1, \ldots, n$ and $\sum_{k=1}^n E_k = I$) so that

$$\mathbf{S}^2 E_k = (S_x^2 + S_y^2 + S_z^2) E_k = l_k(l_k + 1) E_k$$

with $l_k = j_r + j_s + 1 - k$. It follows that $\mathcal{B} = \oplus_{k=1} \mathcal{B} E_k$ and $\mathcal{B} E_k$ is a $((2l_k + 1) \times (2l_k + 1))$ matrix algebra.

Applying these results to the element $I - \mathbf{s}_r \cdot \mathbf{s}_s$ we find

$$\begin{aligned}(I - \mathbf{s}_r \cdot \mathbf{s}_s) E_k &= (j_r j_s)^{-1}(j_r j_s I - \mathbf{S}_r \cdot \mathbf{S}_s) E_k \\ &= (j_r j_s)^{-1}[j_r j_s I + \tfrac{1}{2}\mathbf{S}_r^2 + \tfrac{1}{2}\mathbf{S}_s^2 - \tfrac{1}{2}(\mathbf{S}_r + \mathbf{S}_s)^2] E_k \\ &= \tfrac{1}{2}(j_r j_s)^{-1}[(j_r + j_s)(j_r + j_s + 1) - l_k(l_k + 1)] E_k \\ &= \tfrac{1}{2}(j_r j_s)^{-1}(k - 1)(2j_r + 2j_s + 2 - k) E_k.\end{aligned}$$

Hence, the spectrum of $I - \mathbf{s}_r \cdot \mathbf{s}_s$ consists of the numbers $\{\tfrac{1}{2}(j_r j_s)^{-1}(k - 1)(2j_r + 2j_s + 2 - k)$ for $k = 1, \ldots, n\}$. It follows that $I - \mathbf{s}_r \cdot \mathbf{s}_s$ is positive and zero is in its spectrum.

II. RESISTANCE OF ELECTRICAL NETWORKS

In this section we collect various known results needed to calculate and estimate the resistance of electrical networks. We refer to Ref. 8 as a general reference. Suppose $\mathcal{L}$ is a finite set and G is a graph with vertices $\mathcal{L}$. The graph G is simply a set of lines (i,j) connecting pairs of vertices. We assume G is connected. We assume that with each line (i,j) of G there is associated a positive real number $J(i,j) > 0$. We imagine that each line corresponds to a resistor of $J(i,j)^{-1}$ ohms. Given two vertices i and j, we wish to calculate the resistance $R(i,j)$ between i and j.

There are two physical laws which enable one to calculate the resistance of an electrical network. One is Ohm's law which states that the potential difference $V(i) - V(j)$ between two vertices i and j connected by a line (i,j) of G is equal to the current $I(i,j)$ flowing from i to j times the resistance $J(i,j)^{-1}$ associated with the line (i,j) [i.e., $V(i) - V(j) = J(i,j)^{-1} I(i,j)$]. The second physical law is the Kirchhoff law which states that the total current flowing into a vertex must equal the total current flowing out of that vertex. One can calculate the resistance between two vertices i and j by injecting one ampere of current at the vertex i and extracting one ampere of current at the vertex j. Ohm's law and the Kirchhoff law determine the voltages $V(k)$ at all the vertices $k \in \mathcal{L}$ (to within an additive constant). The resistance between i and j is the difference in voltage $V(i) - V(j)$ between i and j.

We will now restate these ideas in a more mathematical form. Suppose G is a connected graph with vertices $\mathcal{L}$. Suppose with each line (i,j) of G there is associated a positive number $J(i,j) > 0$. The graph G together with the positive numbers $J(i,j)$ will be called a network. We denote by $G(i)$ the set of vertices of $\mathcal{L}$ connected to i by a line of G.

Suppose f is a real function on $\mathcal{L}$, i.e., f is a mapping of $\mathcal{L}$ into the real numbers. We define the Laplacian of f denoted by Δf as the function

$$(\Delta f)(i) = \sum_{j\in G(i)} J(i,j)[f(j) - f(i)]. \tag{2.1}$$

We say f is harmonic if $\Delta f = 0$. Note if f is harmonic, the value of f at each vertex i is a convex combination of the values of f at the vertices connected to i, i.e., if $(\Delta f)(i) = 0$, then

$$\sum_{j\in G(i)} J(i,j)f(i) = \sum_{j\in G(i)} J(i,j)f(j)$$

or

$$f(i) = \sum_{j\in G(i)} \lambda_j f(j)$$

with $\lambda_j = J(i,j)\left[\sum_{k\in G(i)} J(i,k)\right]^{-1}$.

We have that the λ_i are positive and their sum is one. It follows that if f is a real harmonic function on $\mathcal{L}$ which attains its maximum value at the vertex i, then f must also attain its maximum value at all the vertices connected with i. Since the graph G is connected and the number of vertices is finite, it follows that every real harmonic function is constant, i.e., $f(i) = c$ for all $i \in \mathcal{L}$.

With the aid of the Laplacian we define the resistance between two ventices as follows. Let δ_i be the function on $\mathcal{L}$ given by $\delta_i(j) = 0$ if $i \neq j$ and $\delta_i(i) = 1$. Suppose $i, j \in \mathcal{L}$ and V is a solution to the equation

$$-\Delta V = \delta_i - \delta_j. \tag{2.2}$$

The resistance between i and j is defined to be $V(i) - V(j)$.

We will take the time to outline a proof that Eq. (2.2) has a unique solution (up to an additive constant) since the proof also provides a useful way of estimating the resistance between two vertices. Suppose for each line (i,j) of G we specify a real number $I(i,j)$. We may think of $I(i,j)$ as the current flowing along the line (i,j). If the current is flowing from i to j, we have $I(i,j) > 0$. If the current is flowing from j to i we have $I(i,j) < 0$. With this convention we have $I(i,j) = -I(j,i)$.

We wish to calculate the resistance between the vertices i_0 and j_0. We imagine injecting one ampere of current at i_0 and extracting one ampere of current at j_0. Now consider a current flow specified by numbers $I(r,s)$ for each line (r,s) of G. We say this current flow is admissible if it satisfies the Kirchhoff laws, i.e., the total current flow into a vertex equals the total current flowing out of that vertex except at i_0 where the total outflow of current is one ampere and at j_0 where the total inflow is one ampere. In terms of equations these constraints read if $\{I(i,j)\}$ is an admissible current flow, then for $r \neq i_0$ and $r \neq j_0$

$$\sum_{i\in G(r)} I(i,r) = 0$$

and

$$\sum_{i\in G(i_0)} I(i_0,i) = 1, \quad \sum_{i\in G(j_0)} I(i,j_0) = 1.$$

Certainly, there exist admissible current flows. For example, let $(i_0,i_1), (i_1,i_2), \ldots, (i_{n-1},j_0)$ be a sequence of lines of G connecting i_0 to j_0. If we simply set $I(i_0,i_1) = I(i_1,i_2) = \cdots = I(i_{n-1},j_0) = 1$ and set all other $I(i,j) = 0$, we obtain an admissible flow.

Given an admissible flow $\{I(i,j)\}$, we define the dissipation D of this flow as

$$D = D(\{I(i,j)\}) = \sum_{(i,j)\in G} J(i,j)^{-1} I(i,j)^2. \tag{2.3}$$

We consider the problem of minimizing the dissipation. One can prove the existence of minimally dissipative flows as follows. Construct some admissible flow and calculate its dissipation. The set of admissible flows with smaller or equal dissipation is a compact set with the obvious topology on current flows, i.e., two current flows $I(i,j)$ and $I'(i,j)$ are close if $I(i,j) - I'(i,j)$ is small for each line $(i,j) \in G$. Since the dissipation is a continuous function on this compact set the dissipation achieves it minimum value.

Suppose $\{I(i,j)\}$ is a minimally dissipative admissible current flow. Suppose $(i_0,i_1), (i_1,i_2), \ldots, (i_{n-1},i_n), (i_n,i_0)$ is a closed path starting at i_0 and returning to i_0. We show that

$$C = J(i_0,i_1)^{-1} I(i_0,i_1) + J(i_1.\, i_2)^{-1} I(i_1,i_2) + J(i_n,i_0)^{-1} I(i_n,i_0) = 0.$$

To see this, suppose we form a new admissible flow by defining $I'(i_0,i_1) = I(i_0,i_1) + \lambda$, $I'(i_1,i_2) = I(i_1,i_2) + \lambda, \ldots,$ $I'(i_n,i_0) = I(i_n,i_0) + \lambda$ and for all other lines $I'(i,j) = I(i,j)$. The dissipation of the new flow minus the dissipation of the original flow is given by

$$\begin{aligned} D(I') - D(I) = {} & J(i_0,i_1)^{-1}[2\lambda I(i_0,i_1) + \lambda^2] + J(i_1,i_2)^{-1} \\ & \times [2\lambda I(i_1,i_2) + \lambda^2] + \cdots + J(i_n,i_0)^{-1} \\ & \times [2\lambda I(i_n,i_0) + \lambda^2]. \end{aligned}$$

Since $\{I(i,j)\}$ is a minimally dissipative flow, the derivative of the above expression with respect to λ at $\lambda = 0$ must equal zero. Hence, we have $C = 0$. We now define the potential function V by

$$-V(j) = J(i_0,i_1)^{-1} I(i_0,i_1) + \cdots + J(i_n,j)^{-1} I(i_n,j),$$

where $(i_0,i_1), (i_1,i_2), \ldots, (i_n,j)$ is a path from i_0 to j. The number $V(j)$ is independent of the path chosen since the sum C around a closed path is zero. Note that $V(i_0) = 0$. A straightforward computation shows that $-\Delta V = \delta_{i_0} - \delta_{j_0}$. Therefore, a solution to Eq. (2.2) exists. If V' is a second solution to Eq. (2.2), then $-\Delta(V - V') = 0$ and $V - V'$ is harmonic and, therefore, a constant function. Hence, Eq. (2.2) has a unique solution up to a constant function.

The resistance $R(i_0,j_0)$ between i_0 and j_0 is $V(i_0) - V(j_0)$, where V satisfies Eq. (2.2). We also remark that the resistance $R(i_0,j_0)$ is the minimum dissipation. This can be seen a number of ways. One of the simplest is to note that the total energy dissipated by an electrical network [$D =$ voltage difference times current $= V(i_0) - V(j_0) = R(i_0,j_0)$] is the sum of the energies dissipated in each of its parts. We will give a fairly involved proof of this fact not because we will need this fact but because we will need the methods used in the proof in the next section of this paper.

Suppose $i_0, j_0 \in \angle$ and V is a solution to Eq. (2.2). Note V will take its maximum value at i_0 and its minimum value at j_0. We consider the current paths from i_0 to j_0 where a current path is defined as follows. A current path is a sequence of distinct vertices $(i_0, i_1, i_2, \ldots, i_n)$ starting with i_0 and ending at $i_n = j_0$, so that the kth and $(k+1)$th vertices are connected by a line of G and $V(i_k) > V(i_{k+1})$ for $k = 0, 1, \ldots, n-1$. In more intuitive terms a current path is simply a path from i_0 to j_0 along which a particle of current might travel. We define the terms upstream and downstream. We say i is immediately upstream of j if $(i,j) \in G$ and $V(i) > V(j)$. We say i is immediately downstream of j if $(i,j) \in G$ and $V(i) < V(j)$.

Let Q be the set of all possible current paths from i_0 to j_0. If $p \in Q$ is a current path we assign a probability $s(p)$ to p as follows. If $p = (i_0, i_1, \ldots, i_n)$ (with $i_n = j_0$) is a current path from i_0 to j_0, we define

$$s(p) = I(i_0, i_1)A(i_0)^{-1} I(i_1, i_2)A(i_1)^{-1} \circ I(i_{n-1}, i_n)A(i_{n-1})^{-1}, \tag{2.4}$$

where $I(i,j) = J(i,j)[V(i) - V(j)]$ is the current flowing from i to j and $A(i)$ is what we will call the activity of i which is the total current flowing into i (or the total current flowing out of i), i.e.,

$$A(i) = \tfrac{1}{2} \sum_{j \in G(i)} |I(i,j)|, \quad i \neq i_0 \text{ or } i \neq j_0,$$

and $A(i_0) = A(j_0) = 1$. The ratio $I(i,j)/A(i)$ is the probability that a particle of current having arrived at i will flow to j. Intuitively the number $s(p)$ is the probability that a particle of current will follow the path p in its flow from i_0 to j_0.

Lemma 2.1: Suppose G is a finite connected network with vertices $\angle$ and resistances $J(i,j)^{-1}$ associated with each of its lines. Suppose $i_0, j_0 \in \angle$ and V is a solution to Eq. (2.2). Let Q be the set of all current paths from i_0 to j_0 and let $s(\cdot)$ be the function defined by Eq. (2.4). Suppose $r, s \in \angle$ and r is immediately upstream of s. Let $Q(r,s)$ be the set of all current paths which pass through r, s. Then,

$$I(r,s) = J(r,s)[V(r) - V(s)] = \sum_{p \in Q(r,s)} s(p).$$

Sketch of proof: This lemma states that the current $I(r,s)$ is equal to the probability that a particle of current flowing from i_0 to j_0 will pass through the line (r,s). Consider the sum

$$\sum_{p \in Q(r,s)} s(p) = \sum_{p \in Q(r,s)} [I(i_0, i_1)A(i_0)^{-1}] \cdots \times [I(r,s)A(r)^{-1}] \cdots [I(i_{n-1}, j_0)A(i_{n-1})^{-1}].$$

All possible paths $p \in Q(r,s)$ can be generated as follows. Start at r and move upstream one vertex at a time until reaching i_0 and start at s and move downstream one vertex at a time until reaching j_0. For each different set of choices one obtains a different path $p \in Q(r,s)$. In the above sum for each vertex k (with k either equal to r or upstream of r) we sum over all vertices immediately upstream of k. The effect of this summation is to cancel the $A(k)$ in the denominator with the sum of the $I(i,k)$ with the i the upstream vertices of k. Similarly for each k (with k either equal to s or downstream of s) we sum over all vertices i immediately downstream of k. The effect of this summation is to cancel the $A(k)$ in the denominator with the sum of the $I(k,i)$ with 1 immediately downstream of k. The results of these cancellations leaves the single term $I(r,s)$. The completes the sketch of the proof of the lemma.

As a application of this lemma we show that the resistance $R(i_0, j_0)$ between i_0 and j_0 is equal to the minimum dissipation as defined by Eq. (2.3). Consider a path $p \in Q$ with $p = (i_0, i_1, \ldots, i_n)$ with $i_n = j_0$. We consider p as a union of lines $(i_0, i_1), (i_1, i_2), \ldots, (i_{n-1}, i_n)$. Then we have

$$\begin{aligned}\sum_{(i,j) \in p} V(i) - V(j) &= [V(i_0) - V(i_1)] + [V(i_1) - V(i_2)] + \cdots \\ &\quad + [V(i_{n-1}) - V(i_n)] \\ &= V(i_0) - V(j_0).\end{aligned}$$

Then, we have

$$\begin{aligned}\sum_{p \in Q} \sum_{(i,j) \in p} s(p)[V(i) - V(j)] &= \sum_{p \in Q} s(p)[V(i_0) - V(j_0)] \\ &= V(i_0) - V(j_0).\end{aligned}$$

In the first sum we interchange the order of summation as follows. We choose a line (i,j) of G with $V(i) > V(j)$ and then sum over all $p \in Q(i,j)$ and then sum over all lines (i,j) of G. Then, we find

$$V(i_0) - V(j_0) = \sum_{(i,j) \in G} \sum_{p \in Q(i,j)} s(p)[V(i) - V(j)].$$

From Lemma 2.1 we have $\sum_{p \in Q(i,j)} s(p) = I(i,j)$. Hence, we have

$$\begin{aligned}R(i_0, j_0) &= V(i_0) - V(j_0) \\ &= \sum_{(i,j) \in G} I(i,j)[V(i) - V(j)] \\ &= \sum_{(i,j) \in G} J(i,j)^{-1} I(i,j)^2 = D(\{I(i,j)\}).\end{aligned}$$

Since the current flow $I(i,j)$ with $I(i,j) = J(i,j)[V(i) - V(j)]$ minimizes the dissipation, we have the minimum dissipation equals the resistance $R(i_0, j_0)$.

We have given this rather involved proof of the equality of the resistance with the minimal dissipation in order to introduce this method of interchanging the order of summation used above. We will use this method again in the next section.

The equality between the resistance $R(i_0, j_0)$ and the minimal dissipation is useful in determining some of the properties of the resistance $R(i_0, j_0)$ as a function of the resistances $J(i,j)^{-1}$ of the lines of G. For example, suppose the resistance $R(i_0, j_0)$ has been determined for a certain network G. Suppose one of the resistances $J(i,j)^{-1}$ is decreased [i.e., $J(i,j)$ is increased]. Then the resistance $R(i_0, j_0)$ must decrease or remain the same. To see this, consider the admissible flow which minimizes the dissipation for the original network. Using that admissible flow, we calculate the dissipation for the new network in which $J(i,j)^{-1}$ is decreased. We find for the new network a smaller or equal dissipation. Since the resistance $R(i_0, j_0)$ for the new network is the minimal dissipation, we have that the resistance $R(i_0, j_0)$ for the new network is smaller or equal to the resistance $R(i_0, j_0)$ for the original network. Hence, we have shown

$$\frac{\partial R(i_0,j_0)}{\partial J(i,j)} \le 0.$$

Another application of the equality of the resistance with the minimal dissipation is the following. Suppose G is a finite network with vertices $\mathcal{L}$ and $R(i_0,j_0)$ is the resistance between $i_0,j_0 \in \mathcal{L}$. Now, suppose G' is a larger network with vertices $\mathcal{L}'$ containing G, i.e., $\mathcal{L}' \supset \mathcal{L}$, $G' \supset G$, and $J(i,j)=J'(i,j)$ for the lines $(i,j) \in G$. Let $R'(i_0,j_0)$ be the resistance between i_0 and j_0 as calculated in G'. We see that $R'(i_0,j_0) \le R(i_0,j_0)$ since the minimally dissipative admissible flow for G is an admissible flow for G' and, therefore, the minimum dissipation for G' is less than or equal to the minimum dissipation in G.

This fact gives us a convenient way to define the resistance in an infinite network. If G is an infinite network with vertices $\mathcal{L}$, we define the resistance $R(i_0,j_0)$ between $i_0,j_0 \in \mathcal{L}$ as the greatest lower bound of all the resistances $R'(i_0,j_0)$ computed by taking finite subgraphs of G. Since the resistance decreases as the network increases, we have the following. Suppose G is an infinite network and i_0 and j_0 are vertices of G. Suppose $\{G_n\}$ is an increasing sequence of finite connected subnetworks of G containing i_0 and j_0 as vertices. Suppose G_n increases to G as $n \to \infty$, i.e., each line $(i,j) \in G$ is contained in some G_n. Let $R_n(i_0,j_0)$ be the resistance between i_0 and j_0 as calculated in G_n and let $R(i_0,j_0)$ be the resistance between i_0 and j_0 in G. Then $R(i_0,j_0) = \lim_{n\to\infty} R_n(i_0,j_0)$ since the numbers $R_n(i_0,j_0)$ decrease with increasing n and since every finite subnetwork of G is contained in some G_n.

We mention that if the numbers $\{J(i,j)\}$ satisfy inequality (1.3), then $R(i_0,j_0) > 0$ is strictly positive for any two distinct vertices of G. We also mention that the resistance $R(i_0,j_0)$ is equal to the greatest lower bound of all the dissipations calculated from admissible current flows since this is true for finite systems and follows by taking limits for infinite systems. The definition of resistance in terms of the potential V should be used with caution for infinite systems since there are infinite networks for which there exist nonconstant harmonic functions which are bounded.

Next, we would like to mention how to estimate the increase in resistance $R(i_0,j_0)$ if some of the lines (r,s) are removed from a network. Suppose G is a finite network with vertices $\mathcal{L}$ and $R(i_0,j_0)$ is the resistance between i_0 and $j_0 \in \mathcal{L}$. Suppose G' is a connected subnetwork of G with vertices $\mathcal{L}'$ (and $i_0,j_0 \in \mathcal{L}'$) and G' is obtained from G by removing the lines (r_i,s_i), $i=1,\ldots,n$. We wish to estimate the resistance $R'(i_0,j_0)$ as calculated in G' in terms of the resistance $R(i_0,j_0)$ as calculated in G. Let $I(i,j)$ be the admissible current flow in G which minimizes the dissipation. We will show that if $\sum_{i=1}^n |I(r_i,s_i)| < 1$, then

$$R(i_0,j_0) \le R'(i_0,j_0) \le R(i_0,j_0)\left(1 - \sum_{i=1}^n |I(r_i,s_i)|\right)^{-2}. \tag{2.5}$$

This shows that if the lines (r_i,s_i) carry little current, then the resistance does not change much when these lines are removed from the network. To see this we express the current flow $\{I(i,j)\}$ as a sum of simple current flows $s(p)$ along current paths $p \in Q$ as was described in Lemma 2.1. Let $Q(r_i,s_i)$ be the current paths which pass through the line (r_i,s_i) and let $W = \cup_{i=1}^n Q(r_i,s_i)$ be their union. We have

$$\sum_{p\in W} s(p) \le \sum_{i=1}^n \sum_{p\in Q(r_i,s_i)} s(p) = \sum_{i=1}^n |I(r_i,s_i)|.$$

We assume $\sum_{i=1}^n |I(r_i,s_i)| < 1$. We denote by $Q-W$ the paths in Q not in W. We have

$$\lambda = \sum_{p\in Q-W} s(p) = 1 - \sum_{p\in W} s(p) \ge 1 - \sum_{i=1}^n |I(r_i,s_i)| > 0.$$

Let $I'(i,j)$ be the current flow given by

$$I'(i,j) = \lambda^{-1} \sum_{p\in Q(i,j)-W} s(p) \le \lambda^{-1} I(i,j)$$

for each line (i,j) of G' with i upstream of j and if i and j are of the same potential $I'(i,j)=0$. The current flow $\{I'(i,j)\}$ is obtained from $\{I(i,j)\}$ by removing the current paths which pass through the lines (r_i,s_i) and then renormalizing the current flow by multiplying by λ^{-1}. We have $\{I'(i,j)\}$ is an admissible flow in G'. Clearly, we have $\lambda |I'(i,j)| \le |I(i,j)|$ for all lines $(i,j) \in G'$. We estimate the dissipation associated with the flow $\{I'(i,j)\}$.

$$\begin{aligned} D(I') &= \sum_{(i,j)\in G'} J(i,j)^{-1} I'(i,j)^2 \\ &\le \sum_{(i,j)\in G} J(i,j)^{-1}\lambda^{-2} I(i,j)^2 = \lambda^{-2} R(i_0,j_0) \end{aligned}$$

Since $R'(i_0,j_0)$ is the minimum dissipation, we have

$$\begin{aligned} R'(i_0,j_0) &\le D(I') \le \lambda^{-2} R(i_0,j_0) \\ &\le R(i_0,j_0)\left(1 - \sum_{i=1}^n |I(r_i,s_i)|\right)^{-2}. \end{aligned}$$

This establishes inequality (2.5). The left most inequality of (2.5) follows from the fact that $G' \subset G$.

We conclude this section with a discussion of a regular cubic lattice with nearest neighbor connection in three dimensions. Let $\mathcal{L} = Z^3$ be the set of all three-tuples of integers, e.g., $i = (i_x, i_y, i_z) \in \mathcal{L}$, and let G be the graph of all lines connecting each vertex of $\mathcal{L}$ with its six nearest neighbors, i.e., $(i,j) \in G$ if and only if $|i-j| = |i_x - j_x| + |i_y - j_y| + |i_z - j_z| = 1$. We associate a resistance of one ohm with each line of G, i.e., $J(i,j) = 1$ for all $(i,j) \in G$.

If follows from known results[4] that the only bounded harmonic functions on $\mathcal{L} = Z^3$ are the constant functions and, furthermore, the equation $-\Delta V = \delta_{i_0}$ with $V(i_0) = 0$ has a unique bounded solution which we will denote by V_{i_0}. This function has the property that $V_{i_0}(i) \to -\frac{1}{2}R_\infty$ as $|i| = |i_x| + |i_y| + |i_z| \to \infty$ where $R_\infty = 0.50546\cdots$. Also, we have $-\frac{1}{2}R_\infty < V_{i_0}(i) \le 0$ for all $i \in \mathcal{L}$.

Heuristically, this shows that the resistance between i_0 and "infinity" is finite in three dimensions. Given two vertices $i_0, j_0 \in \mathcal{L}$, we define $V = V_{i_0} - V_{j_0}$. We have $-\Delta V = \delta_{i_0} - \delta_{j_0}$. Then, the resistance $R(i_0,j_0)$ between i_0 and j_0 is given by $R(i_0,j_0) = V(i_0) - V(j_0) = -V_{j_0}(i_0) - V_{i_0}(j_0) = -2V_{i_0}(j_0) < R_\infty$. Hence, the resistance between any two vertices is less that R_∞ in three dimensions. In two dimensions the resistance grows logarithmically with the distance between the vertices.

One may show that the resistance calculated from the potential function $[R(i_0,j_0)=V(i_0)-V(j_0)=-2V_{i_0}(j_0)]$ agrees with the definition of resistance for infinite networks given earlier. Consider an increasing sequence $\{G_n\}$ of finite connected subnetworks of G with vertices $\mathcal{L}_n \subset \mathcal{L}$ and with $i_0,j_0 \in \mathcal{L}_n$ for each n. We assume the G_n increase up to G, i.e., $G=\cup_{n=1}^{\infty} G_n$. Let ϕ_n be the solution to the equation $-\Delta\phi_n=\delta_{i_0}-\delta_{j_0}$ on G_n with $\phi_n(i_0)+\phi_n(j_0)=0$. The function ϕ_n are uniformly bounded. In fact, $\phi_n(i_0)\geq\phi_n(i)\geq\phi_n(j_0)$ for all $i\in\mathcal{L}_n$ and $\phi_n(i_0)=-\phi_n(j_0)$ decreases as n increases since the networks G_n are increasing and, therefore, the resistances $R_n(i_0,j_0)=\phi_n(i_0)-\phi_n(j_0)$ are decreasing as n increases. We extend ϕ_n to a function on $\mathcal{L}=Z^3$ by defining $\phi_n(i)=0$ for $i\in\mathcal{L}-\mathcal{L}_n$. Since the ϕ_n are uniformly bounded there is a subsequence $\{\phi_{n(k)},\ k=1,2,\cdots\}$, which converges pointwise to a bounded function $\phi(i)=\lim_{k\to\infty}\times\phi_{n(k)}(i)$. One can easily show that $-\Delta\phi=\delta_{i_0}-\delta_{j_0}$. Since the only bounded harmonic functions on $\mathcal{L}=Z^3$ are the constant functions, we have $\phi-V$ is a constant function. Since $\phi(i_0)+\phi(j_0)=V(i_0)+V(j_0)=0$, we have $\phi=V$. Hence, we have

$$R(i_0,j_0)=\lim_{k\to\infty}R_{n(k)}(i_0,j_0)$$
$$=\lim_{k\to\infty}\phi_{n(k)}(i_0)-\phi_{n(k)}(j_0),$$
$$\phi(i_0)-\phi(j_0)=V(i_0)-V(j_0)=-2V_{i_0}(j_0).$$

Hence, the resistance as calculated from the potential is equal to the resistance as defined earlier for infinite networks.

Finally, we would like to state a lemma which shows that the resistance between vertices distant from the origin is only slightly increased by removing a few lines near the origin.

Lemma 2.2: Let G be the network with vertices $\mathcal{L}=Z^3$ in which each vertex of $\mathcal{L}$ is connected with a one ohm resistor to its six nearest neighbors. Let G' be the network obtained from G by removing the lines (r_i,s_i), $i=1,\ldots,n$. Let $R(i_0,j_0)$ and $R'(i_0,j_0)$ be the resistances between i_0 and j_0 as calculated in G and G', respectively. Then, $R'(i_0,j_0)-R(i_0,j_0)\to 0$ as $|i_0|,|j_0|\to\infty$. Also, given a positive number $\epsilon>0$ there is a finite set S of vertices so that $R'(i_0,j_0)\leq R_\infty+\epsilon$ for $i_0,j_0\in\mathcal{L}-S$.

Proof: Suppose $i_0,j_0\in\mathcal{L}$ and $V=V_{i_0}-V_{j_0}$. Let $I(i,j)=V(i)-V(j)$ be the current flow associated with V. From our previous discussion we have that if $\sum_{i=1}^n\times|I(r_i,s_i)|<1$, then

$$R(i_0,j_0)\leq R'(i_0,j_0)\leq R(i_0,j_0)\left(1-\sum_{i=1}^n|I(r_i,s_i)|'\right)^{-2}.$$

Note that in our previous discussion we only established this inequality for finite networks. The fact that this inequality holds for G may be seen by taking an increasing sequence $\{G_n\}$ of finite connected subnetworks of G which increase up to G and then taking the limit of an appropriate subsequence.

As $|i_0|\to\infty$ we have $V_{i_0}(i)\to-\frac{1}{2}R_\infty$. Then, as $|i_0|,|j_0|\to\infty$ we have $V(i)=V_{i_0}(i)-V_{j_0}(i)\to 0$ for fixed $i\in\mathcal{L}$. It follows that $I(r_i,s_i)=V(r_i)-V(s_i)\to 0$ as $|i_0|,|j_0|\to\infty$. Hence, we have

$$0\leq R'(i_0,j_0)-R(i_0,j_0)$$
$$\leq R(i_0,j_0)\left[\left(1-\sum_{i=1}^n|I(r_i,s_i)|\right)^{-2}-1\right]$$
$$\to 0\quad\text{as } |i_0|,|j_0|\to\infty.$$

Since $R(i_0,j_0)<R_\infty$ for all $i_0,j_0\in\mathcal{L}$ the last statement of the lemma follows. This completes the proof of the lemma.

III. RESISTANCE INEQUALITIES

Let G be a finite network with vertices $\mathcal{L}$. We assume that to each vertex $i\in\mathcal{L}$ there is associated a particle of spin j_i. Let $\mathfrak{A}=\mathfrak{A}_L$ be the Heisenberg spin algebra over $\mathcal{L}$ as described in Sec. 1. The algebra $\mathfrak{A}$ is generated by the elements $\mathbf{S}_i=(S_{ix},S_{iy},S_{iz})$, which satisfy Eq. (1.1) with $j=j_i$ and the elements S_i and S_j commute for $i\neq j$. Let H be the element

$$H=\sum_{(i,j)\in G}J(i,j)(I-\mathbf{s}_i\cdot\mathbf{s}_j),$$

where $\mathbf{s}_i=\mathbf{S}_i/j_i$. We will prove the operator inequality

$$I-\mathbf{s}_i\cdot\mathbf{s}_j\leq R(i,j)H,$$

where $R(i,j)$ is the resistance between i and j. To prove this we will need the following lemma.

Lemma 3.1: Let $\mathfrak{A}$ be the Heisenberg spin algebra describing three particles of spin j_1, j_2, and j_3, i.e., $\mathfrak{A}$ is a C^*-algebra generated by the spin operators $\mathbf{S}_i=(S_{ix},S_{iy},S_{iz})$ for $i=1,2,3$ and the $\mathbf{S}_i$ satisfy Eqs. (1.1) with $j=j_i$. Let $\mathbf{s}_i=\mathbf{S}_i/j_i$. Then, the element

$$H=a(I-\mathbf{s}_1\cdot\mathbf{s}_2)+a(I-\mathbf{s}_1\cdot\mathbf{s}_3)+c(I-\mathbf{s}_2\cdot\mathbf{s}_3)$$

is positive if and only if a, b, and c are real and satisfy

$$a+b+c\geq 0\quad\text{and}\quad ab+ac+bc\geq 0. \qquad (3.1)$$

Furthermore, zero is in the spectrum of H.

Proof: Let $\mathfrak{A}_i$ be the C^*-algebra generated by the three elements $\mathbf{S}_i=(S_{ix},S_{iy},S_{iz})$ satisfying Eqs. (1.1) with $j=j_i$. Then we have $\mathfrak{A}=\mathfrak{A}_1\otimes\mathfrak{A}_2\otimes\mathfrak{A}_3$. Let $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$ be Hilbert spaces of dimension $d_i=2j_i+1$, and let Π_i be the unique (up to unitary equivalence) $*$-representation of $\mathfrak{A}_i$ on $\mathcal{H}_i$ for $i=1,2,3$. Let $\Pi=\Pi_1\otimes\Pi_2\otimes\Pi_3$ be the representation of $\mathfrak{A}$ on $\mathcal{H}=\mathcal{H}_1\otimes\mathcal{H}_2\otimes\mathcal{H}_3$ obtained by taking tensor product of the Π_i, e.g.,

$$\Pi(S_{1x}S_{3y})f_1\otimes f_2\otimes f_3=\Pi_1(S_{1x})f_1\otimes f_2\otimes\Pi_3(S_{3y})f_3.$$

Since Π is a faithful $*$-representation of $\mathfrak{A}$, we have that H is positive (i.e., $H\geq 0$) if and only if $\Pi(H)$ is positive, $\Pi(H)\geq 0$.

Suppose $H\geq 0$. We will show that inequality (3.1) is satisfied. Since $H\geq 0$, we have by convention that $H=H^*$ and therefore a, b, and c are real. Let f_{im} for $m=-j_i,1-j_i,\ldots,j_i-1,j_i$ be an orthonormal basis for $\mathcal{H}_i$, $i=1,2,3$ satisfying Eqs. (1.2) for the operators $\Pi_i(\mathbf{S}_i)$ and let

$$F(m_1,m_2,m_3)=f_{1m_1}\otimes f_{2m_2}\otimes f_{3m_3}.$$

Using the fact that

$$\mathbf{S}_i\cdot\mathbf{S}_j=S_{iz}S_{jz}+\tfrac{1}{2}(S_{i+}S_{j-}+S_{i-}S_{j+})$$

and Eqs. (1.2), it is possible to compute $\Pi(H)\times F(m_1,m_2,m_3)$. We find $\Pi(H)F(j_1,j_2,j_3)=0$. Hence, zero is in the spectrum of H. Let $h_1=F(j_1-1,j_2,j_3)$, $h_2=F(j_1,j_2-1,j_3)$, and $h_3=F(j_1,j_2,j_3-1)$. A straightforward computation shows that $\Pi(H)$ maps the h's into linear combinations of the h's and the (3×3)-matrix $(h_i,\Pi(H)h_j)$ is given by

$$T=(h_i,\Pi(H)h_j)=\begin{pmatrix} j_1^{-1}(a+b) & -(j_1j_2)^{-1/2}a & -(j_1j_3)^{-1/2}b\\ -(j_1j_2)^{-1/2}a & j_2^{-1}(a+c) & -(j_2j_3)^{-1/2}c\\ -(j_1j_3)^{-1/2}b & -(j_2j_3)^{-1/2}c & j_3^{-1}(b+c)\end{pmatrix}.$$

The (3×3)-matrix T can be written in the form $T=RAR$, where

$$R=\begin{pmatrix} j_1^{-1/2} & 0 & 0\\ 0 & j_2^{-1/2} & 0\\ 0 & 0 & j_3^{-1/2}\end{pmatrix},$$

$$A=\begin{pmatrix} a+b & -a & -b\\ -a & a+c & -c\\ -b & -c & b+c\end{pmatrix}.$$

Since R is an invertable positive matrix $T=RAR$ is positive if and only if A is positive. A straightforward computation shows that the eigenvalues of A are $\{0,a+b+c\pm(a^2+b^2+c^2-ab-ac+bc)^{1/2}\}$. In order that the two numbers $a+b+c\pm(a^2+b^2+c^2-ab-ac-bc)^{1/2}$ both be nonnegative, it is necessary and sufficient that inequalities (3.1) be satisfied. Hence, $H\geq 0$ implies inequalities (3.1).

Next suppose inequalities (3.1) are satisfied. We will show $H\geq 0$. We begin by assuming $j_1=j_2=j_3=\frac{1}{2}$. Let the representation Π on $\mathcal{H}$ be as we have constructed and let $h_1,h_2,h_3\in\mathcal{H}$ be defined as before. Let $\mathcal{M}_{+1}$ be the subspace of $\mathcal{H}$ spanned by these three vectors. We have $\Pi(H)\mathcal{M}_{+1}\subset\mathcal{M}_{+1}$ and $\Pi(H)$ is positive on $\mathcal{M}_{+1}$ by the above calculation. Let $k_1=F(\frac{1}{2},-\frac{1}{2},-\frac{1}{2})$, $k_2=F(-\frac{1}{2},\frac{1}{2},-\frac{1}{2})$, and $k_3=F(-\frac{1}{2},-\frac{1}{2},\frac{1}{2})$. A straightforward computation shows that $\Pi(H)$ maps the k's into linear combinations of the k's and the (3×3)-matrix $(k_i,\Pi(H)k_j)$ equals the (3×3)-matrix $(h_i,\Pi(H)h_j)$. Let $\mathcal{M}_{-1}$ be the subspace of $\mathcal{H}$ spanned by k_1, k_2, and k_3. Hence, we have $\Pi(H)\mathcal{M}_{-1}\subset\mathcal{M}_{-1}$ and $\Pi(H)$ is positive on $\mathcal{M}_{-1}$. Let $\mathcal{M}_{+3}$ be the subspace of $\mathcal{H}$ spanned by $F(\frac{1}{2},\frac{1}{2},\frac{1}{2})$ and let $\mathcal{M}_{-3}$ be the subspace of $\mathcal{H}$ spanned by $F(-\frac{1}{2},-\frac{1}{2},-\frac{1}{2})$. We have $\Pi(H)F(\frac{1}{2},\frac{1}{2},\frac{1}{2})=0$ and $\Pi(H)F(-\frac{1}{2},-\frac{1}{2},-\frac{1}{2})=0$. Hence, $\Pi(H)$ maps each of the subspaces $\mathcal{M}_{+3}$, $\mathcal{M}_{+1}$, $\mathcal{M}_{-1}$, $\mathcal{M}_{-3}$ into itself and $\Pi(H)$ is positive on each of these subspaces. Since $\mathcal{H}=\mathcal{M}_{+3}\oplus\mathcal{M}_{+1}\oplus\mathcal{M}_{-1}\oplus\mathcal{M}_{-3}$, we have $\Pi(H)$ is positive. Hence, $H\geq 0$ for $j_1=j_2=j_3=\frac{1}{2}$.

We will complete the proof of the lemma by induction. Suppose the lemma is true for all $j_1\leq n_1$, $j_2\leq n_2$, and $j_3\leq n_3$. We will prove the lemma is true for all $j_1\leq n_1+\frac{1}{2}$, $j_2\leq n_2$, and $j_3\leq n_3$. Let $\mathfrak{A}_0$ be the spin-$\frac{1}{2}$ algebra, i.e., $\mathfrak{A}_0$ is a (2×2)-matrix algebra generated by the three elements $\mathbf{S}_0=(S_{0x},S_{0y},S_{0z})$ satisfying Eq. (1.1) with $j=\frac{1}{2}$. Let $\mathbf{s}_0=2\mathbf{S}_0$. Let $\mathcal{B}=\mathfrak{A}_0\otimes\mathfrak{A}=\mathfrak{A}_0\otimes\mathfrak{A}_1\otimes\mathfrak{A}_2\otimes\mathfrak{A}_3$, where $\mathfrak{A}_i$ are the spin j_i algebras as before. Suppose a, b, and c are real number satisfying inequalities (3.1). Since the lemma is assumed true for $j_1=\frac{1}{2}$ or $j_1\leq n_1$, $j_2\leq n_2$, and $j_3\leq n_3$, we have for the j_i satisfying these inequalities

$$a(I-\mathbf{s}_1\cdot\mathbf{s}_2)+b(I-\mathbf{s}_1\cdot\mathbf{s}_3)+c(I-\mathbf{s}_2\cdot\mathbf{s}_3)\geq 0,$$

$$a(I-\mathbf{s}_0\cdot\mathbf{s}_2)+b(I-\mathbf{s}_0\cdot\mathbf{s}_3)+c(I-\mathbf{s}_2\cdot\mathbf{s}_3)\geq 0.$$

Multiplying the first inequality by j_1 and the second by $\frac{1}{2}$ and adding the resutant inequalities, we find

$$a[(j_1+\tfrac{1}{2})I-(\mathbf{S}_1+\mathbf{S}_0)\cdot\mathbf{s}_2]+b[(j_1+\tfrac{1}{2})I-(\mathbf{S}_1+\mathbf{S}_0)\cdot\mathbf{s}_3]+c(j_1+\tfrac{1}{2})(I-\mathbf{s}_2\cdot\mathbf{s}_3)\geq 0.$$

Let $\mathfrak{A}_1'$ be the C^*-algebra generated by the three elements $\mathbf{S}_1'=\mathbf{S}_1+\mathbf{S}_0$. As we have seen in Sec. 1, $\mathfrak{A}_1'$ has a minimal central profection E so that $\mathbf{S}_1'^2E=(j_1+\frac{1}{2})\times(j_1+1\frac{1}{2})E$. Let $\mathcal{B}_1=\mathfrak{A}_1'E$. We have that $\mathcal{B}_1$ is a spin-$(j_1+\frac{1}{2})$ algebra. Since E commutes with $\mathbf{S}_1'$, $\mathbf{S}_2$, and $\mathbf{S}_3$, the above inequality remains true when multiplied on the right or left by E. Let $\mathbf{s}_1'=(\mathbf{S}_1+\mathbf{S}_0)/(j_1+\frac{1}{2})=\mathbf{S}_1'/(j_1+\frac{1}{2})$. Multiplying the above inequality by E on the left and dividing by $j_1+\frac{1}{2}$, we find

$$(a(I-\mathbf{s}_1'\cdot\mathbf{s}_2)+b(I-\mathbf{s}_1'\cdot\mathbf{s}_3)+c(I-\mathbf{s}_2\cdot\mathbf{s}_3))E\geq 0.$$

Since the algebra generated by $\mathbf{s}_1'E$, $\mathbf{s}_2E$, and $\mathbf{s}_3E$ is the Heisenberg spin algebra describing three particles of spin $j_1+\frac{1}{2}$, j_2, and j_3 we have shown that if the lemma is true for $j_1\leq n_1$, $j_2\leq n_2$, and $j_3\leq n_3$, then the lemma is true for $j_1\leq n_1+\frac{1}{2}$, $j_2\leq n_2$, and $j_3\leq n_3$. Since in our proof we could have just as well have increased n_2 or n_3 by $\frac{1}{2}$, it follows by induction that since the lemma is true for $j_1=j_2=j_3=\frac{1}{2}$ the lemma is true for all $j_1,j_2,j_3=\frac{1}{2},1,1\frac{1}{2},\cdots$. This completes the proof of the lemma.

Lemma 3.2: Let $\mathfrak{A}$ be the Heisenberg spin algebra describing $n+1$ particles of spin j_i, $i=1,\ldots,n+1$, i.e., $\mathfrak{A}$ is a C^*-algebra generated by the Hermitian elements $\mathbf{S}_i=(S_{ix},S_{iy},S_{iz})$ with the S_i satisfying Eqs. (1.1) and the $\mathbf{S}_i$ and $\mathbf{S}_j$ commuting for $i\neq j$. Suppose $a_i>0$ for $i=1,\ldots,n$ and $\mathbf{s}_i=\mathbf{S}_i/j_i$ for $i=1,\ldots,n+1$. Then

$$I-\mathbf{s}_1\cdot\mathbf{s}_{n+1}\leq\sum_{i=1}^{n}a_i^{-1}\sum_{i=1}^{n}a_i(I-\mathbf{s}_i\cdot\mathbf{s}_{i+1}).\qquad(3.2)$$

Proof: For the case $n=2$ inequality (3.2) states

$$(1+a_1/a_2)(I-\mathbf{s}_1\cdot\mathbf{s}_2)-(I-\mathbf{s}_1\cdot\mathbf{s}_3)+(1+a_2/a_1)(I-\mathbf{s}_2\cdot\mathbf{s}_3)\geq 0.$$

The numbers $a=1+a_1/a_2$, $b=-1$, and $c=1+a_2/a_1$ satisfy inequalities (3.1). Hence, the above inequality is valid and the lemma is true for $n=2$.

We complete the proof of the lemma by induction. Suppose the lemma is true for n. We prove it is true for $n+1$. Let $R_n=\sum_{i=1}^n a_i^{-1}$ and $R_{n+1}=R_n+a_{n+1}^{-1}$. Since the lemma is true for $n=2$, we have

$$(I-\mathbf{s}_1\cdot\mathbf{s}_{n+2})\leq R_{n+1}(R_n^{-1}(I-\mathbf{s}_1\cdot\mathbf{s}_{n+1})+a_{n+1}(I-\mathbf{s}_{n+1}\cdot\mathbf{s}_{n+2})).$$

Since the lemma is true for n, we have

$$R_n^{-1}(I-\mathbf{s}_1\cdot\mathbf{s}_{n+1})\leq\sum_{i=1}^{n}a_i(I-\mathbf{s}_i\cdot\mathbf{s}_{i+1}).$$

Combining these inequalities, we have

$$(I-\mathbf{s}_1\cdot\mathbf{s}_{n+2})\leq\left(\sum_{i=1}^{n+1}a_i^{-1}\right)\left(\sum_{i=1}^{n+1}a_i(I-\mathbf{s}_i\cdot\mathbf{s}_{i+1})\right).$$

Hence, the lemma is true for $n+1$. Since the lemma is true for $n=2$, the lemma is true for all n by induction. This completes the proof of the lemma.

Theorem 3.3: Suppose G is a finite connected network with vertices $\mathcal{L}$ and with resistances $J(i,j)^{-1}>0$ associated with each line $(i,j)\in G$. Suppose with each vertex $i\in\mathcal{L}$ there is associated a particle of spin j_i. Let $\mathfrak{A}=\mathfrak{A}_L$ be the Heisenberg spin algebra associated with this system, i.e., $\mathfrak{A}$ is a C^*-algebra generated by the Hermitian elements $\mathbf{S}_i=(S_{ix},S_{iy},S_{iz})$ satisfying Eqs. (1.1) with $j=j_i$ and $\mathbf{S}_i$ and $\mathbf{S}_j$ commuting for $i\neq j$. Let

$$H=\sum_{(i,j)\in G}J(i,j)(I-\mathbf{s}_i\cdot\mathbf{s}_j)$$

with $\mathbf{s}_i=\mathbf{S}_i/j_i$. Then for any two $i,j\in\mathcal{L}$ we have

$$(I-\mathbf{s}_i\cdot\mathbf{s}_j)\leq R(i,j)H,$$

where $R(i,j)$ is the resistance between i and j as calculated in G.

Proof: Suppose $i_0,j_0\in\mathcal{L}$. Let V be a solution (unique up to the addition of a constant function) to the equation $-\Delta V=\delta_{i_0}-\delta_{j_0}$. As we saw in Sec. 2 the resistance between i_0 and j_0 is given by $R(i_0,j_0)=V(i_0)-V(j_0)$. Let Q be the set of current paths from i_0 to j_0, and let $s(p)$ be the probability associated with the path p as described in Sec. 2. Consider a path $p=(i_0,i_1,\ldots,i_n)\in Q$ with $i_n=j_0$. We consider p as a sum of lines (i_{k-1},i_k) for $k=1,\ldots,n$. From Lemma 3.2 we have {setting $a_k=[V(i_{k-1})-V(i_k)]^{-1}$ for $k=1,\ldots,n$}

$$(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\leq[V(i_0)-V(j_0)]\sum_{k=1}^{n}[V(i_{k-1})-V(i_k)]^{-1}\times(I-\mathbf{s}_{i_{k-1}}\cdot\mathbf{s}_{i_k})$$

or

$$(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\leq R(i_0,j_0)\sum_{(i,j)\in p}[V(i)-V(j)]^{-1}\times(I-\mathbf{s}_i\cdot\mathbf{s}_j).$$

Multiplying the above inequality by $s(p)$ and summing over all paths $p\in Q$, we find

$$(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})=\sum_{p\in Q}s(p)(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\leq R(i_0,j_0)\sum_{p\in Q}\sum_{(i,j)\in p}s(p)[V(i)-V(j)]^{-1}\times(I-\mathbf{s}_i\cdot\mathbf{s}_j).$$

Interchanging the order of summation as was described in Sec. 2, we find

$$(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\leq R(i_0,j_0)\sum_{(i,j)\in G'}\sum_{p\in Q(i,j)}s(p)[V(i)-V(j)]^{-1}\times(I-\mathbf{s}_i\cdot\mathbf{s}_j),$$

where G' is the set of all lines (i,j) of G with $V(i)\neq V(j)$ [we use the convention that the pair (i,j) is ordered so that $V(i)\geq V(j)$]. Note the lines with $V(i)=V(j)$ do not occur in the above sum since no path $p\in Q$ passes through a line $(i,j)\in G$ with $V(i)=V(j)$. From Lemma 2.1 we have $\sum_{p\in Q(i,j)}s(p)=I(i,j)=J(i,j)[V(i)-V(j)]$. Hence, we from this fact and the above inequality

$$(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\leq R(i_0,j_0)\sum_{(i,j)\in G'}J(i,j)(I-\mathbf{s}_i\cdot\mathbf{s}_j)\leq R(i_0,j_0)\sum_{(i,j)\in G}J(i,j)(I-\mathbf{s}_i\cdot\mathbf{s}_j)=R(i_0,j_0)H,$$

where the second inequality follows from the fact that the terms $J(i,j)(I-\mathbf{s}_i\cdot\mathbf{s}_j)$ for $(i,j)\in G-G'$ are all positive elements of $\mathfrak{A}$. This completes the proof of the theorem.

Corollary 3.4: Suppose $G,\mathcal{L},\mathfrak{A}=\mathfrak{A}_1$ and $\mathbf{S}_i,\mathbf{s}_i$ for $i\in\mathcal{L}$ are as stated in Theorem 3.3 except now we allow G and $\mathcal{L}$ to be countably infinite. Suppose $i_0,j_0\in\mathcal{L}$ and the resistance $R(i_0,j_0)>0$ is not zero. Then for any state ω of $\mathfrak{A}$ we have

$$\omega(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\leq R(i_0,j_0)\sum_{(i,j)\in G}J(i,j)\omega(I-\mathbf{s}_i\cdot\mathbf{s}_j),$$

where the right-hand side of this inequality may be infinite.

Proof: Let $\{G_n\}$ be an increasing sequence of finite connected subnetworks of G containing i_0 and j_0 as vertices and the sequence $\{G_n\}$ increases up to G as $n\to\infty$. For each n we have by Theorem 3.3 and the fact that $\omega(A)\leq\omega(B)$ for ω a state of $\mathfrak{A}$ and A and B Hermitian elements of $\mathfrak{A}$ with $A\leq B$

$$\omega(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\leq R_n(i_0,j_0)\sum_{(i,j)\in G_n}J(i,j)\omega(I-\mathbf{s}_i\cdot\mathbf{s}_j).$$

where $R_n(i_0,j_0)$ is the resistance between i_0 and j_0 as calculated in G_n. As we saw in Sec. 2, as $n\to\infty$, $R_n(i_0,j_0)\to R(i_0,j_0)$. As $n\to\infty$, the sum over $(i,j)\in G_n$ converges to the sum over all $(i,j)\in G$ or diverges to $+\infty$. This completes the proof of the corollary.

Theorem 3.5: Let G be the network with vertices $\mathcal{L}=Z^3$ with the lines of G connecting all nearest neighbors, i.e., if $i=(i_x,i_y,i_z)$ and $j=(j_x,j_y,j_z)$, then $(i,j)\in G$ if and only if $|i-j|=|i_x-j_x|+|i_y-j_y|+|i_z-j_z|=1$. Let the resistance $J(i,j)$ associated with each line $(i,j)\in G$ be one ohm, i.e., $J(i,j)=1$ for all lines of G. Suppose that with each vertex $i\in\mathcal{L}$ there is associated a particle of spin $j_i=\frac{1}{2},1,1\frac{1}{2},2,\cdots$. Let $\mathfrak{A}=\mathfrak{A}_{\mathcal{L}}$ be the Heisenberg spin algebra associated with this system, i.e., $\mathfrak{A}$ is a C^*-algebra generated by the Hermitian elements $\mathbf{S}_i=(S_{ix},S_{iy},S_{iz})$ satisfying Eq. (1.1) with $j=j_i$ and the $\mathbf{S}_i$ and $\mathbf{S}_j$ commute for $i\neq j$. Let $\mathbf{s}_i=\mathbf{S}_i/j_i$ for $i\in\mathcal{L}$. Suppose ω is a state of $\mathfrak{A}$ of finite energy, i.e.,

$$\sum_{(i,j)\in G}\omega(I-\mathbf{s}_i\cdot\mathbf{s}_j)<\infty.$$

Then for every number $\epsilon>0$

$$\omega(I-\mathbf{s}_i\cdot\mathbf{s}_j)<\epsilon$$

for all $i,j\in\mathcal{L}$ with at most a finite number of exceptions.

Proof: Let $R_\infty=\sup\{R(i,j);i,j\in\mathcal{L}\}$, where $R(i,j)$ is the resistance between i and j as calculated in G. As we have seen in Sec. 2, $R_\infty=0.50546\cdots$ is finite. Suppose $\epsilon>0$. Suppose ω is a state of finite energy and $E=\sum_{(i,j)\in G}\omega(I-\mathbf{s}_i\cdot\mathbf{s}_j)$. Since the sum is finite there are a finite number of lines $(r_k,s_k)\in G$ for $k=1,\ldots,n$ so that

$$\sum_{k=1}^{n}\omega(I-\mathbf{s}_{r_k}\cdot\mathbf{s}_{s_k})>E-\tfrac{1}{2}\epsilon/R_\infty.$$

Let G' be the network obtained from G by removing the lines (r_k,s_k) for $k=1,\ldots,n$. Then, we have

$$\sum_{(i,j)\in G'} \omega(I-\mathbf{s}_i\cdot\mathbf{s}_j) < \tfrac{1}{2}\epsilon/R_\infty.$$

It follows from Lemma 2.2 of Sec. 2 that there exists a finite set S so that for $i,j\in \mathcal{L}-S$ the resistance $R'(i,j)$ between i and j as calculated in G' is less than $2R_\infty$, i.e., $R'(i,j)<2R_\infty$ for $i,j\in\mathcal{L}-S$. From Corollary 3.4 we have for $i_0,j_0\in\mathcal{L}-S$

$$\omega(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0}) \le R'(i_0,j_0) \sum_{(i,j)\in G'} \omega(I-\mathbf{s}_i\cdot\mathbf{s}_j) < 2R_\infty(\tfrac{1}{2}\epsilon/R_\infty)=\epsilon.$$

Hence, $\omega(I-\mathbf{s}_i\cdot\mathbf{s}_j)<\epsilon$ for $i,j\in\mathcal{L}-S$. This completes the proof of the theorem.

Theorem 3.5 shows that all states of finite energy have long range order in three dimensions. We remark that the theorem is false in one and two dimensions.

It is known from the work of Mermin and Wagner[10] that thermal equilibrium states of the Heisenberg model at positive temperature do not have long range order in one and two dimensions. We hope that the notaion of resistance will be useful in understanding the presents of long range order for equilibrium states of the Heisenberg model in three dimensions.

We conclude this paper by showing the constant $R(i,j)$ in Theorem 3.3 is the best possible.

Theorem 3.6: Suppose $G,\mathcal{L},\mathfrak{A}=\mathfrak{A}_\mathcal{L}$ and $\mathbf{S}_i$, $\mathbf{s}_i$, and H are as given in the statement of Theorem 3.3. Suppose $i,j\in\mathcal{L}$ and C is a real number. Then,

$$I-\mathbf{s}_i\cdot\mathbf{s}_j \le CH \tag{3.3}$$

if and only if $C\ge R(i,j)$, where $R(i,j)$ is the resistance between i and j as calculated in G.

Proof: For $C\ge R(i,j)$ inequality (3.3) follows immediately from Theorem 3.3 and the positively of H. Suppose, then, that C is real and inequality (3.3) is satisfied. We will show that $C\ge R(i,j)$.

Let $\mathfrak{A}_i$ be the Heisenberg spin algebra associated with the vertex i. The C^*-algebra $\mathfrak{A}_i$ is generated by the Hermitian elements $\mathbf{S}_i=(S_{ix},S_{iy},S_{iz})$ and has a faithful irreducible $*$-representation Π on a $(2j_i+1)$-dimensional Hilbert space $\mathcal{H}$. Let $\{f_m; m=-j_i,1-j_i,\ldots,j_i-1,j_i\}$ be an orthomormal basis for $\mathcal{H}$ so that the operators $\Pi(\mathbf{S}_i)$ satisfy Eqs. (1.2). Let ω be the state of $\mathfrak{A}_i$ given by $\omega(A)=(f_{j_i},\Pi(A)f_{j_i})$ for $A\in\mathfrak{A}_i$. Let $\mathbf{s}_i=\mathbf{S}_i/j_i$. A straightforward computation shows that $\omega(\mathbf{s}_i)=(\omega(s_{ix}),\omega(s_{iy}),\omega(s_{iz}))=(0,0,1)$. Furthermore, since the eigenvalue 1 for $\Pi(s_{iz})$ had multiplicity one if ω' is a state of $\mathfrak{A}_i$ and $\omega'(\mathbf{s}_i)=(0,0,1)$ then $\omega'=\omega$.

Let $\mathbf{n}\in R^3$ be a unit vector, i.e., $|n|=(n_x^2+n_y^2+n_z^2)^{1/2}=1$. Since the element $\mathbf{n}\cdot\mathbf{s}=n_xs_{ix}+n_ys_{iy}+n_zs_{iz}$ is unitarily equivalent to s_{iz} (in fact, there is a unitary element $U=\exp(i\mathbf{a}\cdot\mathbf{S}_i)\in\mathfrak{A}_i$ so that $Us_{iz}U^*=\mathbf{n}\cdot\mathbf{s}_i$) there is a unique state $\omega_\mathbf{n}$ of $\mathfrak{A}_i$ so that $\omega_\mathbf{n}(\mathbf{s}_i)=(n_x,n_y,n_z)$.

Suppose for each $i\in\mathcal{L}$ we specify a unit vector $\mathbf{n}_i$. We define the state $\omega_{\{n\}}$ on $\mathfrak{A}=\mathfrak{A}_\mathcal{L}$ as $\omega_{\{n\}}=\otimes_{i\in L}\,\omega_{\mathbf{n}_i}$ the tensor product of the states $\omega_{\mathbf{n}_i}$, i.e., if $A_k\in\mathfrak{A}_k$, $\mathcal{L}$ then

$$\omega_{\{\mathbf{n}\}}(A_iA_j\cdots A_r)=\omega_{\mathbf{n}_i}(A_i)\omega_{\mathbf{n}_j}(A_j)\cdots\omega_{\mathbf{n}_r}(A_r).$$

A straightforward computation shows that for $i,j\in\mathcal{L}$

$$\omega_{\{\mathbf{n}\}}(I-\mathbf{s}_i\cdot\mathbf{s}_j)=1-\mathbf{n}_i\cdot\mathbf{n}_j$$

and

$$\omega_{\{\mathbf{n}\}}(H)=\sum_{(i,j)\in G} J(i,j)(1-\mathbf{n}_i\cdot\mathbf{n}_j).$$

Suppose $i_0,j_0\in\mathcal{L}$ and V is a real solution to the equation $-\Delta V=\delta_{i_0}-\delta_{j_0}$. Suppose $s>0$ is a positive number and let the unit vectors $\mathbf{n}_i$ be given by

$$n_{iz}=0,\quad n_{ix}=\cos[sV(i)],\quad n_{iy}=\sin[sV(i)]$$

for all $i\in\mathcal{L}$. Since inequality (3.3) is assumed true and $\omega_{\{\mathbf{n}\}}$ assigns nonnegative numbers to positive elements of we have

$$\omega_{\{\mathbf{n}\}}(I-\mathbf{s}_{i_0}\cdot\mathbf{s}_{j_0})\le C\omega_{\{\mathbf{n}\}}(H)$$

or

$$1-\cos[s(V(i_0)-V(j_0))] \le C\sum_{(i,j)\in G} J(i,j)\{1-\cos[s(V(i)-V(j))]\}.$$

Multiplying both sides of the above inequality by $2/s^2$ and taking the limit as $s\to 0$, we find

$$[V(i_0)-V(j_0)]^2\le C\sum_{(i,j)\in G} J(i,j)[V(i)-V(j)]^2$$

or

$$R(i_0,j_0)^2\le C\sum_{(i,j)\in G} J(i,j)^{-1}I(i,j)^2=D(\{I\})C,$$

where $I(i,j)=J(i,j)[V(i)-V(j)]$. As we saw in Sec. 2, $R(i_0,j_0)=D(\{I\})$. Hence, the above inequality implies $C\ge R(i_0,j_0)$. This completes the proof of the theorem.

ACKNOWLEDGMENT

The author's understanding of the ways for calculating and estimating the resistance of electrical networks discussed in Sec. 2 of this paper grew largely from several stimulating conversations with Professor E. Calabi.

*Work supported in part by the National Science Foundation.

†On leave from the University of Pennsylvania, Philadelphia, Pennsylvania.

[1]D. Ruelle, *Statistical Mechanics Rigorous Results* (Benjamin, New York, 1969).
[2]S. Sakai, *C*-algebras and W*-algebras* (Springer, New York, 1971).
[3]E. P. Wigner, *Group Theory and Its Applications to the Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959).
[4]O. Lanford, *Cargese Lectures in Physics*, edited by D. Kastler (Gordon and Breach, New York, 1970), Vol. 4.
[5]R.F. Streater, Comm. Math. Phys. **6**, 233 (1967).
[6]J. Glimm, Trans. Amer. Math. Soc. **95**, 318 (1960).
[7]R.T. Powers and S. Sakai, Comm. Math. Phys. **39**, 273 (1975).
[8]S. Seshu and M. B. Reed, *Linear Graphs and Electrical Networks* (Addison-Wesley, Reading, Mass., 1961).
[9]F. Spitzer, *Principles of Random Walk* (Van Nostrand, Princeton, N.J., 1964).
[10]N.D. Mermin and H. Wagner, Phys. Rev. Lett. **17**, 1133 (1966).

Lie algebras in the Schrödinger picture and radial matrix elements

E. Chacón,* D. Levi,† and M. Moshinsky*‡

Instituto de Física, Universidad de México (UNAM), México 20, D.F., Mexico
(Received 8 March 1976)

The main objective of this paper is to derive from a unified viewpoint particular one- and two-body radial matrix elements with respect to oscillator and Coulomb states. All the results have been obtained previosly using either generating functions of Laguerre polynomials or group theoretical methods related to particular realizations of the Lie algebras associated with those states. We show, though, in this paper how some of the realizations proposed can be derived from physical considerations. The main idea is to translate well-known realizations, in the Heisenberg picture, to the corresponding ones in the Schrödinger picture. The latter realizations allow us to define indecomposable (i.e., not completely reducible) and irreducible tensors of the Sp(2) [or equivalently the SU(1,1)] group for the one- and two-body problem respectively. The evaluations of the radial matrix elements becomes then just a matter of applying the Wigner–Eckart theorem, giving rise to Wigner coefficients of SU(1,1) that have been discussed extensively in the literature.

1. INTRODUCTION AND SUMMARY

Since the development of quantum mechanics a great deal of work has been expended in showing that matrix elements, that were evaluated through considerable effort of analysis, are actually connected with group theoretical concepts such as Wigner, Racah or $9j$ coefficients of certain groups and, in particular, of $\mathcal{O}(3)$. The best known examples concern the one- and two-body angular matrix elements

$$\langle l'm' | Y^k_q(\theta, \varphi) | lm\rangle, \tag{1.1a}$$

$$\langle l_1' l_2' L' M' | [Y^{k_1}(\theta_1, \varphi_1) Y^{k_2}(\theta_2, \varphi_2)]^k_q | l_1 l_2 LM\rangle, \tag{1.1b}$$

where Y^k_q is a spherical harmonic and $[\;]^k_q$ means coupling of the two irreducible tensors associated with angular coordinates 1 and 2 to a total k and projection q.

When we turn to the radial part of the one- and two-body matrix elements, with respect to harmonic oscillator or Coulomb states, we find that the problem has received considerable attention.[1-6] In some of the references on the subject the approach follows what could be considered a physically natural path. For example, in the calculation of the radial integrals of powers of r with respect to harmonic oscillator wavefunctions, Quesne and Moshinsky[6] start from the dynamical group Sp(6) of the three-dimensional oscillator and its subgroup

$$\mathrm{Sp}(2) \times \mathcal{O}(3), \tag{1.2}$$

where $\mathcal{O}(3)$ is the ordinary three-dimensional orthogonal group and the two-dimensional symplectic group Sp(2) has generators that are linear combinations of $\mathbf{p}^2$, $\mathbf{r}^2$, and $\mathbf{r}\cdot\mathbf{p}$.[6] Both the states of the oscillator and the solid spherical harmonics,

$$\mathcal{Y}^k_q(\mathbf{r}) = r^k Y^k_q(\theta, \varphi), \tag{1.3}$$

can then be expressed as a bases for irreducible representations or irreducible tensors of both Sp(2) and $\mathcal{O}(3)$, and the radial integrals are then given in terms of Wigner coefficients of Sp(2) or, equivalently, of SU(1,1).[6]

The approach followed by Armstrong[3,4] for one body radial integrals seems, at first sight, more mathematical. It uses a realization of the Lie algebra of SU(1,1) proposed by Miller[7] in which besides the radial variable r, another one, designated by t and apparently without physical meaning, also appears. Armstrong then introduces functions of r and t, which from now on we shall designate as Armstrong tensors, whose commutation properties with the generators of SU(1,1) are simple. The matrix elements of Armstrong tensors with respect to bases for irreducible representations (BIR) of SU(1,1) in the Miller realization, turn out to be then the radial integrals we are interested in and they can be evaluated in a group theoretical fashion.

An objective of this paper is to provide a physical justification of Armstrong's approach, as well as for generalizations of it used by Crubellier[2] in the analysis of two-body radial matrix elements. This requires, as we show in Sec. 2, the derivation of the Lie algebras of the Sp(2) group mentioned above, not in the usual Heisenberg but in the Schrödinger picture, in which the parameter t mentioned in the previous paragraph can then be identified with the time.[1]

Furthermore, we discuss in Sec. 3 the indecomposable (i.e., not fully reducible) character of the Armstrong tensor and show that the one-body radial matrix elements can then be evaluated using essentially the Wigner–Eckart theorem for the Sp(2) or, equivalently, the SU(1,1) group.

In Sec. 4 we direct our attention to two-body radial matrix elements and show their relations with those of tensors of SU(1,1), in this case irreducible ones, when the Lie algebra is given in the Schrödinger picture. Again we can use the Wigner–Eckart theorem for the SU(1,1) group in deriving the two-body radial matrix elements.

The discussion in Secs. 2, 3, 4 is always carried in relation with the m-dimensional harmonic oscillator. The particularization to $m=3$ in Sec. 5 gives the physical case, while in Sec. 6 we show that the radial integrals with Coulomb wavefunctions correspond to the oscillator ones with $m=4$.

2. LIE ALGEBRA IN THE SCHRÖDINGER PICTURE OF THE m-DIMENSIONAL OSCILLATOR

As is well known[6] the m-dimensional oscillator has as dynamical group the symplectic group in $2m$—dimensions Sp$(2m)$ whose generators are

$$x_i x_j,\quad p_i p_j,\quad \tfrac{1}{2}(x_i p_j + p_j x_i),\quad i,j=1,2,\ldots,m. \tag{2.1}$$

The group contains as a subgroup the direct product

$$\mathrm{Sp}(2)\times \mathcal{O}(m), \tag{2.2}$$

where $\mathcal{O}(m)$ is the rotation group in m-dimensional space and the generators of Sp(2) are the scalars, with respect to $\mathcal{O}(m)$, we can form from (2.1), i.e.,

$$T_1=\tfrac{1}{4}(p^2-r^2),\quad T_2=\tfrac{1}{4}(\mathbf{r}\cdot\mathbf{p}+\mathbf{p}\cdot\mathbf{r}),$$
$$T_3=\tfrac{1}{4}(p^2+r^2)=\tfrac{1}{2}H. \tag{2.3}$$

In (2.3), $\mathbf{r}$, $\mathbf{p}$, are m-dimensional vectors and we use units in which $\hbar$, the mass of the particle and the frequency of the oscillator are 1. Furthermore, H is the Hamiltonian of the oscillator. The generators of Sp(2) satisfy the commutation relations

$$[T_1,T_2]=-iT_3,\quad [T_3,T_1]=iT_2,\quad [T_2,T_3]=iT_1. \tag{2.4}$$

Turning now our attention to $\mathcal{O}(m)$, its generators are[6]

$$\mathcal{L}_{ij}=x_i p_j - x_j p_i,\quad i,j=1,2,\ldots,m, \tag{2.5}$$

and its Casimir operator becomes

$$\mathcal{L}^2\equiv \tfrac{1}{2}\sum_{i,j=1}^{m}\mathcal{L}_{ij}\mathcal{L}_{ij}$$
$$=-(\mathbf{r}\cdot\mathbf{p})^2+i(m-2)(\mathbf{r}\cdot\mathbf{p})+r^2p^2, \tag{2.6}$$

where we only make use of the relations $[x_j,p_k]=i\delta_{jk}$. From (2.6) it is clear that all homogeneous polynomials $P(\mathbf{r})$ of degree κ, that satisfy the Laplace equation, i.e., $\nabla^2 P=0$, are eigenstates of $\mathcal{L}^2$ with eigenvalue

$$\kappa(\kappa+m-2). \tag{2.7}$$

Again making use of the commutation relation $[x_j,p_k]=i\delta_{jk}$ we see that the Casimir operator of Sp(2) becomes

$$T^2\equiv T_1^2+T_2^2-T_3^2=-\tfrac{1}{4}[\mathcal{L}^2+m(m-4)/4] \tag{2.8}$$

and its eigenvalue is then

$$-\tfrac{1}{4}[\kappa(\kappa+m-2)+m(m-4)/4]. \tag{2.9}$$

From (2.6) we also see that p^2 can be determined in terms of $r^{-2}\mathcal{L}^2$ and $r\partial/\partial r$, and thus the generators T_i can be written in operator form as

$$T_3=\frac{1}{4}\left(-\frac{1}{r^{m-1}}\frac{\partial}{\partial r}r^{m-1}\frac{\partial}{\partial r}+\frac{\mathcal{L}^2}{r^2}+r^2\right), \tag{2.10a}$$

$$T_\pm=T_1\pm iT_2=T_3+\frac{1}{4}\left(-2r^2\pm 2r\frac{\partial}{\partial r}\pm m\right). \tag{2.10b}$$

The radial function $R(r)$ is the eigenfunction of $H=2T_3$ in (2.10a) corresponding to the eigenvalue (2.7) of $\mathcal{L}^2$. To determine it more explicitly let us write

$$R(r)\equiv r^{(1-m)/2}f(r), \tag{2.11}$$

and thus from (2.10) we get

$$T_3R(r)=r^{(1-m)/2}I_3f(r), \tag{2.12a}$$

$$T_\pm R(r)=r^{(1-m)/2}I_\pm f(r), \tag{2.12b}$$

where

$$I_3=\frac{1}{4}\left[-\frac{\partial^2}{\partial r^2}+\frac{(\kappa+(m-2)/2)^2-\frac{1}{4}}{r^2}+r^2\right], \tag{2.13a}$$

$$I_\pm=I_3+\frac{1}{4}\left(-2r^2\pm 2r\frac{\partial}{\partial r}\pm 1\right). \tag{2.13b}$$

In Ref. 8 we showed that the normalized eigenfunction f of I_3, which we designate by $f_n^\nu(r)$, corresponds to the eigenvalue

$$I_3f_n^\nu(r)=[n+\tfrac{1}{2}(\nu+1)]f_n^\nu(r), \tag{2.14}$$

where

$$\nu=\kappa+\tfrac{1}{2}(m-2) \tag{2.15}$$

and that its explicit form is

$$f_n^\nu(r)=[2(n!)/\Gamma(n+\nu+1)]^{1/2}e^{-r^2/2}r^{\nu+1/2}L_n^\nu(r^2), \tag{2.16}$$

where L_n^ν is an associated Laguerre polynomial.[9] Thus the radial wavefunction $R(r)$ can be written as

$$R_\mu^\lambda(r)\equiv r^{(1-m)/2}f_n^\nu(r), \tag{2.17}$$

where, for later convenience, we characterize it by the eigenvalue μ of I_3 in (2.14) and the lowest value λ of this eigenvalue, i.e.,

$$\mu=n+\tfrac{1}{2}(\nu+1),\quad \lambda=\tfrac{1}{2}(\nu+1) \tag{2.18}$$

with ν given by (2.15). As a final point we note[8] that I_3 and $I_\pm$ form a Lie algebra, i.e.,

$$[I_3,I_\pm]=\pm I_\pm,\quad [I_+,I_-]=-2I_3, \tag{2.19}$$

and furthermore that

$$I_\pm f_n^\nu(r)=[(n+\nu+\tfrac{1}{2}\pm\tfrac{1}{2})(n+\tfrac{1}{2}\pm\tfrac{1}{2})]^{1/2}f_{n\pm1}^\nu(r). \tag{2.20}$$

From (2.14), (2.20), and (2.12) we then see that the operators T_3 and $T_\pm$, in which $\mathcal{L}^2$ is replaced by $\kappa(\kappa+m-2)$, when applied to the radial wavefunctions $R_\mu^\lambda(r)$ give

$$T_3R_\mu^\lambda=\mu R_\mu^\lambda, \tag{2.21a}$$

$$T_\pm R_\mu^\lambda=[(\mu\pm\lambda)(\mu\mp\lambda\pm1)]^{1/2}R_{\mu\pm1}^\lambda. \tag{2.21b}$$

The set of radial states R_μ^λ in which

$$\mu=\lambda,\ \lambda+1,\ \lambda+2,\ldots, \tag{2.22}$$

is thus a basis for an irreducible representation (BIR) of the Sp(2) [or equivalently the SU(1,1)[10]] group characterized by the lowest value λ of the μ given by (2.18) and (2.15). The representation is infinite dimensional and unitary.[10]

In all of the previous analysis we have implicitly made use of the variables $\mathbf{r}$, $\mathbf{p}$ in the Heisenberg picture.[11] Thus functions of these variables such as the T_i of (2.3) change with time in accordance with the Heisenberg equations of motion,[12]

$$\frac{dT_3}{dt}=-i[T_3,H]=0, \tag{2.23a}$$

$$\frac{dT_\pm}{dt}=-i[T_\pm,H]=-2i[T_\pm,T_3]=\pm 2iT_\pm, \tag{2.23b}$$

where we made use of the relation $T_3=\frac{1}{2}H$ and the commutations relations (2.4). From (2.23) we can express T_i in terms of their initial values T_i^0 as[1]

$$T_3=T_3^0, \quad T_\pm=T_\pm^0 e^{\pm 2it}. \tag{2.24}$$

The operators T_3^0, $T_\pm^0$ will be the corresponding operators in the Schrödinger picture.[1] We can make use of (2.24) to express them in terms of the T_i. We note that the T_i^0 are applied to the Schrödinger eigenstates of H which contain the extra factor[1]

$$e^{-iEt}=e^{-2\mu t}=e^{-(2n+\nu+1)t}. \tag{2.25}$$

Thus in T_i we can replace

$$\tfrac{1}{2}(p^2+r^2)\to i\frac{\partial}{\partial t}, \tag{2.26}$$

and from (2.24) the generators of Sp(2) in the Schrödinger picture become[1]

$$T_3^0=\frac{i}{2}\frac{\partial}{\partial t}, \tag{2.27a}$$

$$T_\pm^0=\tfrac{1}{2}e^{\mp i2t}\left(i\frac{\partial}{\partial t}-r^2\pm r\frac{\partial}{\partial r}\pm\frac{m}{2}\right)$$
$$=e^{\mp 2it}\left[T_3^0+\frac{1}{2}\left(-r^2\pm r\frac{\partial}{\partial r}\pm\frac{m}{2}\right)\right]. \tag{2.27b}$$

Applying now T_3^0, $T_\pm^0$ to the Schrödinger radial states,[1]

$$\psi_\mu^\lambda(r,t)\equiv R_\mu^\lambda(r)e^{-i2\mu t}, \tag{2.28}$$

we obtain using (2.27) and (2.21) that

$$T_3^0\psi_\mu^\lambda=\mu\psi_\mu^\lambda, \tag{2.29a}$$

$$T_\pm^0\psi_\mu^\lambda=[(\mu\pm\lambda)(\mu\mp\lambda\pm 1)]^{1/2}\psi_{\mu\pm 1}^\lambda. \tag{2.29b}$$

Thus the functions $\psi_\mu^\lambda(r,t)$ are also BIR of Sp(2) [or equivalently of SU(1,1)[13,14]], characterized by λ, but now the Schrödinger realization of the Lie algebra is the one given by (2.27), in terms of first order operators in r,t, rather than the Heisenberg realization (2.13) which contains second order derivatives, but in r alone. We note also that $T_-^0\psi_\lambda^\lambda=0$, and thus this function of the set $\{\psi_\mu^\lambda,\ \mu=\lambda,\lambda+1,\ldots,\}$ is the one of lowest weight.

Having obtained the generators T_3^0 and $T_\pm^0$ in the Schrödinger picture we can also determine the Casimir operator T^{0^2}, which from (2.27) becomes

$$T^{02}\equiv T_2^{0^2}+T_2^{0^2}-T_3^{0^2}=T_+^0T_-^0-T_3^{0^2}+T_3^0$$
$$=-\frac{1}{4}\left[2r^2 i\frac{\partial}{\partial t}-r^4+\frac{1}{r^{m-3}}\frac{\partial}{\partial r}r^{m-1}\frac{\partial}{\partial r}+\frac{m(m-4)}{4}\right], \tag{2.30}$$

and whose eigenfunctions are the $\psi_\mu^\lambda(r,t)$, all of which correspond to the eigenvalue $-\lambda(\lambda-1)$ or equivalently to (2.9).

The volume element in (r,t) space must be selected[1] so that the operators T_3^0, T^{0^2} are Hermitian, which from (2.27a), (2.30) implies that it becomes[1]

$$r^{m-3}\,dr\,dt. \tag{2.31}$$

Furthermore, as the time dependent part of the function $\psi_\mu^\lambda(r,t)$ is periodic, the interval of integration for t can be taken as $0\le t\le 2\pi$. The corresponding interval for r remains $0\le r\le\infty$ and the functions $\psi_\mu^\lambda(r,t)$ satisfy the orthogonality relations

$$(\psi_{\mu'}^{\lambda'},\psi_\mu^\lambda)=\int_0^\infty\int_0^{2\pi}\psi_{\mu'}^{\lambda'}(r,t)\psi_\mu^\lambda(r,t)r^{m-3}\,dr\,dt$$
$$=2\pi(2\lambda-1)^{-1}\delta_{\lambda\lambda'}\delta_{\mu\mu'}. \tag{2.32}$$

The $\delta_{\lambda\lambda'}\delta\delta_{\mu\mu'}$ in (2.32) are due to the fact that the $\psi_\mu^\lambda(r,t)$ are eigenstates of the Hermitian operators T_3^0 and T^{02}. Note that in the radial part alone we get, using the definitions (2.17) and (2.18), the new orthogonality relation[1,3]

$$(\psi_{\mu'}^{\lambda'},\psi_\mu^\lambda)=2\pi\int_0^\infty r^{-2}f_{n+(\nu-\nu')/2}^{\nu'}(r)f_n^\nu(r)\,dr$$
$$=2\pi\nu^{-1}\delta_{\nu\nu'} \tag{2.33}$$

which is completely independent from the standard one in the Heisenberg picture[8]

$$\int_0^\infty f_{n'}^\nu(r)f_n^\nu(r)\,dr=\delta_{nn'}. \tag{2.34}$$

The factor $2\pi\nu^{-1}$ appearing in (2.33) can be evaluated by first considering $\nu=\nu'$ in the integral which implies $\lambda=\lambda'$. As $(\psi_\mu^\lambda,\psi_\mu^\lambda)$ is independent of μ,[15] the integral in (2.33) is independent of n and we can calculate it trivially for $n=0$ obtaining ν^{-1}.

We have established the states $\psi_\mu^\lambda(r,t)$ as BIR of Sp(2) [or equivalently of SU(1,1)[6,10]] and defined their scalar product. In the next section we introduce indecomposable, i.e., not completely reducible, tensors of Sp(2) in (r,t) space and determine their matrix elements with respect to these states with the help of the Wigner–Eckart theorem for Sp(2).

3. ONE-BODY MATRIX ELEMENTS

In view of the Eqs. (2.29) that give the effect of the operators T_3^0, $T_\pm^0$ on the basis $\psi_\mu^\lambda(r,t)$, it is clear that an irreducible tensor of the Sp(2) group in the Schrödinger picture will be a function $P_q^k(r,t)$ that satisfies the commutation relations

$$[T_3^0,P_q^k]=qP_q^k, \tag{3.1a}$$

$$[T_\pm^0,P_q^k]=[(q\pm k)(q\mp k\pm 1)]^{1/2}P_{q\pm 1}^k. \tag{3.1b}$$

In case we were able to find such functions the matrix elements

$$\int_0^\infty\int_0^{2\pi}\psi_{\lambda'\mu'}^*(r,t)P_q^k(r,t)\psi_{\lambda\mu}(r,t)r^{m-3}\,dr\,dt \tag{3.2}$$

will be given[13] in terms of Wigner coefficients of SU(1,1)[10] and a factor independent of μ',q,μ. Armstrong has discussed tensors with properties resembling those of (3.1) and leading to radial integrals of the type we want to discuss. We proceed to analyze these tensors.

A. The Armstrong tensors

Armstrong defines the following functions of r,t,

$$\mathcal{P}_q^k(r,t)\equiv r^{-2k+2}e^{-i2qt}, \tag{3.3}$$

where k takes nonpositive integer or semi-integer values, i.e.,

$$k=0,\ -\tfrac{1}{2},-1,-\tfrac{3}{2},\cdots. \tag{3.4}$$

The values for q go from $-\infty$ to $+\infty$ with the restriction $q-k$ integer.

The functions (3.3) are designated by a script letter to distinguish them from the $P_q^k(r,t)$, as from (2.27) the commutation relations for $\rho_q^k(r,t)$ are not (3.1) but

$$[T_3^0, \rho_q^k] = q\rho_q^k, \tag{3.5a}$$

$$[T_\pm^0, \rho_q^k] = (q \mp k \pm 1)\rho_{q\pm1}^k. \tag{3.5b}$$

Can the $\rho_q^k(r,t)$ be still identified as irreducible tensors of SU(1,1)? To answer this question let us first define a nonnormalized ket

$$|\lambda\mu\rangle \equiv \left[\frac{\Gamma(\mu+\lambda)}{2\Gamma(\mu-\lambda+1)}\right]^{1/2}\psi_\mu^\lambda(r,t)$$

$$= r^{2\lambda-m/2}L_{\mu-\lambda}^{2\lambda-1}(r^2)e^{-i2\mu t}. \tag{3.6}$$

From (2.29) we then immediately have

$$T_3^0|\lambda\mu\rangle = \mu|\lambda\mu\rangle, \tag{3.7a}$$

$$T_\pm^0|\lambda\mu\rangle = (\mu \mp \lambda \pm 1)|\lambda\mu \pm 1\rangle, \tag{3.7b}$$

and furthermore, $T_-|\lambda\lambda\rangle = 0$.

This result seems to suggest that had we multiplied $\rho_q^k(r,t)$ by an appropriate normalization factor, it would satisfy the commutation rules (3.1) and become an irreducible tensor. But the situation turns out to be more complex. In the case of the set of states $|\lambda\mu\rangle$, $\mu = \lambda, \lambda+1, \cdots$, any one of them can be reached from any other with the help of $T_\pm^0$ and furthermore the BIR is bounded from below. For $\rho_q^k(r,t)$ the q is unbounded. Furthermore, if we start with a q in the interval $k \le q \le -k$, we can reach $\rho_q^k(r,t)$ with any other value of q, but if we start with $q > -k$ $(q<k)$ we cannot go below (above) $q = -k+1$ $(q = k-1)$, because $T_-\rho_{-k+1}^k = 0$ $(T_+\rho_{k-1}^k = 0)$. This is illustrated in Fig. 1 where the marks at the points $-k+1$ $(k-1)$ indicate that we cannot go below (above) them in values of q because of the commutation relation (3.5b).

Thus the set $\rho_q^k(r,t)$ for all allowed q is not an irreducible tensor. We shall in fact proceed to show that it is an indecomposable one, by determining explicitly the representation to which it belongs and showing that it is not completely reducible.

B. The indecomposable character of Armstrong's tensor

We are now interested in determining explicitly the representation of Sp(2) [or equivalently SU(1,1)] to which the set of functions $\rho_q^k(r,t)$ belongs. The general element of Sp(2) is given by[1]

$$e^{i\alpha T_3^0}e^{i\beta T_2^0}e^{i\gamma T_3^0}. \tag{3.8}$$

In so far as the part of T_3^0 is concerned, we immediately obtain from (3.5a)

$$e^{i\alpha T_3^0}\rho_q^k(r,t)e^{-i\alpha T_3^0} = e^{iq\alpha}\rho_q^k(r,t)$$

so that the representation of Sp(2) is $\exp(iq\alpha)\delta_{q'q}$. It is the part related with T_2^0 that will be of concern to us, i.e.,

$$e^{i\beta T_2^0}\rho_q^k(r,t)e^{-i\beta T_2^0} = \sum_{q'}\rho_{q'}^k(r,t)D_{q'q}^k(\beta). \tag{3.9}$$

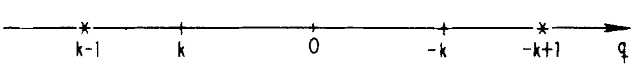


FIG. 1. Values of the component q of the indecomposable tensor $\rho_q^k(r,t)$. If q is between k and $-k$, $k = 0, -\frac{1}{2}, -1, -\frac{3}{2}, \ldots$, then all other values of q can be reached with the help of $T_\pm^0$. If $q > -k$ $(q<k)$ then there is a lower (upper) bound, $-k+1$ $(k-1)$, that can be reached which is indicated by a star.

Taking the derivative of this equation with respect to β and using the commutation relations (3.5b) we immediately obtain for the $D_{q'q}^k(\beta)$ the set of first order linear differential equations

$$\frac{d}{d\beta}D_{q'q}^k(\beta) = \tfrac{1}{2}[(q'-k)D_{q'-1,q}^k(\beta) - (q'+k)D_{q'+1,q}^k(\beta)]. \tag{3.10}$$

This set of equations together with the condition $D_{q'q}^k(0) = \delta_{q'q}$ fully determine the representation $D_{q'q}^k(\beta)$, but they are quite difficult to solve. Fortunately the same Eqs. (3.10) would appear in any realization of the Lie algebra of SU(1,1). Using one proposed by Miller[7] one arrives at the expression [see formulas (5.10) and (5.11), p. 159 of Ref. 7]

$$D_{q'q}^k(\beta) = (-)^{q-q'}[1-\tanh^2(\beta/2)]^{1-k}[\tanh^2(\beta/2)]^{(q-q')/2}$$
$$\times\frac{\Gamma(k+q)}{\Gamma(k+q')\Gamma(q-q'+1)}$$
$$\times F(-k+1-q', -k+1+q; q-q'+1; \tanh^2(\beta/2)), \tag{3.11}$$

where F is the hypergeometric function ${}_2F_1(a,b;c;z)$ and Γ is a gamma function. With the help of well-known formulas for the hypergeometric functions[9] it is possible to show straightforwardly that the $D_{q'q}^k(\beta)$ of (3.11) satisfies (3.10).

Our concern now is to show that $D_{q'q}^k(\beta)$ corresponds to an indecomposable, i.e., not completely reducible representation. We shall analyze first the situation when $q' \le q$ where we can write $D_{q'q}^k(\beta)$ as

$$D_{q'q}^k(\beta)$$
$$= (-)^{q-q'}\left(1-\tanh^2\frac{\beta}{2}\right)^{1-k}\left(\tanh^2\frac{\beta}{2}\right)^{(q-q')/2}$$
$$\times\frac{\Gamma(k+q)}{\Gamma(k+q')}\sum_m\frac{(-k+1-q')_m(-k+1+q)_m}{\Gamma(q-q'+m+1)m!}\left(\tanh^2\frac{\beta}{2}\right)^m, \tag{3.12}$$

where $(\alpha)_m$ stands for the Pochhammer symbol

$$(\alpha)_m = \Gamma(\alpha+m)/\Gamma(\alpha), \tag{3.13}$$

where m is a nonnegative integer. The Pochhammer symbol can never become infinite and as

$$\Gamma(k+q)/\Gamma(k+q') = (k+q')_{q-q'} \tag{3.14}$$

we see that for $q \ge q'$, $D_{q'q}^k(\beta)$ is finite for any value of the indices and $0 \le \beta \le \pi$. We note though that if

$$k+q > 0, \quad k+q' \le 0, \tag{3.15}$$

the ratio (3.14) of Γ functions vanishes. We can then divide the interval for q in three parts,

$$(+):\ q > -k, \quad (0):\ -k \ge q \ge k, \quad (-):\ k > q, \tag{3.16}$$

and similarly for q'. We denote these intervals by the symbols (+), (0), (−) as indicated in (3.16). Thus for $q \ge q'$ we note that if q is in the interval (+) and q' is in (0) or (−), the representation $D_{q'q}^k(\beta)$ vanishes.

We now turn our attention to the problem when $q' \ge q$. Making use of the relation[9]

$$\lim_{\gamma\to -m}F(\alpha,\beta;\gamma;z)$$
$$= \frac{(\alpha)_m(\beta)_m}{(m+1)!}z^{m+1}F(\alpha+m+1, \beta+m+1; m+2; z), \tag{3.17}$$

we immediately see that we can write $D_{q'q}^k(\beta)$ of (3.11) as

$$D_{q'q}^k(\beta) = (-)^{q-q'}\left(1-\tanh^2\frac{\beta}{2}\right)^{1-k}\left(\tanh^2\frac{\beta}{2}\right)^{(q'-q)/2}$$

$$\times \frac{\Gamma(k-q)}{\Gamma(k-q')}$$

$$\times F\left(-k+1-q, -k+1+q'; q'-q+1; \tanh^2\frac{\beta}{2}\right). \quad (3.18)$$

Using this formula for the case $q' \geq q$ we immediately see by reasoning similar to the previous one that the $D^k_{q',q}(\beta)$ is always bounded and it vanishes when

$$k-q>0, \quad k-q'\leq 0, \quad (3.19)$$

i.e., when q is in the interval (–) and q' is either in the interval (+) or (0). We can summarize results in the matrix expression

$\|D^k_{qq'}(\beta)\| =$

q' \ q	$+\infty \cdots -k+1$	$-k \cdots k$	$k-1 \cdots -\infty$
$+\infty \cdots -k+1$	$\mathbf{D}^k_{++}$	$\mathbf{D}^k_{+0}$	0
$-k \cdots k$	0	$\mathbf{D}^k_{00}$	0
$k-1 \cdots -\infty$	0	$\mathbf{D}^k_{-0}$	$\mathbf{D}^k_{--}$

(3.20)

We also decompose the set of functions $\mathcal{P}^k_q(r,t)$ into three subsets corresponding to $q>-k$, $-k\geq q\geq k$, $k>q$ which we may denote as the row vector

$$\{\mathcal{P}^k_q(r,t)\} \equiv [\mathbf{P}^k_+, \mathbf{P}^k_0, \mathbf{P}^k_-]. \quad (3.21)$$

We then see that

$$\exp(i\beta T^0_2)[\mathbf{P}^k_+, \mathbf{P}^k_0, \mathbf{P}^k_-]\exp(-i\beta T^0_2)$$
$$= [\mathbf{P}^k_+ D^k_{++}, \mathbf{P}^k_+ D^k_{+0} + \mathbf{P}^k_0 D^k_{00} + \mathbf{P}^k_- D^k_{-0}, P^k_- D^k_{--}]. \quad (3.22)$$

Under the action of the operator $\exp(i\beta T^0_2)$ the functions $\mathcal{P}^k_q(r,t)$ for $q>-k$ or $q<-k$ transform among themselves. In fact D^k_{++}, D^k_{--} correspond to irreducible representations of Sp(2) [or equivalently SU(1,1)] bounded from below or from above. On the other hand the operation $\exp(i\beta T^0_2)$ on the set of functions $\mathcal{P}^k_q(r,t)$ in the interval (0) involves also those in the intervals (+) and (–).

The Armstrong tensor is then an indecomposable one but, as we proceed to show, the Wigner–Eckart theorem is still applicable to these types of tensors and so the radial matrix elements can be evaluated by purely algebraic means. The procedure followed below closely parallels the analysis of Armstrong.[3,4]

Indecomposable representations bounded either from above or below have been considered in a different context by Gruber and Klymik.[12] Their paper provides a good introduction to what for physicists is still a somewhat esoteric subject.

C. The Wigner-Eckart theorem for indecomposable tensors

From the definition (2.32) of scalar product and (3.7) of the nonnormalized ket $|\lambda\mu\rangle$ we see that

$$\langle\lambda'\mu'|\mathcal{P}^k_q(r,t)|\lambda\mu\rangle$$
$$=\delta_{\mu'-\mu,q}\int_0^\infty r^{2\lambda'-1/2}L^{2\lambda'-1}_{\mu'-\lambda'}(r^2)$$
$$\times \exp(-r^2/2)r^{-2k}r^{2\lambda-1/2}L^{2\lambda-1}_{\mu-\lambda}(r^2)\exp(-r^2/2)\,dr. \quad (3.23)$$

To carry out a group theoretical analysis of the radial integral we take into account the properties (3.7) of the kets and the commutation rules (3.5) to write[3,4]

$$\langle\lambda'\mu'|[T^0_\pm,\mathcal{P}^k_q]|\lambda\mu\rangle = (q\mp k\pm 1)\langle\lambda'\mu'|\mathcal{P}^k_{q\pm1}|\lambda\mu\rangle$$
$$=(\mu'\pm\lambda'\mp 1)\langle\lambda'\mu'\mp 1|\mathcal{P}^k_q|\lambda\mu\rangle$$
$$-\langle\lambda'\mu'|\mathcal{P}^k_q|\lambda\mu\pm 1\rangle\,(\mu\mp\lambda\pm 1). \quad (3.24)$$

A similar relation for T^0_3 provides the selection rule

$$\mu'-\mu=q. \quad (3.25)$$

The recursion relations (3.24) are very similar to those used by Racah[16] in the derivation of the Wigner coefficients of SU(2). Using the same type of approach, we start by writing

$$\langle\lambda'\mu'|\mathcal{P}^k_q|\lambda\mu\rangle = \frac{(-)^{\mu-\lambda}}{(\mu-\lambda)!(q-k)!}f(\mu q;\lambda'\mu'), \quad (3.26)$$

whereby the two recursion formulas (3.24) become

$$(\mu'+\lambda'-1)f(\mu q;\lambda'\mu'-1) = f(\mu q+1;\lambda'\mu') - f(\mu+1\; q;\lambda'\mu'), \quad (3.27a)$$

$$(\mu'-\lambda'+1)f(\mu q;\lambda'\mu'+1)$$
$$=(q-k)(q+k-1)f(\mu\; q-1;\lambda'\mu')$$
$$-(\mu-\lambda)(\mu+\lambda-1)f(\mu-1\; q;\lambda'\mu'). \quad (3.27b)$$

Setting $\mu'=\lambda'$ in (3.27a), and recalling that only matrix elements with $\mu'\geq\lambda'$ can arise, we deduce that

$$f(\mu q;\lambda'\lambda') = A(\lambda' k\lambda) \quad (3.28)$$

with A being independent of the "magnetic" quantum numbers μ', q, μ. Then from (3.27b) and (3.28) we can obtain explicit formulas for $f(\mu q;\lambda'\lambda'+1)$, $f(\mu q;\lambda'\lambda'+2),\ldots,$ which suggest the following general solution for $f(\mu q;\lambda'\mu')$:

$$f(\mu q;\lambda'\mu') = A(\lambda' k\lambda)\sum_t \frac{(-)^t(\mu-\lambda)!(\mu+\lambda-1)!(q-k)!(q+k-1)!}{t!(\mu'-\lambda'-t)!(\mu-\lambda-t)!(\mu+\lambda-1-t)!(q-k-\mu'+\lambda'+t)!(q+k-1-\mu'+\lambda'+t)!}$$
$$=A(\lambda' k\lambda)\delta_{q,\mu'-\mu}\frac{(q+k-1)!(q-k)!}{c!(d-1)!(e-1)!}\,{}_3F_2\left[\begin{matrix}a, & b, & -c\\ d, & e & \end{matrix};1\right], \quad (3.29)$$

where in the last step we have used the selection rule (3.25), and the parameters of the hypergeometric function have the values

$$a=-\mu-\lambda+1,\quad b=-\mu+\lambda,\quad c=\mu'-\lambda',$$
$$d=-k-\mu+\lambda'+1,\quad e=k-\mu+\lambda'. \tag{3.30}$$

If we substitute the "ansatz" (3.29) in (3.27b) we obtain

$${}_3F_2\begin{bmatrix}a, & b, & -c \\ d, & e & \end{bmatrix}\!;1 = {}_3F_2\begin{bmatrix}a, & b, & -c-1 \\ d, & e & \end{bmatrix}\!;1 + \frac{ab}{de}\,{}_3F_2\begin{bmatrix}a+1, & b+1, & -c \\ d+1, & e+1 & \end{bmatrix}\!;1. \tag{3.31a}$$

Hence, if this formula is true, Eq. (3.29) will be the correct solution of the recursion formula (3.27b). But (3.31) is indeed true, since from the identities

$$p(p+1)_m=(p)_{m+1},$$
$$(p+m)(p)_m=(p)_{m+1}, \tag{3.31b}$$
$$(p-1)(p)_m=(p-1)_{m+1},$$

we have

$$\frac{ab}{de}\,{}_3F_2\begin{bmatrix}a+1, & b+1, & -c \\ d+1, & e+1 & \end{bmatrix}\!;1$$
$$=\sum_{m=0}\frac{(a)_{m+1}(b)_{m+1}(-c)_m(-c+m+1+c)}{(m+1)!(d)_{m+1}(e)_{m+1}}$$
$$=\sum_{m=0}\frac{(a)_{m+1}(b)_{m+1}(-c)_{m+1}}{(m+1)!(d)_{m+1}(e)_{m+1}}$$
$$-\sum_{m=0}\frac{(a)_{m+1}(b)_{m+1}(-c-1)_{m+1}}{(m+1)!(d)_{m+1}(e)_{m+1}} \tag{3.31c}$$
$$={}_3F_2\begin{bmatrix}a, & b, & -c \\ d, & e & \end{bmatrix}\!;1 - {}_3F_2\begin{bmatrix}a, & b, & -c-1 \\ d, & e & \end{bmatrix}\!;1.$$

By similar analysis it can be checked that $f(\mu q;\lambda'\mu')$ as given in Eq. (3.29) satisfies the recursion formula (3.27a) as well.

The matrix elements of the tensor $\mathcal{P}^k_q$ are thus given, according to (3.26) and (3.29), by

$$\langle\lambda'\mu'|\mathcal{P}^k_q|\lambda\mu\rangle = A(\lambda'k\lambda)\delta_{q,\mu'-\mu}\frac{(-)^b(q+k-1)!}{b!\,c!\,(d-1)!\,(e-1)!}\times{}_3F_2\begin{bmatrix}a, & b, & -c \\ d, & e & \end{bmatrix}\!;1 \tag{3.32}$$

with a,b,c,d,e defined in (3.30), and where we have still to determine $A(\lambda'k\lambda)$. For this purpose we make a direct evaluation of the integral in (3.23) for the case $\mu=\lambda$, $\mu'=\lambda'$, and identify the result with the right-hand side of (3.32) for the same values of μ,μ'. We find thus that

$$A(\lambda'k\lambda)=\pi(\lambda'-\lambda-k)!\,\Gamma(\lambda+\lambda'-k). \tag{3.33}$$

We have given a group theoretical derivation of the radial matrix elements with respect to harmonic oscillator states. Using the transformation formula[17]

$${}_3F_2\begin{bmatrix}a, & b, & c \\ d, & e & \end{bmatrix}\!;1 = \frac{(-)^c\Gamma(b-e+1)\Gamma(e)}{\Gamma(b+c-e+1)\Gamma(e-c)}\times{}_3F_2\begin{bmatrix}d-a, & b, & c \\ d, & b+c-e+1 & \end{bmatrix}\!;1. \tag{3.34}$$

The matrix elements can also be written as

$$\langle\lambda'\mu'|\mathcal{P}^k_q|\lambda\mu\rangle = \delta_{q,\mu'-\mu}\pi\Gamma(\lambda+\lambda'-k)\Gamma(-k-\lambda+\lambda'+1) \tag{3.35}$$
$$\times\frac{(-)^{\mu'-\lambda'+\mu-\lambda}\Gamma(-k-\lambda'+\lambda+1)}{\Gamma(\mu-\lambda+1)\Gamma(\mu'-\lambda'+1)\Gamma(-k-\mu+\lambda'+1)\Gamma(-k-\mu'+\lambda+1)}$$
$$\times{}_3F_2\begin{bmatrix}-k+\lambda+\lambda', & \lambda-\mu, & \lambda'-\mu' \\ -k-\mu+\lambda'+1, & -k-\mu'+\lambda+1 & \end{bmatrix}\!;1$$

which have a symmetric form with respect to $\lambda,\mu;\lambda',\mu'$ and besides can be derived directly from the radial integral with the help of generating functions of Laguerre polynomials.[9]

The ${}_3F_2$ function appearing in (3.35) is then related to the Wigner coefficients associated with the irreducible representations λ,λ' and the indecomposable one characterized by the index k associated with the $\mathcal{P}^k_q(r,t)$. In particular cases these Wigner coefficients involve only irreducible representations. For example, if $q>-k$ $(q<k)$, then the property $T^0_-\mathcal{P}^k_{-k+1}=0$ $(T^0_+\mathcal{P}^k_{k-1}=0)$ implies that the tensor $\mathcal{P}^k_q(r,t)$ is irreducible and bounded from below (above). The result of the previous paragraph indicates that the Wigner coefficients contain an indecomposable representation only when $k\le q\le -k$. But even here we have that under certain circumstances the representation is irreducible and finite dimensional. This happens, for example, when both $-k-\lambda+\lambda'$ and $-k-\lambda'+\lambda$ are nonnegative integers. In that case the matrix (3.35) can be written as

$$\langle\lambda'\mu'|\mathcal{P}^k_q|\lambda\mu\rangle = \delta_{q,\mu'-\mu}\pi$$
$$\times\frac{(-)^{\mu'-\lambda'+\mu-\lambda}\Gamma(\lambda+\lambda'-k)\Gamma(-k-\lambda+\lambda'+1)\Gamma(-k-\lambda'+\lambda+1)}{\Gamma(\mu-\lambda+1)\Gamma(\mu'-\lambda'+1)}$$
$$\times\sum_\nu\frac{(-k+\lambda+\lambda')_\nu(\lambda-\mu)_\nu(\lambda'-\mu')_\nu}{\Gamma(-k-\mu+\lambda'+\nu+1)\Gamma(-k-\mu'+\lambda+\nu+1)}, \tag{3.36}$$

where from the properties of Γ functions and Pochhammer symbols, ν is limited by the inequalities

$$0\le\mu-\lambda-\nu\le-k+\lambda'-\lambda, \tag{3.37a}$$
$$0\le\mu'-\lambda'-\nu\le-k+\lambda-\lambda'. \tag{3.37b}$$

Changing the sign and order of the inequalities in (3.37a) and summing it with (3.37b) we finally get

$$-k+\lambda-\lambda'\ge(\mu'-\mu)-(\lambda'-\lambda)\ge k+\lambda-\lambda', \tag{3.38}$$

which from the relation $q=\mu'-\mu$ implies that the values of q are limited to the interval

$$-k\ge q\ge k. \tag{3.39}$$

Thus if $-k+\lambda-\lambda'$, $-k-\lambda+\lambda'$ are nonnegative integers, the matrix elements vanish unless q is restricted as in (3.39). The index k characterizes in this case an irreducible and finite dimensional representation of SU(1,1); the Wigner coefficients for this type of representation have been given by Ui.[18]

The importance of the particular case discussed in the previous paragraphs is that for a problem in three-dimensional space the condition

$$|\lambda-\lambda'|\le-k \tag{3.40}$$

corresponds to

$$|\kappa - \kappa'| \leq -2k, \qquad (3.41)$$

where κ, κ' are IR of the $O(3)$ group indicated in (2.7). This condition is automatically satisfied in the approach followed by Quesne and Moshinsky[6] because of the selection rule for the angular part of the matrix elements. Furthermore, the same rule indicates that $-\kappa + \kappa' - 2k$ is an *even* integer which implies that $\lambda - \lambda' - k$ is an integer. Thus the analysis we have carried out in this section, involving the Wigner coefficients in which one of the representations is indecomposable, reduces, when (3.41) is satisfied, to the analysis of Quesne and Moshinsky[6] which involves a corresponding representation that is both irreducible and finite dimensional.

4. TWO-BODY MATRIX ELEMENTS

Once the $R^\lambda_\mu(r)$ are understood as BIR of Sp(2), we can certainly construct from products of two states of this type associated with coordinates r_1, r_2 a new state that corresponds again to a BIR of Sp(2), i.e.,

$$\Phi^{\lambda_1\lambda_2\Lambda}_M(r_1, r_2) = \sum_{\mu_1\mu_2} \langle\lambda_1\lambda_2\mu_1\mu_2|\Lambda M\rangle_{nc} R^{\lambda_1}_{\mu_1}(r_1) R^{\lambda_2}_{\mu_2}(r_2), \qquad (4.1)$$

where $\langle\ |\ \rangle_{nc}$ is a Wigner coefficient of the noncompact group Sp(2), but now all the IR appearing in it are infinite dimensional. We note that $M = \mu_1 + \mu_2$ and thus the minimum value of M is $\lambda_1 + \lambda_2$. Furthermore, we continue to characterize the indices in such a way that $M = \Lambda$ gives the state of lowest weight. Thus Λ is restricted by

$$\lambda_1 + \lambda_2 \leq \Lambda \leq M. \qquad (4.2)$$

The generators of the Sp(2) algebra in the Heisenberg picture for the two particle problem will be the sum of the T_1 of (2.3) associated with particles 1 and 2. Using variables in $2m$-dimensional space defined as

$$\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2), \quad \mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2) \qquad (4.3)$$

they become

$$T_1 = \tfrac{1}{4}(\mathbf{P}^2 - \mathbf{R}^2), \quad T_2 = \tfrac{1}{2}(\mathbf{R}\cdot\mathbf{P} + \mathbf{P}\cdot\mathbf{R}),$$
$$T_3 = \tfrac{1}{4}(P^2 + R^2) = \tfrac{1}{2}H. \qquad (4.4)$$

The volume element in $2m$-dimensional configuration space is

$$R^{2m-1}\, dR \sin^{m-1}\alpha \cos^{m-1}\alpha\, d\alpha\, d\Omega_1\, d\Omega_2, \qquad (4.5)$$

where

$$r_1 = R\cos\alpha, \quad r_2 = R\sin\alpha, \qquad (4.6)$$

and $d\Omega_1$, $d\Omega_2$ are the elements of the solid angle associated with particles 1 and 2.

The operators (4.4) are given in the Heisenberg picture. We note that their commutation relations and equations of motion continue to be given by (2.4) and (2.23) respectively. Thus using (2.24) we can express the Schrödinger operators in terms of the Heisenberg ones. Again, as in Sec. 2, these operators are applied to the time dependent Schrödinger states and we can make the replacements

$$\tfrac{1}{2}(P^2 + R^2) \to i\frac{\partial}{\partial t}. \qquad (4.7)$$

Thus we arrive finally at the Schrödinger operators

$$T^0_3 = \frac{i}{2}\frac{\partial}{\partial t}, \qquad (4.8a)$$

$$T^0_\pm = \tfrac{1}{2}\exp(\mp 2it)\left(i\frac{\partial}{\partial t} - R^2 \pm R\frac{\partial}{\partial R} \pm m\right), \qquad (4.8b)$$

where from (4.6) we have

$$R\frac{\partial}{\partial R} = r_1\frac{\partial}{\partial r_1} + r_2\frac{\partial}{\partial r_2}, \quad \Big| \quad R^2 = r_1^2 + r_2^2. \qquad (4.9)$$

The operators T^0_3, $T^0_\pm$ can now be applied to the two-particle radial Schrödinger wavefunction

$$\Psi^{\lambda_1\lambda_2\Lambda}_M(r_1, r_2, t) = \Phi^{\lambda_1\lambda_2\Lambda}_M(r_1, r_2)\exp(-i2Mt), \qquad (4.10)$$

and give, from the very definition of Wigner coefficient,[10] expressions similar to (2.29), i.e.,

$$T^0_3\Psi^{\lambda_1\lambda_2\Lambda}_M = M\Psi^{\lambda_1\lambda_2\Lambda}_M, \qquad (4.11a)$$

$$T^0_\pm\Psi^{\lambda_1\lambda_2\Lambda}_m = [(M\pm\Lambda)(M\mp\Lambda\pm 1)]^{1/2}\Psi^{\lambda_1\lambda_2\Lambda}_{M\pm 1}. \qquad (4.11b)$$

The Casimir operator for the two-particle case in the Schrödinger picture is obtained from (2.30) if we replace r by R and m by $2m$. Thus the dependence of the volume elements on R, t must be of the form

$$R^{2m-3}\, dR\, dt \qquad (4.12)$$

if we want T^{0^2}, T^0_3 to be Hermitian. We still have to determine the dependence of the volume element on the α defined in (4.6). For this we note that another relevant operator is the Casimir one for the $O(2m)$ group, as two particles in an m-dimensional oscillator are equivalent to one particle in a $2m$-dimensional oscillator. From (2.6) and (4.6) we see that

$$\begin{aligned}
L^2 &= -(\mathbf{r}_1\cdot\mathbf{P}_1 + \mathbf{r}_2\cdot\mathbf{P}_2)^2 + i(m-2)(\mathbf{r}_1\cdot\mathbf{P}_1 + \mathbf{r}_2\cdot\mathbf{P}_2) \\
&\quad + (r_1^2 + r_2^2)(p_1^2 + p_2^2) \\
&= R\frac{\partial}{\partial R}R\frac{\partial}{\partial R} + (m-2)R\frac{\partial}{\partial R} \\
&\quad + R^2\left(-\frac{1}{r_1^{m-1}}\frac{\partial}{\partial r_1}r_1^{m-1}\frac{\partial}{\partial r_1} + \frac{L_1^2}{r_1^2} - \frac{1}{r_2^{m-1}}\frac{\partial}{\partial r_2}r_2^{m-1}\frac{\partial}{\partial r_2} + \frac{L_2^2}{r_2^2}\right) \\
&= -\frac{1}{\sin^{m-1}\alpha\cos^{m-1}\alpha}\frac{\partial}{\partial\alpha}\sin^{m-1}\alpha\cos^{m-1}\alpha\frac{\partial}{\partial\alpha} \\
&\quad + \frac{L_1^2}{\cos^2\alpha} + \frac{L_2^2}{\sin^2\alpha},
\end{aligned} \qquad (4.13)$$

where L_1^2, L_2^2 are respectively the Casimir operators associated with the $O(m)$ groups of particles 1, 2. From (4.13) the operator L^2 will be Hermitian only if the volume element contains the factor $\sin^{m-1}\alpha\cos^{m-1}\alpha\, d\alpha$, and thus finally for the two-particle problem we arrive at the volume element

$$R^{2m-3}\sin^{m-1}\alpha\cos^{m-1}\alpha\, dR\, d\alpha\, dt = R^{-2}r_1^{m-1}r_2^{m-1}\, dr_1\, dr_2\, dt. \qquad (4.14)$$

We have now Schrödinger two-particle radial states $\Psi^{\lambda_1\lambda_2\Lambda}_M(r_1, r_2, t)$ which are BIR of Sp(2), as well as the volume element (4.14) with respect to which we can define scalar products. If we wish to calculate two-body matrix elements using the Wigner–Eckart theorem, we would have to introduce irreducible tensors in a way similar to the one in which indecomposable ones were discussed in the previous section. One of the more interesting examples concerns radial matrix elements associated with the Coulomb interaction $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$. The expansion of this in spherical harmonics gives radial parts of the form $r_<^s/r_<^{s+1}$ where $r_<$ is the smaller and $r_>$ is the larger of r_1, r_2. This expression suggests that, following the lead of Crubellier,[2] we introduce functions of r_1, r_2, t of the type

$$ {}^sP_q^{a/2}(r_1, r_2, t) = \left[\frac{(q-a/2)!(a-1)!}{(a/2+q-1)!}\right]^{1/2} R^2 \frac{r_<^s}{r_>^{s+a}} e^{-i2qt}, \tag{4.15}$$

where a, s, are nonnegative integers and

$$q = a/2, a/2+1, a/2+2, \cdots. \tag{4.16}$$

Using (4.8) we can immediately check that

$$[T_3^0, {}^sP_q^{a/2}] = q\ {}^sP_q^{a/2} \tag{4.17a}$$

$$[T_\pm^0, {}^sP_q^{a/2}] = [(q \pm a/2)(q \pm a/2 \pm 1)]^{1/2}\ {}^sP_{q\pm1}^{a/2}. \tag{4.17b}$$

Thus independently of the value of S we have an irreducible tensor of Sp(2) characterized by the IR $a/2$. We note from (4.16) that the representation is infinite dimensional.

The matrix elements of interest to us come from the following integral

$$\begin{aligned}
&\int_0^\infty \int_0^\infty \int_0^{2\pi} [\Psi_{M'}^{\lambda_1'\lambda_2'\Lambda'}(r_1, r_2, t)]^*\ {}^sP_q^{a/2}(r_1, r_2, t) \\
&\times \Psi_M^{\lambda_1\lambda_2\Lambda}(r_1, r_2, t) R^{-2} r_1^{m-1} r_2^{m-1}\, dr_1\, dr_2\, dt \\
&= (2\pi)\delta_{M+q,M'} \left[\frac{(q-a/2)!(a-1)!}{(a/2+q-1)!}\right]^{1/2} \int_0^\infty \int_0^\infty [\Phi_{M'}^{\lambda_1'\lambda_2'\Lambda'}(r_1 r_2)]^* \\
&\times \frac{r_<^s}{r_>^{s+a}} \Phi_M^{\lambda_1\lambda_2\Lambda}(r_1 r_2) r_1^{m-1} r_2^{m-1}\, dr_1\, dr_2 \\
&= \langle \Lambda(a/2) M q \mid \Lambda' M'\rangle_{nc} \langle \lambda_1'\lambda_2'\Lambda' \| {}^sP^{a/2} \| \lambda_1\lambda_2\Lambda\rangle.
\end{aligned} \tag{4.18}$$

Again $\langle\,|\,\rangle_{nc}$ are the Wigner coefficients for the noncompact group Sp(2), or equivalently SU(1,1),[6,10] discussed in Appendix A. The last term in the equation is the reduced matrix element which can be calculated from the radial integral $M' = \Lambda'$, $M = \Lambda$. The detailed evaluation of the latter is given in Appendix B.

The procedure followed for the evaluation of the two body radial matrix elements in this section exactly parallels the application of the Wigner—Eckart theorem in the two-body angular matrix element (1.1b).

We note that from the orthonormality properties of the Wigner coefficients we can write

$$\begin{aligned}
&\int_0^\infty \int_0^\infty R_{\mu_1'}^{\lambda_1'*}(r_1) R_{\mu_2'}^{\lambda_2'*}(r_2) \frac{r_<^s}{r_>^{s+a}} R_{\mu_1}^{\lambda_1}(r_1) \\
&\times R_{\mu_2}^{\lambda_2}(r_2) r_1^{m-1} r_2^{m-1}\, dr_1\, dr_2 \\
&= \sum_{\Lambda' M'} \sum_{\Lambda M} \langle \lambda_1'\lambda_2'\mu_1'\mu_2' \mid \Lambda' M'\rangle_{nc} \langle \lambda_1\lambda_2\mu_1\mu_2 \mid \Lambda M\rangle_{nc} \\
&\times \int_0^\infty \int_0^\infty [\Phi_{M'}^{\lambda_1'\lambda_2'\Lambda'}(r_1 r_2)]^* \frac{r_<^s}{r_>^{s+a}} \Phi_M^{\lambda_1\lambda_2\Lambda}(r_1 r_2) r_1^{m-1} r_2^{m-1}\, dr_1\, dr_2.
\end{aligned} \tag{4.19}$$

Thus the two-body matrix elements of $(r_<^s/r_>^{s+a})$ with respect to m-dimensional radial harmonic oscillator states can be obtained by purely group theoretical methods in terms of the reduced matrix element in (4.18).

In the next two sections we particularize the present analysis to the ordinary harmonic oscillator, i.e., $m=3$, and to the Coulomb problem which we show, at least in what we have called the pseudo-Coulomb form,[8] to be entirely equivalent to the radial oscillator with $m=4$.

5. MATRIX ELEMENTS FOR THREE-DIMENSIONAL OSCILLATOR STATES

In this section we note only that for $m=3$, κ is the angular momentum l of the particle. Thus

$$\nu = l + \tfrac{1}{2}, \quad \lambda = \tfrac{1}{2}(l + \tfrac{3}{2}), \quad \mu = n + \tfrac{1}{2}(l + \tfrac{3}{2}). \tag{5.1}$$

For the Coulomb interaction with respect to oscillator states the radial part takes the form $r_<^s/r_>^{s+1}$. Thus in formula (4.18), when $m=3$, we are interested in the case $a=1$.

The Wigner coefficients that appear in the previous sections have to be particularized to the above values of the parameters.

6. RADIAL MATRIX ELEMENTS FOR THE THREE-DIMENSIONAL COULOMB PROBLEM

Let us denote by $\mathbf{r}'$, $\mathbf{p}'$ the coordinates and momenta in three-dimensional space and in atomic units ($e = \hbar = m = 1$). The Schrödinger equation for the Hamiltonian of the Coulomb problem becomes then

$$(\tfrac{1}{2}p'^2 - 1/r')\psi = -(1/2N^2)\psi, \quad N = 1, 2, \cdots. \tag{6.1a}$$

We denote by N the total quantum number to distinguish it from the radial quantum number n that appears in Sec. 2 for the oscillator. For the Coulomb problem the total and radial quantum numbers are related by

$$N = n + l + 1, \tag{6.1b}$$

where l is the angular momentum.

Introducing a dilatation canonical transformation for each energy level

$$\rho = (\mathbf{r}'/N), \quad \pi = N\mathbf{p}', \tag{6.2}$$

we obtain the equation

$$H\psi \equiv \tfrac{1}{2}\rho(\pi^2 + 1)\psi = N\psi. \tag{6.3}$$

The H in (6.3) is the well-known Hamiltonian for what, in Ref. 8, we called the "pseudo-Coulomb" problem. We denote the radial eigenstates of H by $\mathcal{R}_N^l(\rho)$ and they satisfy the equation

$$\frac{1}{2}\rho\left[-\frac{1}{\rho^2}\frac{\partial}{\partial\rho}\rho^2\frac{\partial}{\partial\rho} + \frac{l(l+1)}{\rho^2} + 1\right]\mathcal{R}_N^l(\rho) = N\mathcal{R}_N^l(\rho). \tag{6.4}$$

Introducing the change of variable

$$\rho = \tfrac{1}{2}r^2, \tag{6.5}$$

we obtain then the following equation,

$$\begin{aligned}
&\frac{1}{2}\left[-\frac{1}{r^3}\frac{\partial}{\partial r} r^3 \frac{\partial}{\partial r} + \frac{2l(2l+2)}{r^2} + r^2\right]\mathcal{R}_N^l(\tfrac{1}{2}r^2) \\
&= 2N\mathcal{R}_N^l(\tfrac{1}{2}r^2).
\end{aligned} \tag{6.6}$$

Turning now to Eq. (2.21a), i.e.,

$$2T_3 R_\mu^\lambda(r) = 2\mu R_\mu^\lambda(r), \tag{6.7}$$

where T_3 is given by (2.10a) in which $\mathcal{L}^2$ is replaced by its eigenvalue (2.7), we immediately see that

$$\mathcal{R}_N^l(\tfrac{1}{2}r^2)=R_\mu^\lambda(r) \tag{6.8}$$

when $m=4$ and $\kappa=2l$ which implies

$$\lambda=l+1. \tag{6.9}$$

Besides comparing the eigenvalues in (6.6) and (6.7) we have

$$\mu=N=n+l+1=n+\lambda\ . \tag{6.10}$$

The radial pseudo-Coulomb wavefunctions in three-dimensional space are thus related with the radial harmonic oscillator states in four-dimensional space. This indicates that all matrix elements involving the pseudo-Coulomb states $\mathcal{R}_N^l(\rho)$ can be converted into oscillator matrix elements. We note further that for $\mathcal{H}$ to be Hermitian the volume element must be $\rho\, d\rho$. Therefore, in the one-body case, we are interested in matrix elements of the form

$$\int_0^\infty \mathcal{R}_{N'}^{l'}(\rho)\rho^{-k}\mathcal{R}_N^l(\rho)\rho\, d\rho$$
$$=2^{k-1}\int_0^\infty R_{N'}^{l'+1}(r)r^{-2k}R_N^{l+1}(r)r^3\, dr, \tag{6.11}$$

where, to keep up with the notation used in Sec. 3, k takes the nonpositive integer values $k=0,-1,-2,-3,\cdots$. Looking into (3.2), (3.23) we immediately can determine this integral in terms of Wigner coefficients of SU(1,1) associated with the indecomposable representation k and where we take $m=4$, $\lambda=l+1$, $\mu=N$, $\lambda'=l'+1$, and $\mu'=N'$.

Going now to the two-body case, we are interested in particular in the radial matrix elements associated with a Coulomb interaction, i.e., in the integral

$$\int_0^\infty\int_0^\infty \mathcal{R}_{N_1'}^{l_1'}(\rho)\mathcal{R}_{N_2'}^{l_2'}(\rho_2)\,\frac{\rho_<^t}{\rho_>^{t+1}}\mathcal{R}_{N_1}^{l_1}(\rho_1)$$
$$\times\mathcal{R}_{N_2}^{l_2}(\rho_2)\rho_1\, d\rho_1\rho_2\, d\rho_2$$
$$=\frac{1}{2}\int_0^\infty\int_0^\infty R_{N_1'}^{l_1'+1}(r_1)R_{N_2'}^{l_2'+1}(r_2)\,\frac{r_<^{2t}}{r_>^{2t+2}}$$
$$\times R_{N_1}^{l_1+1}(r_1)R_{N_2}^{l_2+1}(r_2)r_1^3r_2^3\, dr_1\, dr_2. \tag{6.12}$$

Looking into (4.19), we immediately see that we can express (6.12) in terms of the Wigner coefficients and reduced matrix elements of (4.18), if we make the replacements

$$s=2t,\quad a=2,\quad \lambda_1=l_1+1,\quad \mu_1=N_1,\quad \lambda_2=l_2+1,\quad \mu_2=N_2, \tag{6.13}$$

and similarly for the primed ones. The required Wigner coefficients are of the type

$$\langle l_1+1,\ l_2+1,\ N_1,\ N_2|\Lambda M\rangle_{nc} \tag{6.14a}$$
$$\langle \Lambda 1 M q|\Lambda' M'\rangle_{nc}, \tag{6.14b}$$

where all the symbols are nonnegative integers. Thus the coefficients contain only true representations[10,11] and as was shown in a recent paper by the present authors[19] they can be expressed in terms of ordinary Wigner coefficients of SU(2) by the relation[19]

$$\langle l_1+1,\ l_2+1,\ N_1,\ N_2|\Lambda M\rangle_{nc}$$
$$=(-1)^{N_2-l_2+1}\left\langle\frac{M+l_1-l_2-1}{2},\ \frac{M+l_2-l_1-1}{2}\ ,\right.$$
$$\left.\times\frac{N_1-N_2+l_1+l_2+1}{2}\ ,\frac{N_2-N_1+l_1+l_2+1}{2}\right|\Lambda-1,l_1+l_2+1\Big\rangle, \tag{6.15}$$

and similarly for (6.14b).

We have determined the one- and two-body radial matrix elements for the pseudo-Coulomb problem for ρ^{-k}, $\rho_<^t/\rho_>^{t+1}$ respectively. We note though from (6.2) that they will give us the corresponding Coulomb matrix elements only in the case when all the radial eigenfunctions belong to the same energy, i.e., the same N.

7. CONCLUSIONS

We have derived from a unified point of view one- and two-body radial matrix elements using the Wigner—Eckart theorem for the Sp(2) [or equivalently the SU(1,1)] group. It proved fundamental for our objective to obtain realizations of the Lie algebra of Sp(2) in the Schrödinger rather than the Heisenberg picture.

The final results for the radial matrix elements are certainly not simpler than those that can be derived by other procedures, but the explicit group theoretical structure of the problem leads to Wigner coefficients of SU(1,1),[10,11] and thus to all the selection rules contained implicitly in them. In fact for the two-body pseudo-Coulomb radial matrix element of $\rho_<^t/\rho_>^{t+1}$, the Wigner coefficients[19] can be reduced to those of SU(2) by expressions such as (6.15). Then all the extensive results for selection and symmetry rules for the latter coefficients can be applied to the corresponding matrix elements.

The technique used suggests the possibility of deriving through the Schrödinger picture, realizations of Lie algebras appropriate to the discussion of properties of other special functions[7] and of the matrix elements that can be defined with respect to them.

ACKNOWLEDGMENTS

The authors are indebted to L. Armstrong, A. Crubellier, B. Gruber, K.B. Wolf, and C. Wulfman for very helpful discussions on this subject. One of the authors (D.L.) would like to thank the Consejo Nacional de Ciencia y Tecnología (CONACYT) for the support it gave him during his stay in México and also the Instituto de Física, UNAM, for its hospitality.

APPENDIX A

In this Appendix we summarize some results on the Wigner coefficients of SU(1,1), discussed by Biedenharn and Holman,[10,11] that are relevant to the analysis of the two-body radial matrix elements discussed in the present paper.

The starting point in Ref. 10 is the set of states $|\lambda\mu\rangle$, $\lambda \geq 0$, $\mu = \lambda, \lambda+1, \ldots$, which constitutes an infinite dimensional and unitary BIR of SU(1,1). Thus the generators T_3, $T_\pm$ of this group acting on the ket give[10,11]

$$T_3|\lambda\mu\rangle = \mu|\lambda\mu\rangle, \quad T_\pm|\lambda\mu\rangle = [(\mu\pm\lambda)(\mu\mp\lambda\pm 1)]^{1/2}|\lambda\mu\pm 1\rangle. \tag{A1}$$

The Wigner coefficients provide us, then, with the set of states $|\Lambda M\rangle$ that are again BIR of SU(1,1), formed from the direct product of $|\lambda_1\mu_1\rangle$, $|\lambda_2\mu_2\rangle$, i.e.,

$$|\Lambda M\rangle = \sum_{\mu_1\mu_2} \langle\lambda_1\lambda_2\mu_1\mu_2|\Lambda M\rangle_{nc}|\lambda_1\mu_1\rangle|\lambda_2\mu_2\rangle. \tag{A2}$$

The generators of SU(1,1) for the direct product are the sums of the generators for systems 1 and 2, i.e.,

$$T_\pm = T_\pm^{(1)} + T_\pm^{(2)}, \quad T_3 = T_3^{(1)} + T_3^{(2)}. \tag{A3}$$

Applying them to (A2) we get recurrence relations for the Wigner coefficients entirely similar to those obtained by Racah[17] for SU(2). These recursion relations were solved in Refs. 10 and 11, and with the phase convention given there, one obtains

$$\begin{aligned}\langle\lambda_1\lambda_2\mu_1\mu_2|\Lambda M\rangle_{nc} = {}& (-1)^{\lambda_1+\lambda_2-\Lambda}(2\Lambda-1)^{1/2}[\Gamma(\mu_2+\lambda_1-\Lambda+1)\Gamma(\mu_2+\Lambda+\lambda_1)]^{-1}\\ &\times\left[\frac{\Gamma(\Lambda+\lambda_1+\lambda_2-1)\Gamma(\lambda_1+\Lambda-\lambda_2)\Gamma(M-\Lambda+1)\Gamma(M+\Lambda)\Gamma(\mu_2-\lambda_2+1)\Gamma(\mu_2+\lambda_2)}{\Gamma(\lambda_2+\Lambda-\lambda_1)\Gamma(1+\Lambda-\lambda_1-\lambda_2)\Gamma(\mu_1-\lambda_1+1)\Gamma(\mu_1+\lambda_1)}\right]^{1/2}\\ &\times {}_3F_2\left[\begin{matrix}\mu_2+\lambda_2,\ \lambda_1-\mu_1,\ \mu_2-\lambda_2+1\\ \mu_2-\Lambda+\lambda_1+1,\ \mu_2+\Lambda+\lambda_1\end{matrix};1\right],\end{aligned} \tag{A4}$$

with

$$\mu_1 = \lambda_1, \lambda_1+1, \cdots, \quad \mu_2 = \lambda_2, \lambda_2+1, \cdots, \quad M = \Lambda, \Lambda+1, \cdots, \tag{A5a}$$

$$\lambda_1 \geq 0, \quad \lambda_2 \geq 0, \quad \Lambda = \lambda_1+\lambda_2, \lambda_1+\lambda_2+1, \cdots, \tag{A5b}$$

i.e., three unitary infinite discrete representations of SU(1,1). We note only that, for comparison of (A4) with formula (2.11) of Ref. 11, we require the identification of $(-1)^{\lambda_1+\lambda_2-\Lambda}$ with the phase factor φ_+ given in (2.10) of the same reference. This can be achieved trivially when we notice from (A5) that $\Lambda-\lambda_1-\lambda_2$ is a nonnegative integer.

The Wigner coefficients (A4) encompass both the one given by the same notation in the present paper as well as $\langle\Lambda(a/2)Mq|\Lambda'M'\rangle_{nc}$.

APPENDIX B

In this appendix we calculate the reduced matrix elements of the two-body m-dimensional harmonic oscillator radial functions, which, from (4.18) are

$$\begin{aligned}\langle\lambda_1'\lambda_2'\Lambda'\|{}^sP^{a/2}\|\lambda_1\lambda_2\Lambda\rangle = {}& [\langle\Lambda(a/2)\Lambda q|\Lambda'\Lambda'\rangle_{nc}]^{-1}\delta_{\Lambda'-\Lambda,q}\left[\frac{(q-a/2)!(a-1)!}{(a/2+q-1)!}\right]^{1/2}\\ &\times 2\pi\sum_{\mu_1\mu_2}\sum_{\mu_1'\mu_2'}\{\langle\lambda_1'\lambda_2'\mu_1'\mu_2'|\Lambda'\Lambda'\rangle_{nc}\langle\lambda_1\lambda_2\mu_1\mu_2|\Lambda\Lambda\rangle_{nc}I\},\end{aligned} \tag{B1}$$

where

$$I = \int_0^\infty\int_0^\infty R_{\mu_1'}^{\lambda_1'}(r_1)R_{\mu_2'}^{\lambda_2'}(r_2)\frac{r_<^s}{r_>^{s+a}}R_{\mu_1}^{\lambda_1}(r_1)R_{\mu_2}^{\lambda_2}(r_2)r_1^{m-1}r_2^{m-1}\,dr_1\,dr_2 = \int_0^\infty\int_0^\infty f_{\mu_1'}^{\nu_1'}(r_1)f_{\mu_2'}^{\nu_2'}(r_2)\frac{r_<^s}{r_>^{s+a}}f_{\mu_1}^{\nu_1}(r_1)f_{\mu_2}^{\nu_2}(r_2)\,dr_1\,dr_2. \tag{B2}$$

The SU(1,1) Wigner coefficients which appear in formula (B1) do not contain a summation and, choosing the normalization and phase as Biedenharn and Holman[10] have, are given by the following formula:

$$\begin{aligned}\langle\lambda_1\lambda_2\mu_1\mu_2|\Lambda\Lambda\rangle_{nc} = {}& (-1)^{\mu_1+\lambda_2-\Lambda}[2\Lambda-1]^{1/2}\\ &\times\left[\frac{\Gamma(\Lambda-\lambda_1+\lambda_2)\Gamma(\Lambda-\lambda_1-\lambda_2+1)\Gamma(\Lambda+\lambda_1-\lambda_2)\Gamma(\lambda_1+\lambda_2+\Lambda-1)}{\Gamma(\mu_2+\lambda_2)\Gamma(\mu_1+\lambda_1)\Gamma(\mu_2-\lambda_2+1)\Gamma(\mu_1-\lambda_1+1)\Gamma(2\Lambda)}\right]^{1/2}.\end{aligned} \tag{B3}$$

Following Moshinsky[21] we can write

$$f_{n'}^{\nu'}(r)f_n^\nu(r) = 2r^2\sum_p\frac{B(n',\nu'-\frac12,n,\nu-\frac12,p)}{\Gamma(p+\frac32)}r^{2p}e^{-r^2}, \tag{B4}$$

where for the case of m and $\nu+\nu'$ odd, the coefficients B are those tabulated by Brody and Moshinsky.[22] It is immediate to extend the program to the calculation of the other cases of interest. Consequently we have

$$I = \sum_{p_1p_2}\frac{4B(n_1',\nu_1'-\frac12,n_1,\nu_1-\frac12,p_1)B(n_2',\nu_2'-\frac12,n_2,\nu_2-\frac12,p_2)}{\Gamma(p_1+\frac32)\Gamma(p_2+\frac32)}J\ , \tag{B5}$$

$$J = \int_0^\infty\int_0^\infty r_1^{2p_1+2}r_2^{2p_2+2}\frac{r_<^s}{r_>^{s+a}}\exp[-(r_1^2+r_2^2)]\,dr_1\,dr_2. \tag{B6}$$

To carry out the last integration we turn to polar coordinates so that we can write J as follows:

$$J=\int_0^\infty dR\,R^{2p_1+2p_2+5-a}\,e^{-R^2}\int_0^{\pi/4}[(\cos\alpha)^{2p_1+2}(\sin\alpha)^{2p_2+2}+(\sin\alpha)^{2p_1+2}(\cos\alpha)^{2p_2+2}]\,\frac{(tg\alpha)^s}{(\cos\alpha)^a}\,d\alpha=\frac{\Gamma(p_1+p_2+3-a/2)}{2}K. \quad \text{(B7)}$$

As a consequence of the Kronecker delta which appears in (B1), the sum p_1+p_2 is always half-integer if a is odd, and integer if a is even. Thus, doing the transformation $x=tg\alpha$ we can write the angular part of the integral in the following way:

$$K=\int_0^1\frac{x^{m_1}+x^{m_2}}{(l+x^2)^n}\,dx,\quad m_1=2p_1+s+2,\quad m_2=2p_2+s+2,\quad n=p_1+p_2+3-a/2. \quad \text{(B8)}$$

Using the formula 2.147.2 and, depending if m_1 and m_2 are even or odd, formula 2.148.4 or 2.124.2 of Ref. 9 we can obtain the final result for K.

Writing

$$K^{\pm}(m)=\int_0^1 x^m\,dx/(1+x^2)^n, \quad \text{(B9)}$$

where the index $\pm$ denotes the parity of m, + if m is even, − if m is odd, we obtain the following result:

$$K^{+}(m)=-\frac{1}{2^n}\sum_{k=1}^{m/2}\frac{\Gamma((m+1)/2)\Gamma(n-(m+1)/2)}{\Gamma(k+\frac12)\Gamma(n-k+\frac12)}+\frac{\Gamma((m+1)/2)\Gamma(n-(m+1)/2)}{2\sqrt{\pi}}\sum_{k=1}^{n-1}\frac{\Gamma(n-k)}{2^{n-k}\Gamma(n)\Gamma(n-k+\frac12)}$$
$$+\frac{\Gamma((m+1)/2)\Gamma(n-(m+1)/2)}{4\Gamma(n)},$$
$$K^{-}(m)=-\frac{1}{2^n}\sum_{k=1}^{(m-1)/2}\frac{\Gamma((m+1)/2)\Gamma(n-(m+1)/2)}{\Gamma(k+1)\Gamma(n-k)}-\frac{\Gamma((m+1)/2)\Gamma(n-(m+1)/2)}{2^n\Gamma(n)}+\frac{\Gamma((m+1)/2)\Gamma(n-(m+1)/2)}{2\Gamma(n)}. \quad \text{(B10)}$$

So for a even, p_1 integer, s even, or p_1 half-integer, s odd, we have

$$K=K^{+}(2p_2+s+2)+K^{+}(2p_1+s+2). \quad \text{(B11)}$$

For the other cases corresponding to a even, the value of K can be obtained changing K^+ into K^- in (B11).

For a odd, p_1 integer, s even, or p_1 half-integer, s odd, we have

$$K=K^{-}(2p_2+s+2)+K^{+}(2p_1+s+2), \quad \text{(B12)}$$

and for the other cases the value of K is given by interchanging p_1 and p_2 in (B12).

*Member of the Instituto Nacional de Energía Nuclear.
†Fellow of Mexican Italian Exchange Program sponsored by Consejo Nacional de Ciencia y Tecnología (CONACYT).
‡Member of El Colegio Nacional.

[1]M. Moshinsky, SIAM J. Appl. Math. 25, 193 (1973).
[2]A. Crubellier, Proceedings of the 3rd International Colloquim on Group Theoretical Methods in Physics, Marseille (1974); J. Phys. A (to be published).
[3]L. Armstrong, J. Math. Phys. 12, 953 (1971).
[4]L. Armstrong, Phys. Rev. A 3, 1546 (1971).
[5]M. Badawi, N. Bessis, G. Bessis, and G. Hadinger, Phys. Rev. A 8, 727 (1973); D. Herrick and O. Sinanoglu, Phys. Rev. A 5, 2309 (1972); N.V.V. Swamy, R.G. Kulkarni, and L.C. Biedenharn, J. Math. Phys. 11, 1165 (1970), and other references given in these papers.
[6]C. Quesne and M. Moshinsky, J. Math. Phys. 12, 1780 (1971).
[7]W. Miller, *Lie Algebras and Special Functions* (Academic, New York, 1968).
[8]M. Moshinsky, T.H. Seligman, and K.B. Wolf, J. Math. Phys. 13, 901 (1972).
[9]I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York, 1965).
[10]W.J. Holman and L.C. Biedenharn, Ann. Phys. 39, 1 (1969); W.J. Holman and L.C. Biedenharn, Ann. Phys. 47, 205 (1968).
[11]P.A.M. Dirac, *The Principles of Quantum Mechanics* (Oxford U.P., London, 1930), pp. 108—16.
[12]B. Gruber and A.U. Klymik, J. Math. Phys. 16, 1816 (1975).
[13]M.E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957).
[14]H.A. Jahn and J. Hope, Phys. Rev. 93, 318 (1954).
[15]E.P. Wigner, *Group Theory* (Academic, New York, 1959), p. 105.
[16]G. Racah, Phys. Rev. 62, 146 (1942).
[17]L.J. Slater, *Generalized Hypergeometric Functions* (Cambridge U.P., Cambridge, 1966), p. 116—20.
[18]H. Ui, Ann. Phys. (N.Y.) 49, 69 (1968).
[19]L. Armstrong, J. Phys. Paris 31, C4 (1970).
[20]E. Chacón, D. Levi, and M. Moshinsky, J. Math. Phys. 16, 1876 (1975).
[21]M. Moshinsky, *The Harmonic Oscillator in Modern Physics: From Atoms to Quarks* (Gordon & Breach, New York, 1969), p. 5.
[22]T.A. Brody and M. Moshinsky, *Tables of Transformation Brackets* (Gordon & Breach, New York, 1967).

ERRATA

Erratum: Proof of the charge superselection rule in local relativistic quantum field theory

[J. Math. Phys. 15, 2198 (1974)]

F. Strocchi

INFN, Sezione di Pisa, Scuola Normale Superiore, Pisa, Italy

A. S. Wightman

Joseph Henry Laboratory of Physics, Princeton University, Princeton, New Jersey 08540

(Received 7 July 1976)

In Eq. (1.8), the round brackets, (,), should be angular brackets, $\langle\, ,\, \rangle$.

At the end of Definition 2.3, add: In the following, we still say there is a special gauge transformation leading from $A_{1\mu}$ to $A_{2\mu}$ even if the vacuum in $\mathcal{H}_2$ is not a cyclic vector for $A_{2\mu}$.

At the end of Proposition 2.1, the phrase "satisfies (2.50)" should stand clear of hypothesis (iii).

On p. 2206, first column, line 14, replace $B(f)^*$ by $B(\bar{f})$.

On pp. 2207, 2209, and 2210, Laudau should be Landau, in six places.

In Eq. (2.111), $G(-x)$ should be $\Box G(-x)$.

On p. 2209, first column, third line from the bottom (line 3–), (2.44) should read (2.46), second column, line 8–, (2.117) should read (2.118).

On p. 2210 in Eq. (2.130), the minus sign should be plus.

On p. 2214, second column, line 5–, (2.168) should be (2.175).

On p. 2215, all script $\mathcal{A}$ should be Gothic $\mathfrak{A}$; same page, first column, line 14–, $\tilde{\mathfrak{F}}(\mathcal{O})$ should be $\mathfrak{F}(\mathcal{O})$; second column, line 5 $(\mathcal{O})$ should be $\mathfrak{F}(\mathcal{O})$.

On p. 2217, second column, line 14, $A_\mu + \partial_\mu \chi$ should be $A_\mu - \partial_\mu\chi$; line 23, $\partial_\mu A^\mu(f)A(g)$ should be $\partial_\mu A^{\mu(+)}(f)A(g)$; Eq. (2.188) should read

$$(B\Psi)^{(n)}_{\mu_1\cdots\mu_n} = b^{\nu_1}_{\mu_1}(k_1)\cdots b^{\nu_n}_{\mu_n}(k_n)\Psi^{(n)}_{\nu_1\cdots\nu_n}(k_1\cdots k_n),$$

where $k_\mu b^\mu_\nu(k) = k_\nu b(k)$, $k^\nu b^\mu_\nu(k) = k^\mu a(k)$ with $a, b\in \mathcal{S}(\mathbb{R}^4)$.

On p. 2218, first column, line 8, $\lfloor\Phi)'$ should be $\lfloor\Phi\rfloor$; second column, line 20–, $T(\underline{\Psi}_1, \Psi_2)$ should be $T(\underline{\Psi}, \underline{\Psi}_2)$.

On p. 2219, first column, line 2, $\langle\psi_1, \chi_2\rangle$ should be $\langle\Psi_1, \chi_2\rangle$.

On p. 2220, first column, line 25, "in one gauge$\cdots$ another." should read "in every gauge$\cdots$"; second column, line 3–, $\partial_\mu \mathbf{f}_{\mu\nu}$ should read $\partial^\mu \mathbf{f}_{\mu\nu}$.

On p. 2221, first column, add after (4.3): [the currents $\mathbf{J}_\nu(x)$ themselves being charged fields].

On p. 2222, first column, in (4.10) and (4.12), F should be G.

On p. 2223, second column, line 22, $\theta = 2\pi$ should be $\theta = \pi/2$.

On pp. 2210–11, the discussion of the $\Phi o \kappa$ space realization of the Landau gauge in (2.132)–(2.153) is not satisfactory for two reasons. First, because the matrix η given in (2.153) has eigenvalues greater than one. As a consequence, the form $\langle\, ,\, \rangle$ is not bounded with respect to (,). This difficulty is easily overcome by inserting a suitable constant N on the right-hand side of (2.132), N being larger than all the eigenvalues of the matrix η. Second, as pointed out to us by Professor G. Rideau, the transversality property (2.146) is not satisfied as an operator identity in the Hilbert space $\mathcal{H}$. [The somewhat confused argument after (2.146) reflects an earlier version of the paper in which a Landau gauge was defined as one for which the two-point function of the vector potential has zero divergence. That is also a legitimate definition, but it is not the one we have finally adopted.] A correct $\Phi o \kappa$ space realization has been given by Rideau in *Letters in Mathematical Physics* 1, 17 (1975). [See also L. Bracci and F. Strocchi, J. Math. Phys. **16**, 2522 (1975), Sec. 4.] Here is another, for the special case $M = 0$, which has some virtues of simplicity.

For any test function f_μ, define

$$\hat{F}_\mu(k) = (g_{\mu\nu}k^2 - k_\mu k_\nu)\hat{f}^\nu(k)$$

and the eight-component wavefunction

$$\Psi_f(k) = \begin{Bmatrix} k_0^{-2}\hat{F}_\mu(k) \\ \\ k_0^{-1}\dfrac{\partial}{\partial k_0}\hat{F}_\mu(k) \end{Bmatrix}.$$

The Hilbert space scalar product, $(\cdot,\cdot)$, and the invariant sesquilinear form, $\langle\cdot,\cdot\rangle$, are then defined as

$$(\Psi_f, \Psi_g) = N\int d\Omega_0(k)\sum_{\alpha=0}^{7}\overline{\Psi_{f\alpha}(k)}\Psi_{g\alpha}(k),$$

$$(\Psi_f, \Psi_g) = \int d\Omega_0(k)\sum_{\alpha,\beta=0}^{7}\overline{\Psi_{f\alpha}(k)}\eta_{\alpha\beta}\Psi_{g\beta}(k),$$

where η is the 8×8 matrix

$$\eta = \frac{1}{8}\begin{Bmatrix} g_{\mu\nu} + 3g_{\mu 0}g_{\nu 0} & -g_{\mu 0}g_{\nu 0} \\ \\ -g_{\mu 0}g_{\nu 0} & g_{\mu\nu} \end{Bmatrix}.$$

Because the largest eigenvalue of this η is $(5+\sqrt{13})/2$,

N should be chosen so that $N > (5+\sqrt{13})/2$. The vector potential is defined as in (2.140) with the $\Pi_{\pm}^{L}(0)f$ of (2.143) replaced by

$$(\Pi_{\pm}^{L}f)_{\mu}(k)=\sqrt{\pi}\begin{Bmatrix} k_0^{-2}\hat{F}_{\mu}(\pm k) \\ k_0^{-1}\dfrac{\partial}{\partial k_0}\hat{F}_{\mu}(\pm k)\end{Bmatrix}.$$

The discussion then runs parallel to that given in the paper.

Regrettably, neither in this realization nor in Rideau's is it obvious the operator η has an inverse with the properties which would enable one to carry out the argument (2.154)–(2.158). However, $c_4=0$ is implied by the following altered version. The extended scalar product is defined as in (2.154) [with the correction: there should be a bar over the $\hat{f}(0)$'s]. The new η is the old with one new row and column consisting of zeros except for the diagonal element $\operatorname{sgn}c_4\, g_{\mu\nu}$. The definition (2.155) is replaced by the definition of the four functionals

$$F_{\mu}(\bar{\Psi}_0)=0,\quad F_{\mu}(A(f)\bar{\Psi}_0)=\sqrt{|c_4|}\hat{f}_{\mu}(0),$$

$$F_{\mu}(:A(f_1)\cdots A(f_n):\Psi_0)=0,\quad n>1.$$

Because the F_{μ} are clearly bounded with respect to $(\ ,\ )^{(1)}$, there exist four vectors Φ_{μ} such that

$$F_{\mu}(A(f)\Psi_0)=(\Phi_{\mu},A(f)\Psi_0)=\sqrt{|c_4|}\hat{f}(0).$$

By the definition of

$$\langle\Phi_{\mu},A(f)\Psi_0\rangle=(\Phi_{\mu},\eta A(f)\Psi_0)=\operatorname{sgn}c_4\sqrt{|c_4|}\hat{f}_{\mu}(0).$$

Thus, the contribution $c_4\hat{f}(0)\hat{g}^{\mu}(0)$ in the two-point function comes from the Φ_{μ}'s:

$$\langle\Psi_0,A(f)A(g)\Psi_0\rangle-(2.130)$$

$$=\sum_{\mu}\langle\Psi_0,A(f)\Phi_{\mu}\rangle\langle\Phi_{\mu},A(g)\Psi_0\rangle$$

$$=c_4\hat{f}_{\mu}(0)\hat{g}^{\mu}(0).$$

Clearly,

$$\langle\Phi_{\mu},U(a)A(f)\Psi_0\rangle=\operatorname{sgn}c_4\ \sqrt{|c_4|}\,(e^{ip\cdot a}\hat{f})_{\mu}(0)$$

$$=\langle\Phi_{\mu},A(f)\Psi_0\rangle,$$

i.e.,

$$(\eta(U(a)\Phi_{\mu}-\Phi_{\mu}),A(f)\Psi_0)=0$$

so since the $A(f)\Psi_0$ span $\mathcal{H}^{(1)}$, and the metric is nondegenerate,

$$U(a)\Phi_{\mu}=\Phi_{\mu}.$$

This is impossible if the vacuum is the unique invariant vector.