Zeros of the partition function for higher-order spin systems

K. Y. Millard and K. S. Viswanathan

Department of Physics, Simon Fraser University, Burnaby, B. C., Canada V5A 1S6

(Received 30 May 1974)

A theorem is proved which is useful in determining information regarding the location of zeros of the partition function for lattice models with arbitrary spin. This theorem is a generalization to higher-order spin systems of a theorem for spin-1/2 systems proved by Ruelle. The total partition function for the system can be constructed by contracting (a generalization of the Asano contraction procedure) a set of lower-order "partition function-like" polynomials. The theorem presented relates information regarding the location of zeros of the lower-order polynomials to the location of zeros of the partition function. This theorem is then used to establish the Lee–Yang unit circle theorem for several higher-order spin models.

I. INTRODUCTION

The Yang and Lee[1] description of a phase transition in terms of the distribution of zeros of the partition function is of fundamental importance to the study of phase transitions. Of particular interest for magnetic systems is the location of the zeros of the partition function in the complex $z=\exp(\beta h)$ plane, where β and h are, respectively, the inverse temperature and the applied magnetic field. Lee and Yang[1] demonstrated that for the spin-$\frac{1}{2}$ Ising model, all the zeros of the canonical partition function lie on the unit circle in the complex z plane. This result was extended to the arbitrary-spin Ising model by Griffiths,[2] using a technique which transformed the arbitrary-spin problem into an analog spin-$\frac{1}{2}$ problem. The unit circle theorem has also been extended by Suzuki and Fisher[3] to several quantum systems, which include the anisotropic Heisenberg model for a class of ferromagnetic coupling parameters.

Ruelle[4] has recently presented an approach to locating regions of the complex z plane which contain no zeros of the partition function. The essence of this approach is to form appropriate "partition function-like" polynomials, involving a small number of lattice sites. The full partition function is then constructed by taking successive Asano[5] contractions of the suitably chosen polynomials. The surprising result is that information regarding the location of zeros of the several particle polynomials implies information regarding the location of zeros of the full partition function. Two intriguing features of Ruelle's approach are that it yields the Lee—Yang unit circle theorem for the spin-$\frac{1}{2}$ Ising model with little effort and also enables the investigation of noncircular regions of the complex z plane which are free of zeros of the partition function.

The Ruelle approach, however, has only been established for spin-$\frac{1}{2}$ magnetic systems or equivalently single component lattice gases. The application of this approach to arbitrary spin systems can take two possible forms: (1) converting the arbitrary-spin problem into an analog spin-$\frac{1}{2}$ problem by means of the Griffiths' transformation[2] or (2) appropriately generalizing Ruelle's technique to be applicable to arbitrary-spin problems directly. Alternative (1) has been pursued by the authors[6] to obtain upper bounds on critical temperatures and magnetic fields for several spin-1 systems. It is the purpose of this article to investigate alternative (2) by presenting an appropriate generalization of Ruelle's approach for arbitrary-spin systems and then using the technique developed to establish the unit circle theorem for several higher order spin systems.

The main theorem, which is the appropriate generalization of Ruelle's theorem to arbitrary-spin systems, is presented in Sec. II. The proof of this theorem is based on Ruelle's theorem for spin-$\frac{1}{2}$ systems and Laguerre's theorem.[7] Section III contains the application of the main theorem to establish the unit circle theorem for several higher order spin models.

II. THE MAIN THEOREM

For completeness, we state without proof the relevant theorems needed for the proof of the main theorem.

Theorem 1 (Laguerre's Theorem[8]): Let $f(z)$ be an nth degree polynomial such that $f(z)$ does not vanish for all $z\notin C$, where C is a closed circular region.[9] Then the first polar derivative of $f(z)$ with respect to ξ_1 defined by

$$f_1(z,\xi_1)=nf(z)+(\xi_1-z)f'(z), \tag{1}$$

does not vanish for $z\notin C$ and $\xi_1\notin C$.

The kth polar derivative of $f(z)$ is defined by

$$\begin{aligned} f_k(z;\xi_1,\ldots,\xi_k)=&(n-k+1)f_{k-1}(z;\xi_1,\ldots,\xi_{k-1})\\ &+(\xi_k-z)f'_{k-1}(z;\xi_1,\ldots,\xi_{k-1}) \end{aligned} \tag{2}$$

where the prime denotes differentiation with respect to z. The nth polar derivative, is then given by

$$f_n(\xi_1,\ldots,\xi_n)=f_{n-1}(z=\xi_n;\xi_1,\xi_2,\ldots,\xi_{n-1}). \tag{3}$$

By repeated application of Laguerre's Theorem, the following lemma is obtained.

Lemma 1: Let $f(z)$ be an nth degree polynomial such that $f(z)$ does not vanish for all $z\notin C$, where C is a closed circular region. Then the nth polar derivative $f_n(\xi_1,\ldots,\xi_n)$ does not vanish for $\xi_j\notin C$, $j=1,2,\ldots,n$.

From (2) and (3) we note that if

$$f(z)=\sum_{k=0}^{n}\binom{n}{k}a_kz^k, \tag{4}$$

then the nth polar derivative can be written as

$$f_n(\xi_1,\ldots,\xi_n)=n!\sum_{k=0}^{n}\sigma(n,k)a_k \tag{5}$$

where $\sigma(n,k)$ is the symmetric function consisting of the sum of all possible products of $\xi_1,\xi_2,\ldots,\xi_n$ taken k at a time. From (5) it immediately follows that

$$f_n(z,z,\ldots,z)=n!\,f(z). \tag{6}$$

Finally, we need a lemma due to Ruelle.[4]

Lemma 2: Let A and B be closed subsets of the complex plane which do not contain the origin. Suppose that the complex polynomial

$$g(z_1, z_2) = a + bz_1 + cz_2 + dz_1z_2 \tag{7}$$

does not vanish for $z_1 \notin A$ and $z_2 \notin B$. Then

$$\tilde{g}(z) = a + dz \tag{8}$$

does not vanish for $z \notin -AB$.[10]

We are now in a position to state and prove the main theorem.

Theorem 2: Let

$$f(z_1, z_2) = \sum_{k_1=0}^{n} \sum_{k_2=0}^{n} A_{k_1k_2} \binom{n}{k_1}\binom{n}{k_2} z_1^{k_1} z_2^{k_2} \tag{9}$$

be a complex polynomial of degree n in each z_1 and z_2. Suppose C_1 and C_2 are closed circular regions in the complex plane such that $0 \notin C_1$ and $0 \notin C_2$ and

$$f(z_1, z_1) \neq 0 \quad \text{for all } z_1 \notin C_1 \text{ and } z_2 \notin C_2.$$

Then the polynomial

$$f(z) = \sum_{k=0}^{n} A_{kk}\binom{n}{k} z^k \tag{10}$$

does not vanish for $z \notin -C_1C_2$.

Proof: First write $f(z_1, z_2)$ as a polynomial in z_1,

$$f(z_1, z_2) = \sum_{k_1=0}^{n} B_{k_1}(z_2) \binom{n}{k_1} z_1^{k_1} \tag{11}$$

where

$$B_{k_1}(z_2) = \sum_{k_2=0}^{n} \binom{n}{k_2} A_{k_1k_2} z_2^{k_2}. \tag{12}$$

From (5) it follows that the nth polar derivative of $f(z_1, z_2)$ with respect to z_1 can be written as

$$\begin{aligned} f_n(\xi_1, \ldots, \xi_n; z_2) &= n! \sum_{k_1=0}^{n} \sigma(n, k_1) B_{k_1}(z_2) \\ &= n! \sum_{k_1=0}^{n} \sum_{k_2=0}^{n} \sigma(n, k_1) A_{k_1k_2} \binom{n}{k_2} z_2^{k_2}. \end{aligned} \tag{13}$$

By assumption $f(z_1, z_2) \neq 0$ for $z_1 \notin C_1$ and $z_2 \notin C_2$. Therefore, by Lemma 1, $f_n(\xi_1, \ldots, \xi_n; z_2) \neq 0$ for $\xi_j \notin C_1$, $j = 1, 2, \ldots, n$, and $z_2 \notin C_2$. Now treat $f_n(\xi_1, \ldots, \xi_n; z_2)$ as a polynomial in z_2 and take the nth polar derivative with respect to z_2 to obtain

$$\begin{aligned} &f_{n;n}(\xi_1, \ldots, \xi_n; \rho_1, \ldots, \rho_n) \\ &\quad = (n!)^2 \sum_{k_1=0}^{n} \sum_{k_2=0}^{n} \sigma(n, k_1)\pi(n, k_2) A_{k_1k_2} \end{aligned} \tag{14}$$

where $\pi(n, k_2)$ is the symmetric function consisting of the sum of all products of $\rho_1, \ldots, \rho_n$ taken k_2 at a time. Again applying Lemma 1, we obtain that $f_{n;n} \neq 0$ for $\xi_j \notin C_1$, $i = 1, \ldots, n$, and $\rho_l \notin C_2$, $l = 1, \ldots, n$. Now, using the Asano contraction procedure, i.e., contracting the pairs (ξ_j, ρ_j), $j = 1, 2, \ldots, n$, as described in Lemma 2, we obtain the function

$$\begin{aligned} &\tilde{f}_{n;n}(\xi_1, \ldots \xi_n) \\ &\quad = (n!)^2 \sum_{k=0}^{n} \sigma(n, k) A_{kk}. \end{aligned} \tag{15}$$

But, by Lemma 2, $\tilde{f}_{n;n}$ does not vanish for $\xi_k \notin -C_1C_2$, $k = 1, 2, \ldots, n$. Finally, setting $\xi_k = z$, for all k, we obtain that

$$f(z) = (n!)^{-2}\tilde{f}_{n;n}(\xi_1 = z, \ldots, \xi_n = z) = \sum_{k=0}^{n} \binom{n}{k} A_{kk} z^k \tag{16}$$

does not vanish for $z \notin -C_1C_2$, which completes the proof.

The contraction defined by (9) and (10) of Theorem 2 provides an appropriate generalization to higher order spin systems of the spin-$\frac{1}{2}$ Asano contraction given by (7) and (8). We now illustrate the potential usefulness of this procedure by using Theorem 2 to establish the unit circle theorem for several higher order spin systems.

III. APPLICATIONS

It is perhaps useful to indicate how Theorem 2 is used to determine information about the location of zeros of the partition function. Consider a lattice consisting of N Ising spin sites, the kth site being a spin-S_k site. The kth site can then be in any of the $2S_k + 1$ states enumerated by $\sigma_k = -S_k, -S_k + 1, \ldots, +S_k$. Split the Hamiltonian H for the system into the interaction with the magnetic field h_k at the kth site ($k = 1, 2, \ldots, N$) plus the remaining interaction $\tilde{H}$

$$H = \tilde{H}(\{\sigma_k\}) - \sum_{k=1}^{N} h_k\sigma_k. \tag{17}$$

The canonical partition function is then given by

$$Q(\beta, \{z_l\}, N) = \sum_{\{\sigma_k\}} \exp\{-\beta\tilde{H}\} \prod_{j=0}^{N} z_j^{\sigma_j} \tag{18}$$

where

$$z_j = \exp(\beta h_j). \tag{19}$$

The quantity

$$\hat{Q}(\beta, \{z_l\}, N) = \left(\prod_{j=0}^{N} z_j^{S_j}\right) Q(\beta, \{z_l\}, N) \tag{20}$$

is then a polynomial of order $2S_k$ in the variable $z_k (k = 1, 2, \ldots, N)$. Now, choose (the choice is *not* unique) a set of polynomials $q_\alpha(\{z_j^{(\alpha)}\})$ such that upon taking the generalized Asano contraction (as described in Theorem 2) of the product

$$\prod_\alpha q_\alpha(\{z_j^{(\alpha)}\}),$$

one obtains the modified partition function $\hat{Q}$. The contractions are of course taken between all pairs of parameters of the form $(z_j^{(\alpha)}, z_j^{(\gamma)})$, $j = 1, 2 \cdots N$. Symbolically, one has

$$\prod_\alpha q_\alpha(\{z_j^{(\alpha)}\}) \xrightarrow[\substack{\text{generalized} \\ \text{Asano contraction}}]{} \hat{Q}(\beta, \{z_j\}, N). \tag{21}$$

If one has information regarding the zeros of the polynomials q_α, Theorem 2 then yields information regarding the zeros of the full partition function. For this technique to be useful, one must make a judicious choice of the q's. Below, we present several examples for which this technique is useful to establish that all the zeros of the partition function lie on the unit circle in the complex z plane.

A. Unit circle theorem for a modified Ising model

In this section we consider the modified Ising Hamiltonian

$$H=-\sum_{i<j} J_{ij}\sigma_i\sigma_j-\sum_j \phi_j(\sigma_j)-\sum_j h_j\sigma_j. \tag{22}$$

The first term on the right-hand side of (23) is the usual Ising interaction. In the second term, the ϕ can be thought of as a "chemical potential" for different states of a spin site. We also have in mind eventually setting all $h_i=h$. The simplest choice for the "few particle" partition functions, i. e., the q_α's of (21), is to choose a "pair" partition function for each pair of sites coupled by the Ising interaction and a "single particle" partition function to accommodate each "chemical potential" term in (22). The "few particle" partition functions corresponding to this choice are conveniently written as:

$$q_{kl}(z_k,z_l)=\sum_{\sigma_k=-S_k}^{S_k}\ \sum_{\sigma_l=-S_l}^{S_l}\binom{2S_k}{S_k+\sigma_k}\binom{2S_l}{S_l+\sigma_l}$$
$$\times\exp\{\beta J_{kl}\sigma_k\sigma_l\}z_k^{S_k+\sigma_k}z_l^{S_l+\sigma_l} \tag{23}$$

and

$$q_k(z_k)=\sum_{\sigma_k=-S_k}^{S_k}\exp\{\beta\phi_k(\sigma_k)\}z_k^{S_k+\sigma_k}. \tag{24}$$

Note that the binomial coefficients (see Theorem 2) are included in (22) but not in (24), so that upon taking the generalized Asano contractions, we obtain the canonical partition function, i. e.,

$$\prod_{k<l} q_{kl}\prod_j q_j \xrightarrow[\text{Asano contraction}]{\text{generalized}} \hat{Q}.$$

To establish that the partition function $Q(\beta,z,N)$ has all zeros on the unit circle, it is sufficient to show that the q_α's do not vanish if all $\{z_l^{(\alpha)}\}$ are contained within the unit circle. This follows by the repeated application of Theorem 2 as described above, where C_1 and C_2 are taken to be the same closed circular region C, the exterior to the unit circle in the complex plane. The proof of this statement for the "pair" partition function is given in the following lemma.

Lemma 3: If $|z_k|<1$ and $|z_i|<1$, then $q_{kl}(z_k,z_l)$ defined by (23) does not vanish provided $J_{kl}\geqslant 0$.

Proof: First define $n_l=S_l+\sigma_l$ and rewrite (23) as

$$\hat{q}_{kl}(z_k,z_l)=\exp\{-\beta J_{kl}S_kS_l\}q_{kl}(z_k,z_l)$$
$$=\sum_{n_l=0}^{2S_l}\binom{2S_l}{n_l}(1+\exp\{\beta J_{kl}(n_l-S_l)\}z_k)^{2S_k}$$
$$\times(\exp\{-\beta J_{kl}S_k\}z_l)^{n_l}. \tag{25}$$

Now, consider $\hat{q}_{kl}(z_k,z_l)$ to be the successive contraction with respect to z_l of the product of polynomials of the form

$$h(z_k,z_l^{(r)})=\sum_{n_l=0}^{2S_l}\binom{2S_l}{n_l}(1+\exp\{\beta J_{kl}(n_l-S_l)\}z_k)$$
$$\times(\exp\{-\tfrac{1}{2}\beta J_{kl}\}z_l^{(\gamma)})^{n_l} \tag{26}$$

where $\gamma=1,2,\ldots,2S_k$. But if $h(z_k,z_l^{(\gamma)})$ vanishes, then

$$|z_k|=\left|\frac{1+\exp\{-\tfrac{1}{2}\beta J_{kl}\}z_l^{(\gamma)}}{\exp\{-\tfrac{1}{2}\beta J_{kl}\}+z_l^{(\gamma)}}\right|^{2S_l}. \tag{27}$$

This expression implies that if $h=0$, $J_{kl}\geqslant 0$, and $|z_k|<1$, then $|z_l^{(\gamma)}|>1$. We therefore conclude that if $J_{kl}\geqslant 0$, $|z_k|<1$, and $|z_l^{(\gamma)}|<1$ then $h(z_k,z_l^{(\gamma)})$ does not vanish. By successive contractions with respect to $z_l^{(\gamma)}$ of

$$\prod_{\gamma=1}^{2S_k} h(z_k,z_l^{(\gamma)})$$

and using Theorem 2, the lemma is proved.

We now examine the zeros of the "single particle" partition function given by (24). To complete the proof of the unit circle theorem for the Hamiltonian (22) it is sufficient to establish conditions for which (24) has no zeros within the unit circle. For any example, one could use the Schur–Cohn criterion[11] to determine the condition for which (24) has no zeros within the unit circle. We will, however, only examine the zeros of (24) for several special cases.

Case (i): If $\phi_k(\sigma_k)=0$ for $\sigma_k=-S_k,\ldots,+S_k$, then the Hamiltonian (22) corresponds to the usual Ising model with arbitrary spin. For this case, we obtain

$$q_k(z)=\sum_{n=0}^{2S}z^n=\frac{1-z^{2S+1}}{1-z} \tag{28}$$

which clearly has all zeros on the unit circle. This, together with Lemma 3 and Theorem 2, implies that all the zeros of the canonical partition function of the arbitrary spin Ising model ($J_{kl}\geqslant 0$) lie on the unit circle in the complex z plane. This result has previously been established by Griffiths.[2]

Case (ii): We now consider the dilute Ising model with arbitrary spin. In this model, each lattice can be either occupied by a magnetic atom or be unoccupied. Each occupied site contributes a weighting factor $\exp(\beta\mu)$ to the partition function, where μ is the chemical potential. For this model, we treat the integral and half-integral spin values separately.

Case (iia): Suppose $2S_k$ is an even integer and

$$\phi_k(\sigma)=\begin{cases}\beta^{-1}\ln(1+e^{\beta\mu}), & \text{for } \sigma=0\\ \mu, & \text{otherwise.}\end{cases}$$

Then, the Hamiltonian (22) corresponds to the dilute Ising model with arbitrary integral spin values. The factor μ represents the chemical potential for an occupied site. For this case, (24) can be written as

$$q_k(z)=\exp(\beta\mu)\sum_{n=0}^{2S}z^n+z^S=\frac{\exp(\beta\mu)}{1-z}$$
$$\times\{1+\exp(-\beta\mu)z^S-\exp(-\beta\mu)z^{S+1}-z^{2S+1}\}. \tag{29}$$

But, it follows from Theorem II of Suzuki[12] that for $\mu>0$, all the zeros of the numerator of (29) lie on the unit circle. We therefore conclude that all the zeros of the partition function for the dilute Ising model ($J_{kl}\geqslant 0$, $\mu>0$) with arbitrary integral spin values lie on the unit circle in the complex z plane.

Case (iib): Consider the dilute Ising model with half-integral spin values. This example is slightly complicated by the "single-particle" partition function not being a polynomial in z but in $z^{1/2}$. We can circumvent this difficulty by introducing the parameters $\sigma_k' = 2\sigma_k$ where $\sigma_k' = -2S_k, -2S_k + 1, \ldots, +2S_k$. Clearly, we have introduced too many states for each site. However, the extraneous states are eliminated by our choice of the "single-particle" partition function. We have gained, however, a convenient designation for the unoccupied site, i.e., $\sigma_k' = 0$. We now choose as the "single-particle" partition function

$$g_k(z) = \exp(\beta\mu) \sum_{n \text{ even}}^{4S} z^n + z^{2S}$$
$$= \frac{\exp(\beta\mu)}{1 - z^2} \{1 + \exp(-\beta\mu) z^{2S} - \exp(-\beta\mu) z^{2S+2} - z^{2(2S+1)}\}. \tag{30}$$

This choice corresponds to choosing ϕ_k such that

$$\phi_k(\sigma') = \begin{cases} \infty, & \text{if } \sigma' \text{ is an odd integer} \\ 0, & \text{if } \sigma' = 0 \\ \mu, & \text{otherwise.} \end{cases}$$

The choice of $\phi_k(\sigma') = \infty$ for σ' an odd integer eliminates (under contraction) all the extraneous states introduced. Again using Theorem II of Suzuki[12] we find that $g_k(z)$ has all zeros on the unit circle if $\mu > 0$. We therefore conclude that all the zeros of the partition function for the dilute Ising model ($J_{kl} \geq 0$, $\mu > 0$) with arbitrary half-integral spin values lie on the unit circle in the complex z plane. We might comment that for the special case $S_k = \frac{1}{2}$

$$g_k(z) = \exp(\beta\mu)(1 + \exp(-\beta\mu) z + z^2)$$

has all zeros on the unit circle for

$$\beta\mu > -\ln 2.$$

This result has previously been obtained by Suzuki.[13]

Case (iii): Suppose $\phi(\sigma_k) = \phi(-\sigma_k)$ and

$$\phi(S_k) \geq \phi(S_k - 1) \geq \cdots \geq \begin{Bmatrix} \phi(0) \\ \phi(\frac{1}{2}) \end{Bmatrix}.$$

This choice corresponds to a model for which Suzuki[14] (using a different method) has established the unit circle theorem. In this case we choose

$$q_j(z) = \sum_{n=0}^{2S} a_n z^n \tag{31}$$

where

$$a_n = \exp[\beta\phi(n - S)].$$

From the above conditions, we observe that

$$a_j = a_{2S-j} \tag{32}$$

and

$$a_j > a_{j+1}, \quad j = 0, 1, \ldots, [S]. \tag{33}$$

where $[S]$ = largest integer less than or equal to S. The following lemma establishes the unit circle theorem for this case.

Lemma 4: If $q_k(z)$ satisfies (31)—(33) then all zeros of q_k lie on the unit circle in the complex z plane.

Proof: We present a proof for $2S$ even, a similar proof holds for $2S$ odd. The proof is accomplished by demonstrating that there are $2S$ distinct solutions of $q_k(z) = 0$ of the form $z = \exp(i\theta)$, i.e., there are $2S$ distinct values of θ satisfying $q_k(\exp(i\theta)) = 0$. By introducing the coefficients

$$b_0 = \tfrac{1}{2} a_S, \tag{34}$$

$$b_j = a_{S-j} \quad \text{for } j = 1, \ldots, n, \tag{35}$$

we obtain the identity

$$\exp(-iS\theta) q_k(\exp(i\theta)) = g(\exp(i\theta)) + g^*(\exp(i\theta))$$
$$= 2 \operatorname{Re} g(\exp(i\theta)) \tag{36}$$

where

$$g(z) = \sum_{l=0}^{S} b_l z^l. \tag{37}$$

From (36) we observe that if $q_k(\exp(i\theta)) = 0$, then $\operatorname{Re} g(\exp(i\theta)) = 0$ or g must be purely imaginary. To complete the proof we only need show that there are $2S$ distinct values of θ, $0 \leq \theta < 2\pi$, for which $g(\exp(i\theta))$ is purely imaginary.

From (33), (34), and (35) we observe that

$$b_j < b_{j+1}, \quad j = 0, \ldots, n-1. \tag{38}$$

[Note that for $2S$ even we can include the possibility $\frac{1}{2} a_S < a_{S-1}$ and still have (38) satisfied.] Using a theorem from Polyá and Szegö,[15] we conclude from (38) that all the zeros of $g(z)$ are contained within the unit circle, $|z| < 1$. Then using the Principle of Argument[16] we observe that as z traverses the unit circle in the complex z plane, $w = g(z)$ winds about the point $w = 0$ in the complex w plane S times, there being S zeros of $g(z)$ contained within the unit circle of the complex z plane. But, each time g winds about the origin, it crosses the imaginary axis in the w plane twice, or a total of $2S$ times for $0 \leq \theta < 2\pi$. Therefore, there are $2S$ distinct values of θ for which $g(\exp(i\theta))$ is purely imaginary or $q_k(z)$ has $2S$ roots on the unit circle.

B. Unit circle theorem for the Lebowitz-Gallavotti model #3[17]

We now consider the conditions for the zeros of the partition function to lie on the unit circle in the complex z plane for the spin-one lattice model with Hamiltonian

$$H = -J \sum_{k<l} \sigma_k \sigma_l (1 - \sigma_k \sigma_l) - \mu \sum_k \sigma_k^2 - h \sum_k \sigma_k \tag{39}$$

where $\sigma_k = 1, 0, -1$ for $k = 1, 2, \ldots, N$, and the first sum on the right-hand side is over nearest-neighbor sites. This is model #3 introduced by Lebowitz and Gallavotti.[17] The authors[6] have established a sufficient condition for the zeros to lie on the unit circle by first converting the model to an analog spin-$\frac{1}{2}$ model using the Griffith transformation.[2] Here, by employing the generalized Asano contraction and Theorem 2, we establish a sufficient condition which includes the previous result.

We choose as our "few particle" partition function for this model, the "pair" function[18]

$$q_{kl}(z_k, z_l) = a_0 + a_1 z_k + a_2 z_k^2 \tag{40}$$

where

$$a_0 = \alpha z_l^2 + \epsilon z_l + 1, \tag{41a}$$

$$a_1 = \epsilon(z_l^2 + \epsilon z_l + 1), \tag{41b}$$

$$a_2 = z_l^2 + \epsilon z_l + \alpha, \tag{41c}$$

and

$$\alpha = \exp(-2\beta J), \tag{42a}$$

$$\epsilon = 2^{(2\nu-1)/2\nu} \exp(-\beta\mu/2\nu). \tag{42b}$$

In these expressions 2ν is number of nearest-neighbor sites (for a square or cubic lattice ν = dimensionality of lattice). The $2^{(2\nu-1)/2\nu}$ in (42a) is included to correct for the binomial coefficients appearing in Theorem 2.

In order to establish a condition for the partition function in this model to have all its zeros on the unit circle, it is sufficient to establish that $q_{kl}(z_k, z_l)$ does not vanish if z_k and z_l are both contained within the unit circle. A convenient method to establish this condition is to use the Schur–Cohn criterion. For a second degree polynomial, the Schur–Cohn criterion states[11]:

The polynomial $f(z) = a_0 + a_1 z + a_2 z^2$ has all its zeros outside the unit circle, provided

$$\Delta_1 = |a_0|^2 - |a_2|^2, \tag{43a}$$

$$\Delta_2 = (\Delta_1)^2 - |a_0 a_1^* - a_1 a_2^*|^2 \tag{43b}$$

are positive.

We now consider (40) as a polynomial in z_k and establish the condition for Δ_1 and Δ_2 to be positive, assuming $|z_l| < 1$. Therefore, define

$$z_l = re^{i\theta}, \quad 0 \leq r < 1. \tag{44}$$

The quantity Δ_1 can then be written as

$$\Delta_1 = (\alpha^2 - 1)(r^2 - 1)\left(r^2 + 2\frac{\epsilon\cos\theta}{\alpha + 1} r + 1\right) \tag{45}$$

which is positive for all θ and r $(0 \leq r < 1)$ provided

$$\alpha < 1 \tag{46}$$

and

$$\epsilon < \alpha + 1. \tag{47}$$

Expression (46) is equivalent to the requirement

$$J > 0.$$

In similar manner one can establish that Δ_2 is positive provided

$$g(u, w) = a_{11} w^2 + a_{22} u^2 + 2a_{12} uw + a_{33} \tag{48}$$

is positive for (u, w) contain in the domain $\mathcal{D}$, defined by $\mathcal{D}$: $\{-1 \leq u \leq 1, \ w \geq 2\}$. In obtaining (48) we have made the substitutions

$$u = \cos\theta \quad \text{and} \quad w = r + 1/r.$$

The coefficients on (48) are given by

$$a_{11} = (\alpha + 1)^2 - \epsilon^2, \tag{49a}$$

$$a_{22} = 4\epsilon^2, \tag{49b}$$

$$a_{12} = \epsilon[2(\alpha + 1) - \epsilon^2], \tag{49c}$$

$$a_{33} = -\epsilon^4. \tag{49d}$$

Expression (48) is the equation for a conic section. In establishing that $g(u, w)$ is positive for (u, w) contained in $\mathcal{D}$, it is sufficient to show that the conic section $g(u, w) = 0$ does not intersect the domain $\mathcal{D}$. After some algebra, one can show that this condition is satisfied for the following situations:

$$\epsilon^2 < 2\alpha \quad \text{for } 1 > \alpha \geq \tfrac{1}{2}, \tag{50a}$$

$$4\alpha^2 + 8(1 - \epsilon)\alpha + (4 - 8\epsilon + 4\epsilon^3 - \epsilon^4) > 0 \quad \text{for } \alpha < \tfrac{1}{2}, \tag{50b}$$

$$\epsilon < 2 \quad \text{for } \alpha = 1. \tag{50c}$$

The condition (50c) for $\alpha = 1$ is obtained directly from the exact partition function [the Schur–Cohn criterion fails for this limiting case as Δ_1 and Δ_2 defined in (43) are both identically zero.] Using the Shur–Cohn criterion, we conclude that the model given by (39) has all zeros of the partition function on the unit circle in the complex z plane, provided (50) is satisfied. Expression (5) includes as special cases the results[6] previously obtained for this model.

IV. CONCLUDING REMARKS

In this article, we have presented a generalization to arbitrary spin of a theorem due to Ruelle regarding the zeros of the partition function. To illustrate the potential usefulness of this theorem, we have used it to establish the Lee–Yang unit circle theorem for several higher-order spin models. Some of these results are new and some reproduce results obtained by other methods. One feature that does stand out, however, is that once one has the basic theorem, the applications follow with relative ease. It is hoped that this technique will find further application.

ACKNOWLEDGMENTS

The authors wish to thank Dr. D. Ruelle for suggesting a proof for Theorem 2, similar to the one presented here. The financial support by the National Research Council of Canada is also gratefully acknowledged.

[1] C.N. Yang and T.D. Lee, Phys. Rev. **87**, 404 (1952); T.D. Lee and C.N. Yang, Phys. Rev. **87**, 410 (1952).

[2] R B. Griffiths, J. Math. Phys. **10**, 1559 (1969).

[3] M. Suzuki and M.E. Fisher, J. Math. Phys. **12**, 235 (1971).

[4] D. Ruelle, Phys. Rev. Lett. **26**, 303 (1971).

[5] T. Asano, J Phys. Soc. Japan **29**, 350 (1970).

[6] K. Millard and K.S. Viswanathan, Phys. Rev. B **9**, 2030 (1974).

[7] M. Marden, *Geometry of Polynomials, Mathematical Surveys*, No. 3 (American Mathematical Society, Providence, Rhode Island, 1966).

[8] This statement of Laguerre's theorem is a slight variation of Theorem (13.1) of Ref. 7.

[9] A closed circular region is the closed interior or exterior of a circle or a closed half-plane.

[10] -AB is the set of points $-z_1 z_2$ with $z_1 \in A$, $z_2 \in B$.

[11] See Sec. 43 of Ref. 7.

[12] M. Suzuki, Prog. Theor. Phys. **41**, 1438 (1969).

[13] M. Suzuki, Prog. Theor. Phys. **40**, 1246 (1968).

[14] M. Suzuki, J. Math. Phys. **14**, 1088 (1973).

[15] G. Polyá and G. Szegö, *Problems and Theorems in Analysis* (Springer, New York, 1972), Vol. I, p. 107, Theorem 23.

[16] See Theorem (1,2) of Ref. 7.

[17] J. L. Lebowitz and G. Gallavotti, J. Math. Phys. **12**, 1129 (1971).

[18] Note that $q_{kl}(z_k, z_l)$ is given only up to a multiplicative constant.

Double spectral representations of single loop amplitudes with k vertices: $k \geq 4$

J. S. Frederiksen

Institute for Theoretical Physics, University of Groningen, The Netherlands

(Received 22 May 1973)

A method developed in several previous papers is combined with the method of induction to derive double dispersion relations, with Mandelstam boundary, for the class of single loop amplitudes with four or more vertices. The spectral functions are expressed as integral representations and restrictions on the masses and kinematic invariants for which dispersion relations are valid are found. It is also discussed how representations for the low order single loop amplitudes can be obtained for wider ranges of these variables.

1. INTRODUCTION

Since the late 1950's, when it became apparent that dispersion relations for nuclen nucleon scattering and the nucleon electromagnetic form factor could not be proved on the basis of the general principles of field theory,[1] there has been a flood of literature on the analytic properties of Feynman amplitudes.[2-4] Because of their relative simplicity, special attention was initially devoted to the study of the low order single loop amplitudes in ϕ^3 theory. These investigations led to the introduction of some important new concepts. In particular Karplus, Sommerfield, and Wichmann,[5] Nambu,[6] and Oehme[7] discovered the anomalous threshold of the triangle diagram vertex function and Mandelstam[8] showed that, for a restricted range of masses, the box diagram amplitude satisfies the famous double spectral representation that bears his name.[9] The single loop diagrams have also played a central role in the majorization procedure,[2-4] which is aimed at showing that all Feynman amplitudes contributing to a particular process involving a given number of external particles are regular functions in a domain whose extent is determined by one or more of the simple diagrams.

With the advent of the Landau—Cutkosky rules[10,11] it became possible, in principle, to determine the singularities of a general Feynman integral and as well to obtain the discontinuities across the corresponding branch cuts. While these rules have been enormously useful in studying the analytic properties of Feynman amplitudes[2-4] and for obtaining heuristic dispersion relations for certain processes,[12,13] they are, by themselves, not sufficient for a rigorous derivation of dispersion relations. One of the main problems is that they do not determine on which Riemann sheets the singularities lie,[13-15] and in particular which singularities lie on the physical sheet. Further the discontinuity can in general only be determined up to a sign factor.

To overcome these problems, Fotiadi, Froissart, Lascoux, and Pham[16] proposed in 1963 that homology theory be used as a rigorous way of studying the analytic properties of individual Feynman integrals. Again, the investigations made using this method have been mainly restricted to the single loop diagrams and especially the low order single loop diagrams,[17] since the application of homology theory to more complicated diagrams has proved to be much more difficult.[18]

It is the aim of this paper to show that, within ϕ^3 theory, double dispersions relation with Mandelstam boundary can, for a restricted range of masses and kinematic invariants, be proved for any Feynman amplitude arising from a single loop diagram with four or more vertices. Further we obtain integral representations for the weight functions and discuss the significance of the above restrictions on the masses and kinematic invariants. The method used to derive these results is a combination of a method developed in several previous papers[19-21] with the method of induction; it involves the direct transformation of the Feynman parametrized form of the kth order single loop amplitude ($k \geqslant 4$) into the required form. (Refs. 19, 20, 21, are referred to as VF, I and P respectively.)

In Sec. 2 the first of the two Cauchy kernels needed for the double dispersion relation is introduced by changing the variables in the Feynman parametrized form of the kth order single loop amplitude. The restrictions on the masses and kinematic invariants for which this new form of the amplitude is valid are also discussed in this section. The boundary of the region of integration in the multiple integral representation derived in Sec. 2 is studied in Sec. 3, and in Sec. 4 we obtain some results necessary for reversing the orders of integration.

In Sec. 5, the orders of some of the integrations are reversed and the second Cauchy kernel is introduced by changing one of the variables of integration. The boundary of the region of integration in the resultant new multiple integral representation is studied in Sec. 6 and in Sec. 7 further results necessary for the reversal of the orders of integration are obtained.

Finally in Sec. 8 the required double dispersion relation for the kth order single loop amplitude is derived by changing the orders of integration in the integral representation obtained in Sec. 5. The spectral function is expressed in the form of a multiple integral, and it is found that the boundary of the double spectral representation is the usual Mandelstam boundary for the box diagram amplitude. In this section we also discuss in detail the implications of the restrictions on the masses and kinematic invariants made in Sec. 2 and how these restrictions may to a certain extent be relaxed.

2. TRANSFORMATION OF k^{th} ORDER SINGLE LOOP AMPLITUDE: $k \geqslant 4$

With plane wave states normalized, so that $\langle \mathbf{p}' | \mathbf{p} \rangle = \delta^3(\mathbf{p}' - \mathbf{p})$, we define the scalar invariant amplitude T_k for the multiparticle production process in which i initial particles produce $f = (k - i)$ final particles by

$$\langle \mathbf{p}_1, \ldots, \mathbf{p}_i | S-1 | -\mathbf{p}_{i+1}, \ldots, -\mathbf{p}_k\rangle$$
$$= -i(2\pi)^4\delta^{(4)}\left(\sum_{i=1}^{k} p_i\right)(2\pi)^{-3k/2}2^{-k/2}\left(\prod_{i=1}^{k} E_i\right)^{-1/2}T_k. \tag{1}$$

Our object is to show that the contribution to T_k from the kth order single-loop diagram shown in Fig. 1 or equivalently in Fig. 2 can, for a restricted range of masses and kinematic invariants, be written as a double spectral representation with Mandelstam boundary. Further we shall obtain an integral representation for the spectral function.

The k external momenta shown in Figs. 1 and 2 are labeled by the subscripts of the adjacent internal masses and the external mass squared of a particular external line is, of course, just the square of the external momentum of that line. The other variables on which the single loop amplitude depends are most conveniently defined in terms of the external momenta in Fig. 1 by

$$q_{ij}^2 = \left(\sum_{r=1}^{j-1} q_{r,r+1}\right)^2 \quad (1 \leqslant i < j \leqslant k). \tag{2}$$

It should be noted that when $k \geqslant 6$ the above kinematic invariants are not independent but satisfy algebraic constraints.[22] Notice also that when $j = i+1$ so that q_{ij}^2 is an external mass squared, Eq. (2) becomes an identity. Further, with the powers 2 removed, Eq. (2) is just the energy momentum conservation law when $i=1$, $j=k$ (since $-q_{1k}$ rather than q_{1k} is the ingoing 4-momentum in Fig. 1).

We shall find, however, that many of the subsequent expressions needed to obtain the double spectral representations take a simpler form in terms of the asymmetrically labelled variables shown in Fig. 2. The relations between the two sets of variables can be seen from Figs. 1 and 2. With $n = k-2$, they are as follows:

$$m_{-1} = M_1, \quad m_0 = M_{(k-1)}, \quad m_1 = M_k,$$
$$p_{-10} = q_{1(k-1)}, \quad p_{-11} = q_{1k}, \quad p_{01} = q_{(k-1)k},$$
$$m_j = M_j, \quad p_{-1j} = q_{1j}, \quad p_{1j} = q_{jk}, \quad p_{0j} = q_{j(k-1)},$$
$$(2 \leqslant j \leqslant k-2 = n)$$
$$p_{ij} = q_{ij} \quad (2 \leqslant i < j \leqslant k-2 = n), \tag{3}$$

where the q_{ij} are given in Eq. (2). We now define

$$y_{ij} = -(2m_i m_j)^{-1}[p_{ij}^2 - m_i^2 - m_j^2] \quad (-1 \leqslant i < j \leqslant n)$$
$$y_{ij} = y_{ji}, \quad y_{ii} = 1 \tag{4}$$

and as well

$$x_1 \equiv -y_{-10}, \quad x_2 \equiv -y_{12}. \tag{5}$$

We shall write the dispersion relations in x_1 and x_2 which are linearly related to the usual Mandelstam variables s and t.

Then, using standard Feynman rules[23] (see also Ref. 24 and Section 1.5 of Ref. 2) we find that the amplitude arising from the kth order single loop diagram shown in Fig. 2 takes the form

$$T_{(n+2)\,\text{loop}}(y_{ij}) = \frac{g}{16\pi^2}\cdot\frac{(-1)^n}{2m_{-1}m_0m_1m_2(n-1)!}I_{n+2}(y_{ij}), \tag{6}$$

where

$$I_{n+2}(y_{ij}) = 2m_{-1}m_0m_1m_2(-1)^n(n-1)!$$
$$\times\int_{R_n}\prod_{\substack{-1\\ i\neq 1}}^{n} d\alpha_i[D_n(\alpha_{-1}, \alpha_0, \alpha_2, \ldots, \alpha_n)]^{-n}, \tag{7}$$

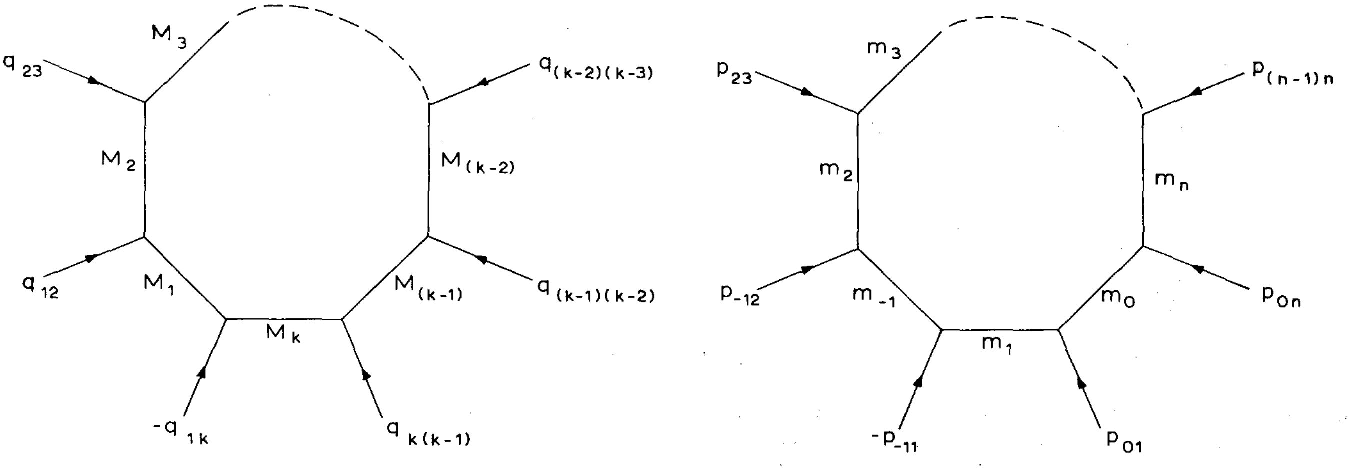
$$D_n(\alpha_{-1}, \alpha_0, \alpha_2, \ldots, \alpha_n) = \sum_{\substack{-1\\ i\neq 1}}^{n} m_i^2\alpha_i^2 + m_1^2(1-\sum_{\substack{-1\\ i\neq 1}}^{n}\alpha_i)^2$$
$$+ \sum_{\substack{i<j\\ i,j\neq 1}}^{n} 2m_im_jy_{ij}\alpha_i\alpha_j + \sum_{\substack{-1\\ j\neq 1}}^{n} 2m_1m_jy_{1j}\alpha_j(1-\sum_{\substack{-1\\ i\neq 1}}^{n}\alpha_i) \tag{8}$$

and

$$R_n = \{(\alpha_{-1}, \alpha_0, \alpha_2, \ldots, \alpha_n) |$$
$$\alpha_{-1} \geqslant 0, \alpha_0 \geqslant 0, \alpha_2 \geqslant 0, \ldots, \alpha_n \geqslant 0,$$
$$\sum_{\substack{-1\\ i\neq 1}}^{n}\alpha_i = 1\}. \tag{9}$$

The constant g is the product of the coupling constants acting at the vertices in Fig. 2.

We begin by generalizing the transformation used in Sec. 2 of P. The change of variables is



FIGS. 1 and 2. Single loop diagrams for the multiparticle production process in which i initial particles produce $f = (k-i) = (n+2-i)$ final particles.

$$\nu = \alpha_0^{-1}(\alpha_{-1} + \alpha_0),$$

$$\lambda_1 = (\alpha_{-1} + \alpha_0)^{-1}\left(1 - \sum_{\substack{-1 \\ i\neq 1}}^{n}\alpha_i\right),$$

$$\lambda_i = (\alpha_{-1} + \alpha_0)^{-1}\alpha_i \quad (2 \leq i \leq n), \tag{10}$$

and it is shown by induction in Appendix A that the inverse is

$$\alpha_{-1} = \nu^{-1}(\nu - 1)\left(1 + \sum_{i=1}^{n}\lambda_i\right)^{-1},$$

$$\alpha_0 = \nu^{-1}\left(1 + \sum_{i=1}^{n}\lambda_i\right)^{-1},$$

$$\alpha_i = \lambda_i\left(1 + \sum_{i=1}^{n}\lambda_i\right)^{-1} \quad (2 \leq i \leq n). \tag{11}$$

Further, from Appendix A, we find that the Jacobian of the transformation is

$$\left|\frac{\partial(\alpha_{-1}, \alpha_0, \alpha_2, \ldots, \alpha_n)}{\partial(\nu, \lambda_1, \ldots, \lambda_n)}\right| = \left(1 + \sum_{i=1}^{n}\lambda_i\right)^{-(n+2)}\nu^{-2} \tag{12}$$

and

$$D_n(\alpha_{-1}, \alpha_0, \alpha_2, \ldots, \alpha_n) \equiv \Delta_n(\nu, \lambda_1, \ldots, \lambda_n)$$
$$= \nu^{-1}\left(1 + \sum_{i=1}^{n}\lambda_i\right)^{-2}[(\nu - 1)\phi(\lambda_1, \ldots, \lambda_n) + \psi(\lambda_1, \ldots, \lambda_n)$$
$$- \nu^{-1}(\nu - 1)v(x_1)]. \tag{13}$$

Here

$$\phi(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} m_i^2\lambda_i^2 + \sum_{i<j}^{n} 2m_i m_j y_{ij}\lambda_i\lambda_j$$
$$+ \sum_{i=1}^{n} 2m_{-1}m_i y_{-1i}\lambda_i + m_{-1}^2, \tag{14}$$

$$\psi(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} m_i^2\lambda_i^2 + \sum_{i<j}^{n} 2m_i m_j y_{ij}\lambda_i\lambda_j$$
$$+ \sum_{i=1}^{n} 2m_0 m_i y_{0i}\lambda_i + m_0^2, \tag{15}$$

and

$$v(x_1) = 2m_{-1}m_0 x_1^2 + m_{-1}^2 + m_0^2 \tag{16}$$

with x_1 given in Eq. (5). Now $I_{n+2}(y_{ij})$ takes the form

$$I_{n+2}(y_{ij}) = 2m_{-1}m_0m_1m_2(-1)^n(n-1)!\int_0^\infty \prod_{i=1}^{n} d\lambda_i \int_1^\infty d\nu$$
$$\times \nu^{-2}\left(1 + \sum_{i=1}^{n}\lambda_i\right)^{n-2}$$
$$\times\{\nu^{-1}[(\nu - 1)\phi(\lambda_1, \ldots, \lambda_n) + \psi(\lambda_1, \ldots, \lambda_n)$$
$$- \nu^{-1}(\nu - 1)v(x_1)]\}^{-n} \tag{17a}$$
$$= 2m_{-1}m_0m_1m_2\int_0^\infty \prod_{i=1}^{n} d\lambda_i\left(\prod_{i=3}^{n}\lambda_i\right)^{-1}\int_1^\infty d\nu$$
$$\times\left(\prod_{i=3}^{n}\frac{\partial}{\partial m_i^2}\right)[(\nu - 1)\phi(\lambda_1, \ldots, \lambda_n) + \psi(\lambda_1, \ldots, \lambda_n)$$
$$- \nu^{-1}(\nu - 1)v(x_1)]^{-2} \tag{17b}$$

for $n \geq 3$. That the expression for $I_{n+2}(y_{ij})$ in Eq. (17b) is, for $n \geq 3$, equivalent to that in Eq. (17a) can be seen by using Eq. (4) in Eqs. (14) and (15). Equation (17a) of course holds for all $n \geq 2$, but as the case $n = 2$ was treated in detail in I, we shall concentrate on the case $n \geq 3$. Note also that the form of Eq. (17b) is similar to Eq. (P-5); in fact the structures of many of the subsequent equations will be similar to those in P. [Equations from P (resp. I) are denoted by placing a P- (resp. I-) in front of the equation number.]

To simplify the proof of the spectral representations, we restrict the y_{ij} defined in Eq. (4) to

$$y_{ij} > 0 \quad (-1 \leq i < j \leq n). \tag{18}$$

Equation (18) ensures that $\phi(\lambda_1, \ldots, \lambda_n) > 0$, $\psi(\lambda_1, \ldots, \lambda_n) > 0$ for $\lambda_1 \geq 0, \ldots, \lambda_n \geq 0$; in fact the term in square brackets in Eq. (17b) is always positive and $I_{n+2}(y_{ij})$ is well defined. That the conditions in Eq. (18) can, for sufficiently large internal masses, be satisfied for finite physical values of the kinematic invariants and external masses is shown in Sec. 8. Equation (18) in fact gives sufficient conditions for the external masses to be stable. In general, however, for a physical single-loop amplitude corresponding to i initial particles producing $f = (k - i)$ final particles, some of the kinematic invariants defined in Eq. (3) can be positive and unbounded. Thus, for finite internal masses it is possible for some of the kinematic invariants to have physical values such that the corresponding y_{ij} are negative. However, it can be seen from Eqs. (17b), (14), (15), and (16) that when some of the y_{ij} are negative, a spectral representation for $I_{n+2}(y_{ij})$ cannot in general be proved by using real analysis only. To obtain a representation in such cases, for physical values of the invariants, one might then start with the double spectral representation derived in Sec. 8 [Eq. (67)] and attempt to do an analytic continuation in the required kinematic invariants using, for example, a generalization of the method of Ref. 25 (referred to as II). Such a procedure might be feasible for the pentagon diagram amplitude, at least for some specific processes,[26] but for a general kth order single-loop amplitude this method does not seem practical for obtaining a representation for all possible configurations involving physical invariants. Of course, some continuation, namely in x_1 and x_2, can easily be carried out since these variables appear only in the Cauchy kernels in Eq. (67). Further, as discussed in Sec. 8, Eq. (67) is expected to be valid under more general conditions [on the other variables defined in Eq. (4) as well as on x_1 and x_2] than those given in Eq. (18).

The argument leading to Eqs. (I-19) and (I-20) can now be used to show that

$$I_{n+2}(y_{ij}) = \int_0^\infty\left(\prod_{\substack{3\\ j\neq i}}^{n} d\lambda_j\right)^{-1}\left(\prod_{\substack{3\\ j\neq i}}^{n}\frac{\partial}{\partial m_j^2}\right)\lim_{\epsilon_i\downarrow 0}\frac{\partial}{\partial m_i^2}\int_{\epsilon_i}^\infty\frac{d\lambda_i}{\lambda_i}$$
$$\times J_{n+2}(y_{ij}, \lambda_3, \ldots, \lambda_n) \tag{19a}$$
$$= \prod_{i=3}^{n}\left(\lim_{\epsilon_i\downarrow 0}\frac{\partial}{\partial m_i^2}\int_{\epsilon_i}^\infty\frac{d\lambda_i}{\lambda_i}\right)J_{n+2}(y_{ij}, \lambda_3, \ldots, \lambda_n), \tag{19b}$$

where

$$J_{n+2}(y_{ij}, \lambda_3, \ldots, \lambda_n)$$
$$= \int_0^\infty\frac{d\lambda_2}{\lambda_2}\lim_{\epsilon_1\downarrow 0}\frac{\partial}{\partial x_2}\int_{\epsilon_1}^\infty\frac{d\lambda_1}{\lambda_1}\int_{h(\lambda_1,\ldots,\lambda_n)}^\infty$$
$$\frac{d\xi}{(\xi - x_1)[U(\xi, \lambda_1, \ldots, \lambda_n)]^{1/2}}. \tag{20}$$

In Eq. (20)

$$U(\xi, \lambda_1, \ldots, \lambda_n) = [\xi - h(\lambda_1, \ldots, \lambda_n)][\xi - k(\lambda_1, \ldots, \lambda_n)], \tag{21}$$

$$\begin{matrix} h(\lambda_1, \ldots, \lambda_n) \\ k(\lambda_1, \ldots, \lambda_n) \end{matrix} = (2m_{-1}m_0)^{-1}\{[\sqrt{\phi(_1, \ldots, \lambda_n)} \pm \sqrt{\psi(\lambda_1, \ldots, \lambda_n)}\,]^2 - m_{-1}^2 - m_0^2\} \tag{22}$$

and $\phi(\lambda_1, \ldots, \lambda_n)$, $\psi(\lambda_1. \ldots, \lambda_n)$ are given in Eqs. (14) and (15). The required Cauchy kernel now appears in Eqs. (19) and (20) and to obtain a dispersion relation the orders of integration must be reversed so that the lower limit of the ξ integration becomes a constant.

3. STUDY OF h (λ)

To reverse the order of integration in Eqs. (19) and (20), we need to examine the function $h(\lambda_1, \ldots, \lambda_n)$ for $\lambda_1 \geqslant 0, \ldots, \lambda_n \geqslant 0$. For convenience we introduce the following notation:

$$(\boldsymbol{\lambda}) \equiv (\lambda_1, \ldots, \lambda_n), \tag{23}$$

$$({}_{i\ldots l}\boldsymbol{\lambda}) \equiv (\lambda_1, \ldots, \lambda_{i-1}, 0, \lambda_{i+1}, \ldots, \lambda_{l-1}, 0, \lambda_{l+1}, \ldots, \lambda_n), \tag{24}$$

and

$${}_{i\ldots l}\boldsymbol{\lambda} \geqslant 0 \equiv \{\lambda_1 \geqslant 0, \ldots, \lambda_{i-1} \geqslant 0, \lambda_{i+1} \geqslant 0, \ldots, \lambda_{l-1} \geqslant 0, \lambda_{l+1} \geqslant 0, \ldots, \lambda_n \geqslant 0\} \tag{25}$$

where $i, \ldots, l \in \{1, \ldots, n\}$. Then from Eqs. (14) and (15),

$$\phi(\boldsymbol{\lambda}) = p_i\lambda_i^2 + 2q_i({}_i\boldsymbol{\lambda})\lambda_i + r_i({}_i\boldsymbol{\lambda}) \quad (>0),$$

$$\psi(\boldsymbol{\lambda}) = p_i\lambda_i^2 + 2q_i'({}_i\boldsymbol{\lambda})\lambda_i + r_i'({}_i\boldsymbol{\lambda}) \quad (>0), \tag{26}$$

where $i \in \{1, \ldots, n\}$ is fixed and

$$p_i = m_i^2,$$

$$q_i(\boldsymbol{\lambda}) = m_i\left(\sum_{j=1}^{n} m_j y_{ij}\lambda_j + m_{-1}y_{-1i}\right) (>0),$$

$$q_i'(\boldsymbol{\lambda}) = m_i\left(\sum_{j=1}^{n} m_j y_{ij}\lambda_j + m_0 y_{0i}\right) (>0),$$

$$r_i(\boldsymbol{\lambda}) \equiv \phi(\boldsymbol{\lambda}), \quad r_i'(\boldsymbol{\lambda}) \equiv \psi(\boldsymbol{\lambda}). \tag{27}$$

We have chosen to define $q_i(\boldsymbol{\lambda})$ etc. although only $q_i({}_i\boldsymbol{\lambda})$ etc. are needed in Eq. (26). The functions $r_i({}_i\boldsymbol{\lambda})$ and $r_i'({}_i\boldsymbol{\lambda})$ are determined recursively from Eq. (26) by putting λ_i equal to zero and using in addition Eq. (27) and the fact that $r_i(0) = m_{-1}^2$, $r_i'(0) = m_0^2$.

The argument of Sec. 4 of I (or of VF) then shows that for fixed ${}_i\boldsymbol{\lambda} \geqslant 0$, $h(\boldsymbol{\lambda})$ increases strictly from $h({}_i\boldsymbol{\lambda})$ to $+\infty$ as λ_i increases from 0 to $+\infty$, whenever $h_{\lambda_i}({}_i\boldsymbol{\lambda}) \geqslant 0$. Now

$$h_{\lambda_i}({}_i\boldsymbol{\lambda}) = (m_{-1}m_0)^{-1}[\sqrt{r_i({}_i\boldsymbol{\lambda})} + \sqrt{r_i'({}_i\boldsymbol{\lambda})}]l_i({}_i\boldsymbol{\lambda}), \tag{28}$$

where

$$l_i({}_i\boldsymbol{\lambda}) = \frac{q_i({}_i\boldsymbol{\lambda})}{\sqrt{r_i({}_i\boldsymbol{\lambda})}} + \frac{q_i'({}_i\boldsymbol{\lambda})}{\sqrt{r_i'({}_i\boldsymbol{\lambda})}} \quad (>0), \tag{29}$$

which is positive whenever Eq. (18) holds. Thus we have established that for fixed $i \in \{1, \ldots, n\}$ and fixed ${}_i\boldsymbol{\lambda} \geqslant 0$, $h(\boldsymbol{\lambda})$ increases strictly from $h({}_i\boldsymbol{\lambda})$ to $+\infty$ as λ_i increases from 0 to $+\infty$. Next we shall find the inverse of $\xi = h(\boldsymbol{\lambda})$ for fixed ${}_i\boldsymbol{\lambda} \geqslant 0$.

4. SOLUTIONS OF U (ξ, λ) = 0

In this section we study the behavior of the zeros of $U(\xi, \boldsymbol{\lambda})$ for fixed $\xi \geqslant h({}_i\boldsymbol{\lambda})$ and for fixed ${}_i\boldsymbol{\lambda} \geqslant 0$. From Eqs. (21), (22), (26), and (27) we have, for fixed $i \in \{1, \ldots, n\}$,

$$4m_{-1}^2m_0^2U(\xi, \boldsymbol{\lambda}) = a_i(\xi)\lambda_i^2 + 2b_i(\xi, {}_i\boldsymbol{\lambda})\lambda_i + c_i(\xi, {}_i\boldsymbol{\lambda}), \tag{30}$$

where

$$\begin{aligned} a_i(\xi) &= 4[(q_i(\boldsymbol{\lambda}) - q_i'(\boldsymbol{\lambda}))^2 - p_i v(\xi)] \\ &= 4m_i^2[(m_{-1}y_{-1i} - m_0 y_{0i})^2 - v(\xi)], \\ b_i(\xi, \boldsymbol{\lambda}) &= 2\{[q_i(\boldsymbol{\lambda}) - q_i'(\boldsymbol{\lambda})][r_i(\boldsymbol{\lambda}) - r_i'(\boldsymbol{\lambda})] - [q_i(\boldsymbol{\lambda}) \\ &\quad + q_i'(\boldsymbol{\lambda})]v(\xi)\} \\ &= \sum_{j=1}^{n} \alpha_{ij}(\xi, -y_{ij})\lambda_j + b_i(\xi, \mathbf{0}), \\ \alpha_{ij}(\xi, -y_{ij}) &= 4m_i m_j[(m_{-1}y_{-1i} - m_0 y_{0i})(m_{-1}y_{-1j} - m_0 y_{0j}) \\ &\quad - y_{ij}v(\xi)], \\ b_i(\xi, \mathbf{0}) &= -4m_i m_{-1}m_0[(m_{-1}y_{-1i} + m_0 y_{0i})\xi + m_{-1}y_{0i} \\ &\quad + m_0 y_{-1i}] \end{aligned} \tag{31}$$

and $v(\xi)$ is given in Eq. (16). The functions $c_i(\xi, {}_i\boldsymbol{\lambda})$ are determined recursively from Eq. (30) by putting λ_i equal to zero and using in addition Eq. (31) and the fact that

$$c_i(\xi, \mathbf{0}) = 4m_{-1}^2m_0^2(\xi^2 - 1) \quad (1 \leqslant i \leqslant n). \tag{32}$$

The argument of Sec. 5 of VF (or I) shows that for each $\xi \geqslant h({}_i\boldsymbol{\lambda})$, where ${}_i\boldsymbol{\lambda} \geqslant 0$, the quadratic equation in λ_i,

$$U(\xi, \boldsymbol{\lambda}) = 0,$$

has two real roots given by

$$\lambda_{i\pm}(\xi, {}_i\lambda) = [a_i(\xi)]^{-1}\{-b_i(\xi, {}_i\boldsymbol{\lambda}) \mp [(b_i(\xi, {}_i\boldsymbol{\lambda}))^2 - a_i(\xi)c_i(\xi, {}_i\boldsymbol{\lambda})]^{1/2}\}. \tag{33}$$

From Eqs. (31), (22), (25), and (27) we see that

$$b_i(h({}_i\boldsymbol{\lambda}), {}_i\boldsymbol{\lambda}) = -4[\sqrt{r_i({}_i\boldsymbol{\lambda})} + \sqrt{r_i'({}_i\boldsymbol{\lambda})}]\sqrt{r_i({}_i\boldsymbol{\lambda})}\ \sqrt{r_i'({}_i\boldsymbol{\lambda})}l_i({}_i\boldsymbol{\lambda}), \tag{34}$$

where $l_i({}_i\boldsymbol{\lambda})$ is given in Eq. (29). Since $l_i({}_i\boldsymbol{\lambda}) > 0$ when Eq. (18) holds it follows that $\lambda_{i+}(h({}_i\boldsymbol{\lambda}), {}_i\boldsymbol{\lambda}) = 0 \neq \lambda_{i-}(h({}_i\boldsymbol{\lambda}), {}_i\boldsymbol{\lambda})$ and in fact for fixed ${}_i\boldsymbol{\lambda} \geqslant 0$ $\lambda_{i+}(\xi, {}_i\boldsymbol{\lambda})$ is the inverse of the strictly increasing function $h(\boldsymbol{\lambda})$ on $0 \leqslant \lambda_i < \infty$. Thus $\lambda_{i+}(\xi, {}_i\boldsymbol{\lambda})$ increases strictly from 0 to $+\infty$ as ξ increases from $h({}_i\boldsymbol{\lambda})$ to $+\infty$. We are now in a position to reverse the orders of the ξ and λ_i integrations ($1 \leqslant i \leqslant n$) in Eqs. (19) and (20).

5. REVERSAL OF THE ORDER OF INTEGRATION

We begin this section by reversing the orders of the ξ and λ_1 integrations in Eq. (20). From Secs. 3 and 4 it follows in particular that $h_{\lambda_1}({}_1\boldsymbol{\lambda}) > 0$ for all ${}_1\boldsymbol{\lambda} \geqslant 0$ and that, for fixed ${}_1\boldsymbol{\lambda} \geqslant 0$, $\lambda_{1+}(\xi, {}_1\boldsymbol{\lambda})$ is the inverse of the strictly increasing function $h(\boldsymbol{\lambda})$ on $0 \leqslant \lambda_1 < \infty$. Thus

$$J_{n+2}(y_{ij}, {}_{12}\lambda) = 2m_{-1}m_0 \int_0^\infty \frac{d\lambda_2}{\lambda_2} \lim_{\epsilon_1 \downarrow 0} \frac{\partial}{\partial x_2} \times \int_{h(\epsilon_1, {}_1\lambda)}^\infty \frac{d\xi}{\xi - x_1} \cdot \Lambda(\xi, \epsilon_1, {}_1\lambda) \tag{35}$$

where

$$\Lambda(\xi, \epsilon_1, {}_1\lambda) = \int_{\epsilon_1}^{\lambda_{1+}(\xi, {}_1\lambda)} \frac{d\lambda_1}{\lambda_1[a_1(\xi)\lambda_1^2 + 2b_1(\xi, {}_1\lambda)\lambda_1 + c_1(\xi, {}_1\lambda)]^{1/2}}, \tag{36}$$

and $a_1(\xi)$, $b_1(\xi, {}_1\lambda)$, and $c_1(\xi, {}_1\lambda)$ can be obtained from Eqs. (30)–(32). Note that $h(\epsilon_1, {}_1\lambda)$, $b_1(\xi, {}_1\lambda)$, $\lambda_{1+}(\xi, {}_1\lambda)$, and $\Lambda(\xi, \epsilon_1, {}_1\lambda)$ depend on x_2, where x_2 is given in Eq. (5).

Now, since, for fixed ${}_{12}\lambda \geqslant 0$, $\lambda_{2+}(\xi, {}_{12}\lambda)$ is the inverse of the strictly increasing function $h({}_1\lambda)$ on $0 \leqslant \lambda_2 < \infty$, the argument of Sec. 6 of I (or Sec. 5 of P) can be used to show that

$$J_{n+2}(y_{ij}, {}_{12}\lambda) = \int_{h({}_{12}\lambda)}^\infty \frac{d\xi}{\xi - x_1} X(\xi, {}_{12}\lambda), \tag{37}$$

where

$$X(\xi, {}_{12}\lambda) = 8m_{-1}m_0m_1m_2v(\xi) \int_{f+(\xi, {}_{12}\lambda)}^\infty \frac{d\eta}{(\eta - x_2)[\overline{F}(\xi, \eta, {}_{12}\lambda]^{1/2}} \tag{38}$$

In Eq. (38)

$$\begin{aligned}\overline{F}(\xi, \eta, {}_{12}\lambda) = {} & [\alpha_{12}(\xi, x_2)]^2 c_2(\xi, {}_{12}\lambda) \\ & - 2\alpha_{12}(\xi, x_2) b_1(\xi, {}_{12}\lambda) b_2(\xi, {}_{12}\lambda) \\ & + [b_1(\xi, {}_{12}\lambda)]^2 a_2(\xi) + [b_2(\xi, {}_{12}\lambda)]^2 a_1(\xi) \\ & - a_1(\xi) a_2(\xi) c_2(\xi, {}_{12}\lambda) \\ = {} & 16m_1^2m_2^2[v(\xi)]^2 c_2(\xi, {}_{12}\lambda)[x_2 - f_+(\xi, {}_{12}\lambda)][x_2 - f_-(\xi, {}_{12}\lambda)],\end{aligned} \tag{39}$$

where $f_\pm(\xi, {}_{12}\lambda)$ are defined by

$$\begin{aligned}\alpha_{12}(\xi, f_\pm(\xi, {}_{12}\lambda)) = {} & [c_2(\xi, {}_{12}\lambda)]^{-1}(b_1(\xi, {}_{12}\lambda)b_2(\xi, {}_{12}\lambda) \\ & \pm \{[b_1(\xi, {}_{12}\lambda)]^2 - a_1(\xi)c_1(\xi, {}_{12}\lambda)\}^{1/2}\{[b_2(\xi, {}_{12}\lambda)]^2 \\ & - a_2(\xi)c_2(\xi, {}_{12}\lambda)\}^{1/2})\end{aligned} \tag{40}$$

and the argument of Sec. 5 of I (or of VF) shows that

$$\begin{aligned}&(b_1(\xi, {}_{12}\lambda))^2 - a_1(\xi)c_1(\xi, {}_{12}\lambda) > 0, \\ &(b_2(\xi, {}_{12}\lambda))^2 - a_2(\xi)c_2(\xi, {}_{12}\lambda) > 0\end{aligned} \tag{41}$$

for $\xi \geqslant h({}_{12}\lambda)$. The following points should now be noted. Firstly $\alpha_{12}(\xi, x_2)$ given in Eq. (31) is linear in x_2 so that the explicit expression for $f_\pm(\xi, {}_{12}\lambda)$ can easily be obtained from Eqs. (31) and (40). Secondly, as in Sec. 6 of I (or Sec. 5 of P) it is important to note that

$$c_1(\xi, {}_{12}\lambda) = c_2(\xi, {}_{12}\lambda) > 0 \quad \text{for } \xi > h({}_{12}\lambda)$$

in order to obtain Eqs. (37)–(40). Finally

$$\overline{F}(\xi, x_2, 0) = 64m_{-1}^2m_0^2m_1^2m_2^2[v(\xi)]^2 F(\xi, x_2), \tag{42}$$

where $F(\xi, x_2)$, corresponding to the usual Mandelstam spectral function, is given in Eq. (B1) of Appendix B and $v(\xi)$ is given in Eq. (16).

The second Cauchy kernel now appears in the expression for $I_{n+2}(y_{ij})$ given by Eqs. (38), (37), and (19). To obtain the desired double spectral representation, it remains to reverse the order of the λ_i ($3 \leqslant i \leqslant n$) and ξ integrations and as well the λ_i and η integrations. The interchange of the λ_i ($3 \leqslant i \leqslant n$) and ξ integrations can be carried out by using the results of Secs. 3 and 4. Thus, using the expression for $I_{n+2}(y_{ij})$ in Eq. (19a), we have

$$\begin{aligned}I_{n+2}(y_{ij}) = {} & \int_0^\infty \prod_{\substack{3 \\ j \neq i}}^n d\lambda_j \Bigg(\prod_{\substack{3 \\ j \neq i}}^n \lambda_j\Bigg)^{-1} \Bigg(\prod_{\substack{3 \\ j \neq i}}^n \frac{\partial}{\partial m_j^2}\Bigg) \lim_{\epsilon_i \downarrow 0} \frac{\partial}{\partial m_i^2} \\ & \times \int_{h({}_{12}\lambda)|_{\lambda_i = \epsilon_i}}^\infty \frac{d\xi}{\xi - x_1} \int_{\epsilon_i}^{\lambda_{i+}(\xi, {}_{12i}\lambda)} \frac{d\lambda_i}{\lambda_i} \\ & \times \int_{f_+(\xi, {}_{12}\lambda)}^\infty \frac{d\eta}{\eta - x_2} \frac{8m_{-1}m_0m_1m_2v(\xi)}{[\overline{F}(\xi, \eta, {}_{12}\lambda)]^{1/2}}.\end{aligned} \tag{43}$$

The ξ and the other λ_j ($3 \leqslant j \leqslant n, j \neq i$) integrations can, of course, be reversed in a similar way, but it will be more convenient to interchange the order of the λ_i and η integrations before this is done.

6. STUDY OF $f_+(\xi, {}_{12}\lambda)$

To reverse the order of the λ_i and η integrations in Eq. (43), we need to examine the function $f_+(\xi, {}_{12}\lambda)$ for $0 \leqslant \lambda_i \leqslant \lambda_{i+}(\xi, {}_{12i}\lambda)$ with fixed ${}_{12i}\lambda \geqslant 0$, $\xi \geqslant h({}_{12i}\lambda)$, and $i \in \{3, \ldots, n\}$. First we study the behavior of $f_+(\xi, {}_{12}\lambda)$ as $\lambda_i \uparrow \lambda_{i+}(\xi, {}_{12i}\lambda)$. From Eq. (34) and the fact that $l_1({}_{12}\lambda) > 0$ when Eq. (18) holds it follows that

$$b_1(\xi, {}_{12}\lambda)|_{\lambda_i = \lambda_{i+}(\xi, {}_{12i}\lambda)} < 0. \tag{44}$$

Similarly

$$b_2(\xi, {}_{12}\lambda)|_{\lambda_i = \lambda_{i+}(\xi, {}_{12i}\lambda)} < 0 \tag{45}$$

and hence from Eqs. (30) and (40) and the fact that

$$v(\xi) > 0 \tag{46}$$

for $\xi \geqslant h({}_{12i}\lambda)$ ($\geqslant 1$) it follows that $f_+(\xi, {}_{12}\lambda) \to +\infty$ as $\lambda_i \uparrow \lambda_{i+}(\xi, {}_{12i}\lambda)$.

Next, from Eq. (40) we see that the derivative of $f_+(\xi, {}_{12}\lambda)$ with respect to λ_i is

$$\begin{aligned}f_{+\lambda_i}(\xi, {}_{12}\lambda) = {} & [4m_{-1}m_0v(\xi)]^{-1}[c_2(\xi, {}_{12}\lambda)]^{-2} \\ & \times(-b_1(\xi, {}_{12}\lambda)\{[b_2(\xi, {}_{12}\lambda)]^2 - a_2(\xi)c_2(\xi, {}_{12}\lambda)\}^{1/2} \\ & - b_2(\xi, {}_{12}\lambda)\{b_1(\xi, {}_{12}\lambda)]^2 - a_1(\xi)c_1(\xi, {}_{12}\lambda)\}^{1/2}). \\ & \times L_i(\xi, {}_{12}\lambda),\end{aligned} \tag{47}$$

where

$$L_i(\xi, {}_{12}\lambda) = \frac{Q_i(\xi, {}_{12}\lambda)}{\sqrt{R_i(\xi, {}_{12}\lambda)}} + \frac{Q_i'(\xi, {}_{12}\lambda)}{\sqrt{R_i'(\xi, {}_{12}\lambda)}}, \tag{48}$$

$$\begin{aligned}Q_i(\xi, {}_{12}\lambda) &= \lambda_i(a_ib_1 - \alpha_{1i}b_i) + (b_ib_1 - \alpha_{1i}c_i), \\ Q_i'(\xi, {}_{12}\lambda) &= \lambda_i(a_ib_2 - \alpha_{2i}b_i) + (b_ib_2 - \alpha_{2i}c_i), \\ R_i(\xi, {}_{12}\lambda) &= \lambda_i^2(\alpha_{1i}^2 - a_1a_i) + 2\lambda_i(\alpha_{1i}b_1 - a_1b_i) + b_1^2 - a_1c_1, \\ R_i'(\xi, {}_{12}\lambda) &= \lambda_i^2(\alpha_{2i}^2 - a_2a_i) + 2\lambda_i(\alpha_{2i}b_2 - a_2b_i) + b_2^2 - a_2c_2.\end{aligned} \tag{49}$$

In Eq. (49), a_j has been written for $a_j(\xi)$, c_j for $c_j(\xi, {}_{12i}\lambda)$, b_j for $b_j(\xi, {}_{12i}\lambda)$ ($j = 1, 2, i$) and α_{li} for $\alpha_{li}(\xi, -y_{li})$ ($l = 1, 2$).

The term in square brackets in Eq. (47) is always positive as can be seen as follows. From Eqs. (34), (29), and (18) and the fact that $b_1(\xi, {}_{12i}\lambda)$ is linear in ξ, we have

$$b_1(\xi, {}_{12i}\lambda) < 0 \qquad (50)$$

for all $\xi \geqslant h({}_{12i}\lambda)$, ${}_{12i}\lambda \geqslant 0$. Then from Eqs. (44) and (50) and the fact that $b_1(\xi, {}_{12}\lambda)$ is linear in λ_i it follows that

$$b_1(\xi, {}_{12}\lambda) < 0 \qquad (51)$$

for $0 \leqslant \lambda_i \leqslant \lambda_{i+}(\xi, {}_{12i}\lambda)$ with fixed $\xi \geqslant h({}_{12i}\lambda)$, ${}_{12i}\lambda \geqslant 0$. Similarly

$$b_2(\xi, {}_{12}\lambda) < 0 \qquad (52)$$

for $0 \leqslant \lambda_i \leqslant \lambda_{i+}(\xi, {}_{12i}\lambda)$.

The argument of Sec. 6 of P can now be used to show that, for fixed ${}_{12i}\lambda \geqslant 0$, $\xi \geqslant h({}_{12i}\lambda)$, $f_+(\xi, {}_{12i}\lambda)$ is strictly increasing on $0 \leqslant \lambda_i \leqslant \lambda_{i+}(\xi, {}_{12i}\lambda)$ if and only if $L_i(\xi, {}_{12i}\lambda) > 0$. That $L_i(\xi, {}_{12i}\lambda)$ is always positive for $\xi \geqslant h({}_{12i}\lambda)$, ${}_{12i}\lambda \geqslant 0$ can be seen from Eq. (48) and Eqs. (B7) and (B8) of Appendix B. Next we shall find the inverse of the strictly increasing function $\eta = f_+(\xi, {}_{12}\lambda)$ for fixed $\xi \geqslant h({}_{12i}\lambda)$, ${}_{12i}\lambda \geqslant 0$.

7. SOLUTIONS OF $\bar{F}(\xi, \eta, {}_{12}\lambda) = 0$

To obtain the inverse of $\eta = f_+(\xi, {}_{12}\lambda)$, we need to study the behavior of the zeros $\bar{F}(\xi, \eta, {}_{12}\lambda)$ for fixed $\xi. \eta$ and fixed ${}_{12i}\lambda \geqslant 0$. From Eqs. (29) and (31) we find that

$$F(\xi, \eta, {}_{12}\lambda) = A_i(\xi, \eta)\lambda_i^2 + 2B_i(\xi, \eta, {}_{12i}\lambda)\lambda_i + C_i(\xi, \eta, {}_{12i}\lambda), \qquad (53)$$

where

$$\begin{aligned} A_i(\xi, \eta) &= a_i([\alpha_{12}(\xi, \eta)]^2 - a_1a_2) + \alpha_{1i}^2 a_2 + \alpha_{2i}^2 a_1 \\ &\quad - 2\alpha_{12}(\xi, \eta)\alpha_{1i}\alpha_{2i}, \\ B_i(\xi, \eta, {}_{13i}\lambda) &= b_i([\alpha_{12}(\xi, \eta)]^2 - a_1a_2) + b_1\alpha_{1i}a_2 + b_2\alpha_{2i}a_1 \\ &\quad - \alpha_{12}(\xi, \eta)b_1\alpha_{2i} - \alpha_{12}(\xi, \eta)b_2\alpha_{1i}, \\ C_i(\xi, \eta, {}_{12i}\lambda) &= c_i([\alpha_{12}(\xi, \eta)]^2 - a_1a_2) + b_1^2a_2 + b_2^2a_1 \\ &\quad - 2\alpha_{12}(\xi, \eta)b_1b_2. \end{aligned} \qquad (54)$$

The functions $A_i(\xi, \eta)$ and $B_i(\xi, \eta, {}_{12i}\lambda)$ in fact depend on m_i^2 whereas $C_i(\xi, \eta, {}_{12i}\lambda)$ does not. The abbreviations described after Eq. (49) have again been used.

As in Sec. 7 of P, we find that the discriminant of the quadratic function of λ_i in Eq. (53) is

$$\begin{aligned} &[B_i(\xi, \eta, {}_{12i}\lambda)]^2 - A_i(\xi, \eta)C_i(\xi, \eta, {}_{12i}\lambda) \\ &\quad = \{[\alpha_{12}(\xi, \eta)]^2 - a_1a_2\}\bar{P}_i(\xi, \eta, {}_{12i}\lambda), \end{aligned} \qquad (55)$$

where

$$\begin{aligned} \bar{P}_i(\xi, \eta, {}_{12i}\lambda) &= 16m_1^2m_2^2[v(\xi)]^2(b_i^2 - a_ic_i) \\ &\quad \times[\eta - p_{i+}(\xi, {}_{12i}\lambda)][\eta - p_{i-}(\xi, {}_{12i}\lambda)]. \end{aligned} \qquad (56)$$

The functions $p_{i\pm}(\xi, {}_{12i}\lambda)$ are given by

$$\begin{aligned} \alpha_{12}(\xi, p_{i\pm}(\xi, {}_{12i}\lambda)) &= (b_i^2 - a_ic_i)^{-1}\{b_i(b_1\alpha_{2i} + b_2\alpha_{1i}) - a_ib_1b_2 \\ &\quad - c_i\alpha_{1i}\alpha_{2i} \pm [D_i(\xi, {}_{12i}\lambda)D_i'(\xi, {}_{12i}\lambda)]^{1/2}\}, \end{aligned} \qquad (57)$$

where

$$D_i(\xi, {}_{12i}\lambda) = c_i(\alpha_{1i}^2 - a_1a_i) + b_1^2a_i + b_i^2a_1 - 2\alpha_{1i}b_1b_i,$$

$$D_i'(\xi, {}_{12i}\lambda) = c_i(\alpha_{2i}^2 - a_2a_i) + b_2^2a_i + b_i^2a_2 - 2\alpha_{2i}b_2b_i, \qquad (58)$$

and $\alpha_{12}(\xi, \eta)$ given in Eq. (31) is linear in η.

The discriminant in Eq. (55) is always nonnegative for $\eta \geqslant f_+(\xi, {}_{12i}\lambda)$, $\xi \geqslant h({}_{12i}\lambda)$, ${}_{12i}\lambda \geqslant 0$, since the inverse of $\eta = f_+(\xi, {}_{12}\lambda)$ is real. That it is in fact positive can be seen as follows.

In Sec. 7 of P we showed that

$$\{[\alpha_{12}(\xi, \eta)]^2 - a_1(\xi)a_2(\xi)\} > 0 \qquad (59)$$

for all $\xi \geqslant h(0) = 1$, $\eta \geqslant f_+(\xi, 0)$. Since for ${}_{12i}\lambda \geqslant 0$, $h({}_{12i}\lambda) \geqslant 1$ and for $\xi \geqslant h({}_{12i}\lambda)$, $f_+(\xi, {}_{12i}\lambda) \geqslant f_+(\xi, 0)$ Eq. (59) holds in particular for $\xi \geqslant h({}_{12i}\lambda)$, $\eta \geqslant f_+(\xi, {}_{12i}\lambda)$. Further, it is shown in Eq. (B3) of Appendix B that $\bar{P}(\xi, \eta, {}_{12i}\lambda)$ is positive for $\xi \geqslant h({}_{12i}\lambda)$, $\eta \geqslant f_+(\xi, {}_{12i}\lambda)$.

The two real solutions of

$$\bar{F}(\xi, \eta, {}_{12}\lambda) = 0 \qquad (60)$$

are

$$\begin{aligned} &\lambda_{i\substack{a\\b}}(\xi, \eta, {}_{12i}\lambda) \\ &\quad = [A_i(\xi, \eta)]^{-1}(-B_i(\xi, \eta, {}_{12i}\lambda) \mp \{[B_i(\xi, \eta, {}_{12i}\lambda)]^2 \\ &\qquad - A_i(\xi, \eta)C_i(\xi, \eta, {}_{12i}\lambda)\}^{1/2}). \end{aligned} \qquad (61)$$

From Eqs. (54), (48), (B7), (B8), (51), and (52) we see that

$$\begin{aligned} &B_i(\xi, f_+(\xi, {}_{12i}\lambda), {}_{12i}\lambda) \\ &= c_i^{-2}[b_1(R_i'(\xi, {}_{12i}\lambda))^{-1/2} + b_2(R_i(\xi, {}_{12i}\lambda))^{-1/2}]L_i(\xi, {}_{12i}\lambda) < 0 \end{aligned} \qquad (62)$$

so that

$$\lambda_{ia}(\xi, f_+(\xi, {}_{12i}\lambda), {}_{12i}\lambda) = 0 \neq \lambda_{ib}(\xi, f_+(\xi, {}_{12i}\lambda), {}_{12i}\lambda). \qquad (63)$$

Also as $\eta \to +\infty$

$$\lambda_{i\substack{a\\b}}(\xi, \eta, {}_{12i}\lambda) \sim \lambda_{i\pm}(\xi, {}_{12i}\lambda),$$

where $\lambda_{i\pm}(\xi, {}_{12i}\lambda)$ are given in Eq. (33). Thus $\lambda_{ia}(\xi, \eta, {}_{12i}\lambda)$ is the inverse of the strictly increasing function $f_+(\xi, {}_{12}\lambda)$ on $0 \leqslant \lambda_i \leqslant {}_{i+}(\xi, {}_{12i}\lambda)$, where ξ is fixed such that $\xi \geqslant h({}_{12i}\lambda)$ and ${}_{12i}\lambda \geqslant 0$. The function $\lambda_{ia}(\xi, \eta, {}_{12i}\lambda)$ increases strictly from 0 to $\lambda_{i+}(\xi, {}_{12i}\lambda)$ as η increases from $f_+(\xi, {}_{12i}\lambda)$ to $+\infty$.

8. DOUBLE SPECTRAL REPRESENTATION OF THE k^{th} ORDER SINGLE LOOP AMPLITUDE

The order of the λ_i and η integrations in Eq. (43) can now be interchanged, and we find, on taking $i = 3$, that

$$\begin{aligned} I_{n+2}(y_{ij}) &= \int_0^\infty \prod_{j=4}^n d\lambda_j \left(\prod_{j=4}^n \lambda_j\right)^{-1} \left(\prod_{j=4}^n \frac{\partial}{\partial m_j^2}\right) \lim_{\epsilon_3 \downarrow 0} \frac{\partial}{\partial m_3^2} \\ &\quad \times \int_{h({}_{12}\lambda)|_{\lambda_3=\epsilon_3}}^\infty \frac{d\xi}{\xi - x_1} \int_{f_+(\xi, {}_{12}\lambda)|_{\lambda_3=\epsilon_3}}^\infty \frac{d\eta}{\eta - x_2} \\ &\quad \times \int_{\epsilon_3}^{\lambda_{3a}(\xi, \eta, {}_{123}\lambda)} \frac{d\lambda_3}{\lambda_3} \frac{8m_{-1}m_0m_1m_2v(\xi)}{[\bar{F}(\xi, \eta, {}_{12}\lambda)]^{1/2}}. \end{aligned} \qquad (64)$$

The results of Appendix C can be used to interchange the order of $\lim_{\epsilon_3 \downarrow 0} \partial/\partial m_3^2$ and integration with respect to ξ and η. Defining the operators

$$O_j \equiv \lim_{\epsilon_j \downarrow 0} \frac{\partial}{\partial m_j^2} \int_{\epsilon_j}^{\lambda_{ja}(\xi,\eta,{}_{12\cdots j}\lambda)} \frac{d\lambda_j}{\lambda_j} \quad (3 \leqslant j \leqslant n), \tag{65}$$

we find that

$$I_{n+2}(y_{ij}) = \int_0^\infty \prod_{j=4}^n d\lambda_j \left(\prod_{j=4}^n \lambda_j\right)^{-1} \left(\prod_{j=4}^n \frac{\partial}{\partial m_j^2}\right) \times \int_{h({}_{123}\lambda)}^\infty \frac{d\xi}{\xi - x_1} \int_{f_+(\xi,{}_{123}\lambda)}^\infty \frac{d\eta}{\eta - x_2} O_3 \times \frac{8m_{-1}m_0m_1m_2v(\xi)}{[\overline{F}(\xi, \eta, {}_{12}\lambda)]^{1/2}} . \tag{66}$$

Note that the limit $\epsilon_3 \downarrow 0$ is now inside the ξ and η integrations and in fact $O_3[\overline{F}(\xi, \eta, {}_{12}\lambda)]^{-1/2}$ can be evaluated as in Eq. (C1).

With the expression for $I_{n+2}(y_{ij})$ given in Eqs. (19b), (37), and (38), we can use the results of Secs. 3 and 4 to reverse the order of the λ_4 and ξ integrations and the results of Secs. 6 and 7 to interchange the order of the λ_4 and η integrations. The order of $\lim_{\epsilon_4 \downarrow 0} \partial/\partial m_4^2$ and the integration with respect to ξ and η can then be interchanged by using the results of Appendix C. Repeating the process, we find that

$$I_{n+2}(y_{ij}) = \int_1^\infty \frac{d\xi}{\xi - x_1} \int_{f_+(\xi)}^\infty \frac{d\eta}{\eta - x_2} O_n O_{n-1} \cdots O_3 \times \frac{8m_{-1}m_0m_1m_2}{[\overline{F}(\xi, \eta, {}_{12}\lambda)]^{1/2}} \tag{67}$$

giving the required double spectral representation with Mandelstam boundary for the single loop amplitude of order $k = n + 2$ (>4). In Eq. (67) $f_+(\xi) \equiv f_+(\xi, \mathbf{0})$ is given in Eq. (B2) and $h(\mathbf{0}) = 1$.

From Eq. (42) it can be seen that with all the O_j operators missing $(3 \leqslant j \leqslant n)$, ${}_{12}\lambda \to \mathbf{0}$ and $n = 2$, Eq. (67) is just the double spectral representation obtained for the box diagram amplitude in Eq. (I-97). Also with only O_3 appearing in Eq. (67), ${}_{12}\lambda \to \lambda_3$ and $n = 3$, we see from Eq. (C1) that Eq. (67) is just the double spectral representation for the pentagon diagram amplitude given in Eq. (P-64) and hence in Eq. (P-65). While it is much more tedious to evaluate some of the higher order spectral functions from Eq. (67), such calculations would provide interesting checks on the Cutkosky rules which, to the best of my knowledge, have only been checked for the lowest order amplitudes (see, for example, also Chap. 4, Sec. 3 of Ref. 4).

It remains to show that the conditions in Eq. (18) under which the double spectral representation for the kth order single-loop amplitude has been proved, can, for sufficiently large internal masses, be satisfied for finite physical values of the kinematic invariants and external masses. Further we shall discuss how one can obtain a representation for the amplitude when the conditions in Eq. (18) are, at least to some extent, relaxed. We begin by considering the box diagram amplitude. For a particular channel reaction, the equations determining the region in which the kinematic invariants take on physical values (given, for example, in Ref. 8) depend only on the kinematic invariants and external masses. Thus, for finite physical values of the kinetic invariants and external masses, Eq. (18) can be satisfied provided the internal masses are sufficiently large. To obtain a representation for more general values of the y_{ij}, one can do an analytic continuation in these variables as was done in II. While, as found there, it is rather laborious to obtain a representation for almost all real kinematic invariants and for all possible mass configurations involving stable external particles, the continuation in the kinematic invariants can readily be carried out when the mass variables satisfy Eq. (18). Then, the representation from which the continuation is started is the double spectral representation which contains x_1 and x_2 only in the Cauchy kernels. Thus for the x_1 channel reaction, which in Fig. 2 corresponds to $n = 2$, p_{-11} (rather than $-p_{-11}$) and p_{01} being incoming and p_{-12} and p_{02} (rather than their negatives) being outgoing 4-momenta, the amplitude has, for physical invariants, the representation given in Eq. (I-97) and in Eq. (67) with $x_1 \to x_1 + i \circ O$ [and as described earlier, with the O_j operators missing $(3 \leqslant j \leqslant n)$ and ${}_{12}\lambda \to 0$]. In fact, as shown in Secs. 6 and 8 of I, Eq. (I-97) holds under slightly more general conditions on the mass variables than those given in Eq. (18). In a similar way, for a general kth order single-loop amplitude one would expect Eq. (67) to be valid under more general conditions than those given in Eq. (18); that continuation in x_1 and x_2 can easily be carried out is of course obvious.

For the pentagon diagram amplitude, the equations which, for a particular channel reaction, define the region in which the kinematic invariants take on physical values again depend only on the kinematic invariants and external masses (see, for example, Section 4.3 of Ref. 4). Thus Eq. (18) can be satisfied for finite physical values of the kinematic invariants and external masses provided the internal masses are sufficiently large. It should, however, be noted that, for pentagon diagram amplitudes associated with most physically interesting production reactions involving hadrons, for example $\pi N \to \pi\pi N$, the lowest mass intermediate particles which can be exchanged are such that complex singularities appear on the physical sheet,[26] even for the smallest possible physical values of the invariants, causing a breakdown of the double (and even single) dispersion relations in x_1 and x_2. Hence one would not expect dispersion relations, over real contours, in these variables to be valid for the total production amplitudes. In fact, for the reaction $\pi N \to \pi\pi N$, complex singularities are also produced by lower order contracted diagrams.[27] To obtain a representation for the pentagon diagram amplitudes when the internal masses are small, one might attempt to generalize the method of continuation used in II. On the basis of the work of Cook and Tarski,[26] it seems that at least a numerical study of the motion of the singularities for specific processes is feasible.

Finally, for the kth order single-loop amplitude where $k \geqslant 6$, we mentioned in Sec. 2 that the kinematic invariants defined in Eqs. (2) or (3) are not independent but satisfy algebraic constraints.[22] These constraints, however, involve only the kinematic invariants and the external masses. Further the equations which, for a particular channel reaction, define the region in which the kinematic invariants take on physical values again depend only on the kinematic invariants and the external masses.[4] Thus, for finite physical values of the kinematic invariants and external masses, the conditions in Eq. (18), under which the double spectral representation was proved, can be satisfied provided the internal masses are sufficiently large. In fact, as mentioned earlier, Eq. (67) is expected to hold under slightly more general conditions than those given in Eq. (18). The double (and even single) dispersion relations in x_1 and x_2 will of course break down for sufficiently small internal masses. In such cases the method of analytic continuation unfortunately seems of little use for finding a representation of the amplitude, simply because of the increased number of singularities and the more complicated nature of the spectral function in Eq. (67).

ACKNOWLEDGMENTS

It is my pleasure to thank Dr. W.S. Woolcock for discussing several aspects of this work with me. I also wish to thank F.O.M. (Foundation for Fundamental Research on Matter) for the receipt of a grant.

APPENDIX A

In this appendix we outline the method of induction used to obtain Eqs. (10)—(17). It was shown in Sec. 2 of P that these equations hold for the case $n=3$ (where $n=k-2$ and k is the number of vertices of the single loop amplitude). We suppose that Eqs. (6)—(17), with the replacement $n\to l$, hold for all $3\leqslant l\leqslant n-1$ and show that they are then valid as well for $l=n$. The steps in the proof are as follows.

(1) In Eqs. (7)—(9) make the change of variables $\zeta=(1-\alpha_n)^{-1}$ and then $x_i=\zeta\alpha_i$ $(i\neq 1,\ -1\leqslant i\leqslant n-1)$. The Jacobian of the transformation is $\zeta^{-(n+1)}$ and in terms of the new variables

$$\begin{aligned}D_n(\alpha_{-1},\alpha_0,\alpha_2,\ldots,\alpha_n)\equiv{}& m_n^2(\zeta-1)^2\zeta^{-2}\\&+\sum_{\substack{-1\\i\neq1}}^{n-1}2m_im_ny_{in}(\zeta-1)\zeta^{-2}x_i\\&+2m_1m_ny_{1n}\Big(1-\sum_{\substack{-1\\i\neq1}}^{n-1}x_i\Big)(\zeta-1)\zeta^{-2}\cdot\\&+\zeta^{-2}D_{n-1}(x_{-1},x_0,x_2,\ldots x_{n-1}).\end{aligned}\tag{A1}$$

(2) Make the change of variables given in Eqs. (10) and (11) with the replacements $n\to n-1$, $\alpha_i\to x_i$. With these replacements, the Jacobian of the transformation is, by assumption, given in Eq. (12) and using Eq. (13), again with the above replacements, we find that

$$\begin{aligned}&D_n(\alpha_{-1},\alpha_0,\alpha_2,\ldots,\alpha_n)\\&\equiv\zeta^{-2}\Bigg[m_n^2(\zeta-1)^2+\sum_{i=2}^{n=1}2m_im_ny_{in}\lambda_i(\zeta-1)\Big(1+\sum_{j=1}^{n=1}\lambda_j\Big)^{-1}\\&\quad+2m_0m_ny_{0n}\nu^{-1}(\zeta-1)\Big(1+\sum_{j=1}^{n-1}\lambda_j\Big)^{-1}\\&\quad+2m_{-1}m_ny_{-1n}\nu^{-1}(\nu-1)(\zeta-1)\Big(1+\sum_{j=1}^{n-1}\lambda_j\Big)^{-1}\\&\quad+2m_1m_ny_{1n}\lambda_1(\zeta-1)\Big(1+\sum_{j=1}^{n-1}\lambda_j\Big)^{-1}\\&\quad+\nu^{-1}\Big(1+\sum_{j=1}^{n-1}\lambda_j\Big)^{-2}[(\nu-1)\phi(\lambda_1,\ldots,\lambda_{n-1},0)\\&\quad+\psi(\lambda_1,\ldots,\lambda_{n-1},0)\\&\quad-\nu^{-1}(\nu-1)v(x_1)]\Bigg].\end{aligned}\tag{A2}$$

(3) Make the change of variable

$$\lambda_n=(\zeta-1)\Big(1+\sum_{j=1}^{n-1}\lambda_j\Big)$$

with the inverse

$$\zeta=\Big(1+\sum_{j=1}^{n}\lambda_j\Big)\Big(1+\sum_{j=1}^{n-1}\lambda_j\Big)^{-1}.$$

The Jacobian of the transformation is $(1+\sum_{j=1}^{n-1}\lambda_j)^{-1}$ and we find that $D_n(\alpha_{-1},\alpha_0,\alpha_2,\ldots,\alpha_n)\equiv\Delta_n(\nu,\lambda_1,\ldots,\lambda_n)$ as given in Eqs. (13)—(16).

Compounding the transformations and Jacobians in steps (1)—(3) we find that the resultant transformation is just that given in Eq. (10) with the inverse as in Eq. (11). Further the resultant Jacobian is as given in Eq. (12). It is then readily seen that the new region of integration and expression for $I_{n+2}(y_{ij})$ are as given in Eq. (17).

APPENDIX B

We collect here a number of results involving the various functions needed in the main body of the paper. It is assumed throughout that Eq. (18) holds. From Eqs. (42) and (39)

$$F(\xi,x_2)=(\xi^2-1)[x_2-f_+(\xi)][x_2-f_-(\xi)],\tag{B1}$$

where from Eqs. (40) and (31)

$$\begin{aligned}f_\pm(\xi)\equiv f_\pm(\xi,0)={}&(\xi^2-1)^{-1}[(\xi-1)(y_{-11}y_{02}+y_{01}y_{-12})\\&+(y_{-11}+y_{01})(y_{-12}+y_{02})\\&\pm(\xi^2+2y_{-11}y_{01}\xi+y_{-11}^2+y_{01}^2-1)^{1/2}\\&\times(\xi^2+2y_{-12}y_{02}\xi+y_{-12}^2+y_{02}^2-1)^{1/2}].\end{aligned}\tag{B2}$$

The above functions, with a relabelling of variables, were also defined in Eqs. (I-11) and (I-12) and their properties were discussed in detail in Sec. 8 of I and in Sec. 4 of II.

Next we show that, for fixed $i\in\{3,\ldots,n\}$,

$$\overline{P}_i(\xi,\eta,{}_{12i}\lambda)>0\tag{B3}$$

for all $\xi\geqslant h({}_{12i}\lambda)$, $\eta\geqslant f_+(\xi,{}_{12i}\lambda)$, where $\overline{P}_i(\xi,\eta,{}_{12i}\lambda)$ is defined in Eqs. (56), (57), and (58) and the abbreviations described after Eq. (49) have again been used. In the same way as Eq. (41) was established, we find that

$$b_i^2-a_ic_i>0\tag{B4}$$

for $\xi\geqslant h({}_{12i}\lambda)$ and hence Eq. (B3) will hold if $p_{i\pm}(\xi,{}_{12i}\lambda)$, given in Eq. (57), are either complex conjugates or if

$$p_{i-}(\xi,{}_{12i}\lambda)\leqslant p_{i+}(\xi,{}_{12i}\lambda)<f_+(\xi,{}_{12i}\lambda).\tag{B5}$$

That both alternatives are in fact possible can be seen from Appendix A of P. Thus we have the following cases to consider.

(i) $D_i(\xi,{}_{12i}\lambda)D_i'(\xi,{}_{12i}\lambda)<0$. Then $p_{i\pm}(\xi,{}_{12i}\lambda)$ are complex conjugates and Eq. (B3) holds.

(ii) $D_i(\xi, {}_{12i}\lambda) \geqslant 0$, $D_i'(\xi, {}_{12i}\lambda) \geqslant 0$. From Eqs. (40), (57), (58), (49) and the fact that $c_1 = c_2 = c_i$, we have

$$\begin{aligned}\alpha_{12}(\xi, p_{i\pm}(\xi, {}_{12i}\lambda)) - \alpha_{12}(\xi, f_+(\xi, {}_{12i}\lambda))\\ = (b_i^2 - a_i c_i)^{-1} c_2^{-1}[- Q_i(\xi, {}_{12i}\lambda) Q_i'(\xi, {}_{12i}\lambda)\\ \pm (\{[Q_i(\xi, {}_{12i}\lambda)]^2 - (b_i^2 - a_i c_i)(b_1^2 - a_1 c_1)\}\\ \times \{[Q_i'(\xi, {}_{12i}\lambda)]^2 - (b_i^2 - a_i c_i)(b_2^2 - a_2 c_2)\})^{1/2}\\ - (b_i^2 - a_i c_i)(b_1^2 - a_1 c_1)^{1/2}(b_2^2 - a_2 c_2)^{1/2}]. \qquad \text{(B6)}\end{aligned}$$

Now from Eqs. (49), (31), (30), (21) and (27) it can be shown that

$$\begin{aligned}Q_i(\xi, {}_{12i}\lambda)\\ = 4v(\xi)\{y_{1i}c_1 + 2(q_1 q_i' + q_1' q_i)[v(\xi) - r_1 - r_1'] + 4q_i' q_1' r_1\\ + 4q_i q_1 r_1'\} > 0 \qquad \text{(B7)}\end{aligned}$$

for $\xi \geqslant h({}_{12i}\lambda)$. Here we have used the abbreviation q_j for $q_j({}_{12i}\lambda)$ $(j = 1, i)$ etc. It is also important to note that $r_1 = r_i$ $(= r_2)$ and $c_1 = c_i$ $(= c_2)$. Similarly

$$Q_i'(\xi, {}_{12i}\lambda) > 0 \qquad \text{(B8)}$$

for $\xi \geqslant h({}_{12i}\lambda)$. Defining

$$\cosh\kappa_1 = (b_1^2 - a_1 c_1)^{-1/2}(b_i^2 - a_i c_i)^{-1/2} Q_i(\xi, {}_{12i}\lambda), \qquad \text{(B9)}$$

$$\cosh\kappa_2 = (b_2^2 - a_2 c_2)^{-1/2}(b_i^2 - a_i c_i)^{-1/2} Q_i'(\xi, {}_{12i}\lambda), \qquad \text{(B10)}$$

we can then use the method of Appendix A of P and Eq. (31) to show that Eq. (B5) and hence Eq. (B3) hold.

(iii) $D_i(\xi, {}_{12i}\lambda) < 0$, $D_i'(\xi, {}_{12i}\lambda) < 0$. In this case we define $\cos\phi_1$ (resp. $\cos\phi_2$) by the right-hand side of Eq. (B9) [resp. (B10)] and again Eq. (B5) and hence Eq. (B3) hold.

APPENDIX C

In this appendix we outline the method of interchanging the order of $\lim_{\epsilon_j \downarrow 0} \partial/\partial m_j^2$ $(3 \leqslant j \leqslant n)$ and the integrations with respect to ξ and η, which is needed to obtain Eq. (67) in Sec. 8. The method is very similar to that described in Secs. 6, 7 and Appendix B of I and in Sec. 8 of P. From Eqs. (22), (14), and (15) we find that $(\partial/\partial m_j^2) h(\lambda) = O(\lambda_j)$ as $\lambda_j \downarrow 0$ and from Eqs. (40), (31), and (30) $(\partial/\partial m_j^2) f_+(\xi, {}_{12}\lambda) = O(\lambda_j)$ as $\lambda_j \downarrow 0$. Further, as noted after Eq. (54), $A_j(\xi, \eta)$ and $B_j(\xi, \eta, {}_{12j}\lambda)$ depend on m_j^2 whereas $C_j(\xi, \eta, {}_{12j}\lambda)$ $[\equiv \bar{F}(\xi, \eta, {}_{12j}\lambda)]$ does not. Thus the argument of Sec. 8 of P can be used to show that $I_{n+2}(y_{ij})$ given in Eq. (64) can also be written as in Eq. (66). In fact, it follows from Sec. 8 of P that, with O_3 defined in Eq. (65),

$$O_3[\bar{F}(\xi, \eta, {}_{12}\lambda)]^{-1/2} = \frac{-\frac{1}{2}(\partial/\partial m_3^2)\{[B_3(\xi, \eta, {}_{123}\lambda)]^2 - A_3(\xi, \eta) C_3(\xi, \eta, {}_{123}\lambda)\}}{[\bar{F}(\xi, \eta, {}_{123}\lambda)]^{1/2}\{[B_3(\xi, \eta, {}_{123}\lambda)]^2 - A_3(\xi, \eta) C_3(\xi, \eta, {}_{123}\lambda)\}}. \qquad \text{(C1)}$$

The term in square brackets in Eq. (C1) never vanishes in the region of integration in Eq. (66) since, as can be seen from Eqs. (55), (56), (59), and (B3),

$$\{[B_j(\xi, \eta, {}_{12j}\lambda)]^2 - A_j(\xi, \eta) C_j(\xi, \eta, {}_{12j}\lambda)\} > 0 \qquad \text{(C2)}$$

for all $\xi \geqslant h({}_{12j}\lambda)$, $\eta \geqslant f_+(\xi, {}_{12j}\lambda)$, ${}_{12j}\lambda \geqslant 0$. In fact, $O_3[\bar{F}(\xi, \eta, {}_{12}\lambda)]^{-1/2}$ can be majorized by $M_3[\bar{F}(\xi, \eta, {}_{123}\lambda)]^{-1/2}$, where M_3 is a positive constant. To repeat the process of interchanging the order of $\lim_{\epsilon_j \downarrow 0} \partial/\partial m_j^2$ $(3 < j \leqslant n)$ and the ξ and η integrations, it is necessary in addition to use the theorem given, for example, in Section 225 of Hobson[28] and the fact that $O_j \ldots O_3[\bar{F}(\xi, \eta, {}_{12}\lambda)]^{-1/2}$ can be majorized by $M_j[\bar{F}(\xi, \eta, {}_{123\cdots j}\lambda)]^{-1/2}$, where M_j is a positive constant.

[1] N. N. Bogoliubov, B. V. Medvedev, and M. K. Polivanov, Fortschr. Phys. 6, 159 (1959); H. J. Bremermann, R. Oehme, and J. G. Taylor, Phys. Rev. 109, 2178 (1958); H. Lehmann, Nuovo Cimento 10, 579 (1958).

[2] R. J. Eden, P. V. Landshoff, D. I. Olive, and J. C. Polkinghorne, *The Analytic S-Matrix* (Cambridge U. P., Cambridge, 1966).

[3] N. Nakanishi, *Graph Theory and Feynman Integrals* (Gordon and Breach, New York, 1971).

[4] I. T. Todorov, *Analytic Properties of Feynman Diagrams in Quantum Field Theory* (Pergamon, New York, 1971).

[5] R. Karplus, C. M. Sommerfield, and E. H. Wichmann, Phys. Rev. 111, 1187 (1958).

[6] Y. Nambu, Nuovo Cimento 9, 610 (1958).

[7] R. Oehme, Phys. Rev. 111, 1430 (1958).

[8] S. Mandelstam, Phys. Rev. 115, 1741, 1752 (1959).

[9] S. Mandelstam, Phys. Rev. 112, 1344 (1959).

[10] L. D. Landau, Nucl. Phys. 13, 181 (1959).

[11] R. E. Cutkosky, J. Math. Phys. 1, 429 (1960).

[12] P. V. Landshoff and S. B. Treiman, Phys. Rev. 127, 649 (1962); J. J. Kubis and J. L. Gammel, Phys. Rev. 172, 1664 (1968).

[13] R. E. Norton, Phys. Rev. 135, B1381 (1964).

[14] R. E. Cutkosky, Rev. Mod. Phys. 33, 448 (1961).

[15] R. C. Hwa and V. L. Teplitz, *Homology and Feynman Integrals* (Benjamin, New York, 1966).

[16] D. Fotiadi, M. Froissart, J. Lascoux, and F. Pham, "Analytic Properties of some integrals over a complex manifold," Paris (1963) (preprint); Topology 4, 159 (1965) (reprinted in Ref. 15).

[17] D. Fotiadi and F. Pham, pp. 192—244 of Ref. 15; G. Ponzano, T. Regge, E.R. Speer, and M.J. Westwater, Commun. Math. Phys. 18, 1 (1970); further references are given in Refs. 4 and 15.

[18] P. Federbush, J. Math. Phys. 6, 274 (1965); J.M. Westwater, Helv. Phys. Acta 40, 596 (1967).

[19] J.S. Frederiksen and W.S. Woolcock, Nucl. Phys. B 28, 605 (1971).

[20] J.S. Frederiksen and W.S. Woolcock, Ann. Phys. (N.Y.) 75, 503 (1973).

[21] J.S. Frederiksen, "Spectral representation of the pentagon diagram amplitude," to be published in J. Math. Phys.

[22] V.E. Asribekov, Zh. Eksp. Teor. Fiz. 42 565 (1962) [Sov. Phys. JETP 15, 394 (1962)].

[23] S.S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961).

[24] J. Tarski, J. Math. Phys. 1, 149 (1960).

[25] J.S. Frederiksen and W.S. Woolcock, Ann. Phys. (N.Y.) 80, 86, (1973).

[26] L. F. Cook and J. Tarski, J. Math. Phys. 3, 1 (1962).

[27] D. Bessis and F. Pham, J. Math. Phys. 4, 1253 (1963).

[28] E.W. Hobson, *The Theory of Functions of a Real Variable and the Theory of Fourier's Series* (Dover, New York, 1957).

Principle of compensation of dangerous diagrams for boson systems. III. Finite temperature*

Donald H. Kobe and Gerald W. Goble

Department of Physics, North Texas State University, Denton, Texas 76203

(Received 5 November 1973)

The principle of compensation of dangerous diagrams (PCDD) is derived at finite temperature for boson systems by minimizing the average number of Bogoliubov quasiparticles in the system. The conditions obtained state that (a) the amplitude for the creation (or annihilation) of a single quasiparticle is zero and (b) the amplitude for the creation (or annihilation) of a pair of quasiparticles is zero. These conditions are expanded in finite-temperature perturbation theory, using both the density matrix and Green's function methods. In first order the resulting equations are the Hartree–Fock–Bogoliubov equations for a homogeneous boson system at finite temperature which can also be obtained from a free energy variational principle.

I. INTRODUCTION

Bogoliubov[1] originally formulated the *principle of compensation of dangerous diagrams* (PCDD) as a means of determining the coefficients in his canonical transformation to quasiparticles, which we will call bogolons. By choosing the coefficients such that the sum of all the vacuum to two-bogolon diagrams vanished,[2] he was able to eliminate diagrams in the perturbation expansion of the ground-state energy which diverged and were hence "dangerous." Although motivated originally by boson systems,[3] the PCDD had its first applications to superconductivity[1] where it was shown that the compensation of the lowest order dangerous diagrams (CLODD) gave the same result as the energy variational principle.[4] Higher order corrections in the PCDD were later shown to be important in both fermion systems[5] and boson systems.[6] Hence it became important to justify the PCDD on more fundamental grounds than the vanishing of divergent diagrams in an expansion whose convergence was unknown.

In two previous papers on boson systems,[7,8] the PCDD was justified on the basis of some variational principles. In Paper I the overlap between the true ground-state wavefunction and the bogolon vacuum state was maximized to obtain a form of the PCDD. In Paper II the expected number of bogolons in the true ground state was minimized to obtain the PCDD. These papers were extensions of previous work on fermion systems[9] to boson systems.

In the present paper it is shown that the PCDD II can be extended to finite temperatures by minimizing the grand canonical average number of bogolons in the system. The conditions obtained state that (a) the amplitude for the creation (or annihilation) of a single bogolon is zero, and (b) the amplitude for the creation (or annihilation) of a pair of bogolons is zero. These amplitudes can then be expanded in finite-temperature perturbation theory. The density matrix perturbation expansion[10] is developed for a temperature dependent unperturbed Hamiltonian and perturbation, which is shown to be the same as for a temperature independent unperturbed Hamiltonian and perturbation. The bogolon Green's functions developed in II are extended to finite temperature by replacing the unperturbed single-bogolon propagator at zero temperature by the unperturbed finite-temperature propagator.[11] These perturbation methods are used to expand the PCDD to first order to obtain the *compensation of the lowest order dangerous diagrams* (CLODD). The resulting equations are the Hartree—Fock—Bogoliubov (HFB) equations[12] for a homogeneous boson system at finite temperature. These equations were apparently first derived by Tolmachev.[13]

The HFB equations or modifications of them have been obtained by previous authors by a variety of methods. After the sucess of the pairing theory of superconductivity, many authors attempted similar theories of superfluidity, with and without explicit treatment of the zero-momentum single-particle condensate. The pair theory of Girardeau and Arnowitt,[14] based on the energy variational principle, gave a theory in which both the single-particle condensate and pair correlations were taken into account at zero temperature. Unfortunately, their theory had the unphysical feature of a gap in the energy spectrum at zero momentum. This theory was extended to finite temperature by Wentzel,[15] who used the concept of the thermodynamically equivalent Hamiltonian, and by Girardeau,[16] who used a free energy variational principle. Wentzel's theory was further developed and discussed by Luban,[17] who treated the zero-momentum condensate in a different way than by replacing the particle operators for the state by c-numbers.[3] The treatment of the condensate by Valatin and Butler[18] at zero temperature, and later by Valatin[19] at finite temperature, was in such a way as to eliminate the gap in the excitation spectrum. Their approach leads to other difficulties, as was pointed out by Kobe.[20] The HFB equations for a homogeneous boson system at finite temperature have also been obtained by the Green's function method by several authors.[13,21]

In the Soviet Union, work of Bogoliubov, Zubarev, and Tserkovnikov[22] on phase transitions using a form of the PCDD at nonzero temperature was applied by the latter two authors to the nonideal Bose gas.[23] The equations they obtained are very similar to the equations of Tolmachev,[13,24] who used both Green's functions and a free energy variational principle, as well as making a remark about the PCDD in lowest order. The equations he obtained are the HFB equations for a homogeneous boson system at finite temperature. A modification of this theory was also developed by Gelikman,[25] who did not obtain a gap in the single-particle spectrum.

In the next section a canonical transformation, which treats the zero momentum state exactly, is made on the full Hamiltonian to obtain the bogolon Hamiltonian. A free energy, which is an upper bound to the true free energy, is constructed in Sec. III from an unperturbed Hamiltonian in terms of free bogolons. This free energy is varied in Sec. IV to obtain the HFB equations for a homogeneous boson system at nonzero temperature. In Sec. V the PCDD is derived by minimizing the average number of bogolons in the system at nonzero temperature. A finite temperature perturbation theory is developed in Sec. VI for the density matrix. The PCDD is expanded to first order in Sec. VII to obtain the HFB equations. In Sec. VIII the same set of equations is shown to follow from the finite-temperature generalization of the bogolon Green's functions. Finally the conclusions are given in Sec. IX.

II. BOGOLON HAMILTONIAN

The grand canonical Hamiltonian for a system of bosons interacting with each other through the two-body potential V is[26]

$$H=\sum_k(e_k-\mu)a_k^\dagger a_k+\tfrac{1}{2}\sum_{1234}\langle 12|V|34\rangle a_1^\dagger a_2^\dagger a_3 a_4, \qquad (2.1)$$

where $e_k=k^2/2m$ is the kinetic energy of a particle of mass m and momentum **k**, and μ is the chemical potential. The creation and annihilation operators, $a_1^\dagger$ and a_1, respectively, for a particle with momentum $(1)=(\mathbf{k}_1)$, satisfy the usual boson commutation relations. The matrix elements of the potential $\langle 12|V|34\rangle$ in Eq. (2.1) are symmetrized.

A partial diagonalization of the Hamiltonian can be obtained by making a canonical transformation[3] to Bogoliubov quasiparticles or bogolons. The canonical transformation[27]

$$a_k=\phi_0\delta_{k0}+u_k\gamma_k+v_k\gamma_{-k}^\dagger \qquad (2.2)$$

expresses the particle annihilation operator as a linear combination of the bogolon creation and annihilation operators, $\gamma_{-k}^\dagger$ and γ_k, respectively. The zero-momentum state can be macroscopically occupied, which is taken into account by the c-number ϕ_0. There is, of course, a gain in generality by using the ϕ_0 since, as a special case, it can be zero. The use of the bogolon operators for zero momentum insures that the particle creation and annihilation operators for zero momentum satisfy the commutation relations. In order for the bogolons to be bosons, the bogolon creation and annihilation operators must also satisfy boson commutation relations. The coefficients u_k and v_k in the canonical transformation must then satisfy

$$u_k^2-v_k^2=1, \qquad (2.3)$$

and also be even functions of **k**.

When the canonical transformation in Eq. (2.2) is made on the Hamiltonian of Eq. (2.1), and all the terms are normal ordered, the Hamiltonian can be written as[7]

$$H=\sum_{jk}H_{jk}, \qquad (2.4)$$

where $j,k=0,1,2,3,4$ and $j+k=0,1,2,3,$ and 4. The term H_{jk} has j bogolon creation operators and k bogolon annihilation operators,

$$H_{jk}=\sum h_{jk}(1,2,\ldots,j;j+1,\ldots,j+k)$$
$$\times\gamma_1^\dagger\cdots\gamma_j^\dagger\gamma_{j+1}\cdots\gamma_{j+k}, \qquad (2.5)$$

where the sum is over all momenta. The coefficient $h_{jk}(1,2,\ldots,j;j+1,\ldots,j+k)$ is symmetric with respect to the interchange of the first j variables, and also with respect to the interchange of the last k variables. The coefficients involve the matrix elements of the potential and the coefficients in the canonical transformation, and are given in Table II of I.

III. FREE ENERGY

A free energy function is constructed in this section which is an upper bound to the exact free energy in the grand canonical ensemble. This variational free energy is the sum of the free energy of the unperturbed system and the average in the unperturbed system of the perturbation. That this variational free energy is an upper bound to the exact free energy is based on an inequality for the partition function due to Peierls,[28] and has variously been attributed to Bogoliubov by Kvasnikov[29] and to Schultz[29b] by Valatin.[4] Girardeau[16] gives a brief sketch of a proof due to Mühlschlegel.[30] Because of its importance and for the sake of completeness it is given in Appendix A. The exact free energy of the system in the grand canonical ensemble is

$$F=-k_BT\ln Z, \qquad (3.1)$$

where k_B is Boltzmann's constant and T is the absolute temperature. The grand partition function Z is defined as

$$Z=\mathrm{Tr}[\exp(-\beta H)], \qquad (3.2)$$

since H in Eq. (2.1) contains μN, where N is the number operator, and $\beta=(k_BT)^{-1}$.

An unperturbed temperature-dependent Hamiltonian H_0 can be defined which is diagonal in the bogolon representation,

$$H_0(\beta)=U+\sum_k E_k\gamma_k^\dagger\gamma_k, \qquad (3.3)$$

where the bogolon kinetic energy E_k, the bogolon operators γ_k, and U are all dependent on the temperature. The best choice of E_k, γ_k, and U are made later on the basis of the variational principle. The unperturbed Hamiltonian H_0 is added to and subtracted from the Hamiltonian H, so that the perturbation,

$$H'(\beta)=H-H_0(\beta), \qquad (3.4)$$

is also temperature dependent. The full Hamiltonian in Eq. (2.1) is not temperature dependent.

The upper bound to the exact free energy,[29,30] discussed above, is (see Appendix A)

$$F\leqslant F_0+\langle H'\rangle_0\equiv F_{\mathrm{var}}, \qquad (3.5)$$

which defines the variational free energy F_{var}. The free energy F_0 for the unperturbed system is

$$F_0=-k_BT\ln Z_0, \qquad (3.6)$$

where the unperturbed grand partition function is

$$Z_0=\mathrm{Tr}[\exp(-\beta H_0)]. \qquad (3.7)$$

The unperturbed partition function can be easily evaluated in the bogolon representation, and the unperturbed free energy is thus

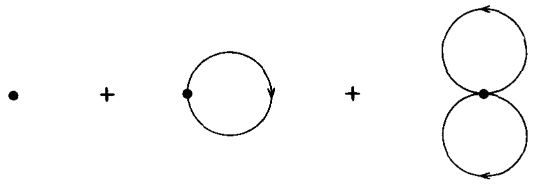


FIG. 1. The graphical representation of the internal energy in Eq. (3.14). The lines denote bogolon lines, and the vertex with j lines in and k lines out denotes h_{jk}.

$$F_0 = U + k_B T \sum_k \ln[1 - \exp(-\beta E_k)]. \tag{3.8}$$

The average in the unperturbed system is defined as

$$\langle \cdots \rangle_0 = Z_0^{-1} \operatorname{Tr}[\exp(-\beta H_0) \cdots]. \tag{3.9}$$

Thus the average of the perturbation H', needed in Eq. (3.5), is

$$\langle H' \rangle_0 = \langle H \rangle_0 - U - \sum_k E_k \bar{n}_k, \tag{3.10}$$

from Eq. (3.4), where the average bogolon occupation number is

$$\bar{n}_k = [\exp(\beta E_k) - 1]^{-1}. \tag{3.11}$$

The variational free energy in Eq. (3.5) is obtained by adding Eqs. (3.8) and (3.10), and can be written in the familiar form[31]

$$F_{\text{var}} = \langle H \rangle_0 - TS_0. \tag{3.12}$$

The entropy S_0 for a system of noninteracting bogolons is

$$S_0 = k_B \sum_k [(\bar{n}_k + 1) \ln(\bar{n}_k + 1) - \bar{n}_k \ln \bar{n}_k], \tag{3.13}$$

and is obtained by eliminating βE_k in terms of $\bar{n}_k$ by means of Eq. (3.11).

The internal energy (minus $\mu \langle N \rangle_0$ where N is the particle number operator) in Eq. (3.12) can be evaluated in the unperturbed ensemble by substituting Eq. (2.4) into Eq. (3.9), which gives

$$\langle H \rangle_0 = H_{00} + \sum_k h_{11}(k,k) \bar{n}_k + 2 \sum_{kp} h_{22}(k,p,p,k) \bar{n}_k \bar{n}_p. \tag{3.14}$$

The average of the operators $\gamma_k^\dagger \gamma_p^\dagger \gamma_p \gamma_k$ was used in obtaining Eq. (3.14) and is given in Eq. (B4) of Appendix B. The internal energy in this approximation is shown in Fig. 1, where the three terms are in one-to-one correspondence with the terms in Eq. (3.14). The lines shown are bogolon lines and the vertices are bogolon vertices. When the expressions for H_{00}, h_{11}, and h_{22} from Table II of I are substituted into Eq. (3.14), the result is

$$\begin{aligned}
\langle H \rangle_0 = {} & -\mu \phi_0^2 + \tfrac{1}{2} \langle 00 | V | 00 \rangle \phi_0^4 \\
& + \sum_k (e_k - \mu)[v_k^2 + (u_k^2 + v_k^2) \bar{n}_k] \\
& + \phi_0^2 \sum_k \langle 00 | V | -kk \rangle u_k v_k (1 + 2\bar{n}_k) \\
& + 2\phi_0^2 \sum_k \langle 0k | V | k0 \rangle [v_k^2 + (u_k^2 + v_k^2) \bar{n}_k] \\
& + \tfrac{1}{2} \sum_{kp} \langle k, -k | V | -p, p \rangle u_k v_k (1 + 2\bar{n}_k) u_p v_p (1 + 2\bar{n}_p) \\
& + \sum_{kp} \langle k, p | V | p, k \rangle [v_k^2 + (u_k^2 + v_k^2) \bar{n}_k][v_p^2 + (u_p^2 + v_p^2) \bar{n}_p],
\end{aligned} \tag{3.15}$$

as shown in Appendix C. Equation (3.12) together with Eqs. (3.14) and (3.15) are used in the following section.

IV. MINIMIZATION OF THE FREE ENERGY

The inequality in Eq. (3.5) shows that F_{var} in Eq. (3.12) is an upper bound to the exact free energy.[32] To obtain the least upper bound of this form, F_{var} is minimized with respect to $\bar{n}_k$, u_k, v_k, and ϕ_0, subject to the constraint of Eq. (2.3). The equations resulting from the variation will, of course, determine these parameters. The $\bar{n}_k$ in Eq. (3.11) can be considered an unknown parameter, since it is a function of the unknown bogolon kinetic energy E_k in Eq. (3.3).

The minimization of F_{var} in Eq. (3.12) with respect to $\bar{n}_k$ gives

$$\frac{\partial F_{\text{var}}}{\partial \bar{n}_k} = \xi_k + \beta^{-1} \ln[\bar{n}_k / (\bar{n}_k + 1)] = 0. \tag{4.1}$$

The term ξ_k is defined as

$$\xi_k = h_{11}(k,k) + 4 \sum_p h_{22}(k,p,p,k) \bar{n}_p, \tag{4.2}$$

where h_{11} and h_{22} are given in Eqs. (C2) and (C3), and is shown graphically in Fig. 2. The differentiation of Eq. (3.14) with respect to $\bar{n}_k$ corresponds to cutting a bogolon line in Fig. 1 which gives Fig. 2. Equation (4.1) implies that

$$\bar{n}_k = [\exp(\beta \xi_k) - 1]^{-1}, \tag{4.3}$$

so that on comparison with Eq. (3.11) the bogolon kinetic energy E_k in Eq. (3.3) is determined to be

$$E_k = \xi_k. \tag{4.4}$$

When Eqs. (C2) and (C3) are substituted into Eq. (4.2) and Eq. (4.4) is used, the bogolon energy,

$$E_k = \xi_k = U_k (u_k^2 + v_k^2) + \Delta_k 2 u_k v_k, \tag{4.5}$$

is obtained. The single-particle energy U_k is the kinetic energy minus the chemical potential plus dressing,

$$U_k = e_k - \mu + 2 \langle k0 | V | 0k \rangle \phi_0^2 + f_k, \tag{4.6}$$

where the noncondensate Hartree–Fock dressing is

$$f_k = 2 \sum_p \langle kp | V | pk \rangle [v_p^2 + (u_p^2 + v_p^2) \bar{n}_p]. \tag{4.7}$$

The pair potential Δ_k in Eq. (4.5) is

$$\Delta_k = \langle k, -k | V | 00 \rangle \phi_0^2 + g_k, \tag{4.8}$$

where the noncondensate contribution

$$g_k = \sum_p \langle k, -k | V | -p, p \rangle u_p v_p (1 + 2\bar{n}_p) \tag{4.9}$$

describes the scattering of pairs of particles with equal but opposite momentum.

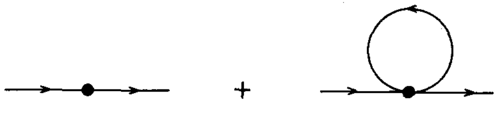


FIG. 2. The graphical representation of the bogolon energy in Eq. (4.2).

The minimization of $F_{\rm var}$ in Eq. (3.12) with respect to u_k and v_k subject to the constraint in Eq. (2.3) gives the condition

$$U_k 2u_k v_k + \Delta_k(u_k^2 + v_k^2) = 0, \tag{4.10}$$

from which the u_k and v_k are determined. Equation (4.10) together with Eq. (2.3) shows that the coefficients in the canonical transformation of Eq. (2.2) satisfy

$$2u_k v_k = -\Delta_k(U_k^2 - \Delta_k^2)^{-1/2} \tag{4.11}$$

and

$$u_k^2 + v_k^2 = U_k(U_k^2 - \Delta_k^2)^{-1/2}. \tag{4.12}$$

When these expressions are substituted into the bogolon energy in Eqs. (4.4) and (4.5) the result is

$$E_k = (U_k^2 - \Delta_k^2)^{1/2}. \tag{4.13}$$

The free energy in Eq. (3.12) can also be minimized with respect to the condensate amplitude ϕ_0,

$$\frac{\partial F_{\rm var}}{\partial \phi_0} = [-\mu + \langle 00|V|00\rangle\phi_0^2 + g_0 + f_0]2\phi_0 = 0. \tag{4.14}$$

If the zero-momentum state is macroscopically occupied, then $\phi_0 \neq 0$ and Eq. (4.14) determines the chemical potential to be[33]

$$\mu = \langle 00|V|00\rangle\phi_0^2 + f_0 + g_0. \tag{4.15}$$

When this expression for the chemical potential is used in Eq. (4.13), a gap in the energy spectrum at $\mathbf{k} = 0$ is obtained,

$$E_0 = 2\phi_0[-g_0\langle 00|V|00\rangle]^{1/2}, \tag{4.16}$$

which was first found by Girardeau and Arnowitt.[14] This gap is not physical and violates the Hugenholtz and Pines[34] theorem. If the variation principle in Eq. (4.14) is not used, the chemical potential can be chosen to eliminate the gap. The result is the right-hand side of Eq. (4.15) minus $2g_0$.

The density of particles in the unperturbed system can be obtained by taking the average of the particle number operator N in the unperturbed ensemble, and dividing by the volume Ω. On substituting the number operator N into Eq. (3.9), we obtain

$$\langle N\rangle_0/\Omega = \phi_0^2/\Omega + \Omega^{-1}\sum_p [v_p^2 + (u_p^2 + v_p^2)\bar{n}_p]. \tag{4.17}$$

If $\langle N\rangle_0/\Omega$ is taken to be the density of the original system, then ϕ_0 can be determined. As Girardeau[16] has pointed out this condition does not follow from the variational principle since Eq. (3.5) is valid for fixed μ. He uses Eq. (4.15) to determine ϕ_0 and then the exact chemical potential is unknown. It may, however, be determined approximately by using perturbation theory.

V. MINIMUM NUMBER OF BOGOLONS

The coefficients in the canonical transformation of Eq. (2.2) can be obtained from a variational principle other than the one of the last section. The principle used here is an extension of the PCDD II to finite temperature. The average number of bogolons in the system is

$$\langle n\rangle = \sum_k \langle\gamma_k^\dagger\gamma_k\rangle, \tag{5.1}$$

where the average is with respect to the grand canonical ensemble. The average in the grand canonical ensemble is defined as

$$\langle\cdots\rangle = Z^{-1}\,{\rm Tr}[\exp(-\beta H)\cdots], \tag{5.2}$$

where Z is the grand partition function in Eq. (3.2). Since H does not depend on the coefficients in the canonical transformation in Eq. (2.2), they enter only through the dependence of γ_k and $\gamma_k^\dagger$ on them.

A criterion for the choice of the coefficients in the transformation is to minimize the average number of bogolons in the system. With fewer bogolons present, the bogolon interactions will not be as important and the bogolons will behave more like an ideal gas. Then it can be expected that the free bogolon model will be a better approximation to the true system. Equation (2.2) can be used to express the $\gamma_k^\dagger$ and γ_k in terms of the particle creation and annihilation operators and u_k, v_k, and ϕ_0.

When Eq. (5.1) is minimized with respect to ϕ_0, the result is[8]

$$\langle\gamma_0^\dagger\rangle = 0, \tag{5.3}$$

which states that the amplitude for the creation (or annihilation) of a single bogolon is zero. In graphical form it is shown in Fig. 3a. In the language of perturbation theory, Eq. (5.3) states that the sum of all the diagrams leading to the creation (or annihilation) of a single bogolon is zero. This condition is the formulation of the augmented PCDD for boson systems[8] at finite temperatures, and is due to the exact treatment of the zero-momentum state.

Minimizing Eq. (5.1) with respect to u_k and v_k subject to the constraint in Eq. (2.3), we obtain[8]

$$\langle\gamma_k^\dagger\gamma_{-k}^\dagger\rangle = 0, \tag{5.4}$$

which states that the amplitude for the creation (or annihilation) of a pair of bogolons is zero. In the language of perturbation theory, Eq. (5.4) states that the sum of all the diagrams leading to the creation (or annihilation) of two bogolons is zero. This condition is the formulation of Bogoliubov's principle of compensation of dangerous diagrams at finite temperature, and is shown graphically in Fig. 3b.

The question arises as to the connection between the PCDD of this section and the free energy variational principle of the last section. The next section develops

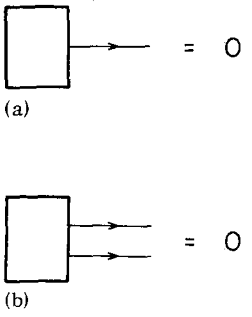


FIG. 3. The principle of compensation of dangerous diagrams (PCDD) in graphical form at finite temperature: (a) creation of a single bogolon; (b) creation of a pair of bogolons.

a finite temperature perturbation theory which is applied in Sec. VII to show that in first order the PCDD gives the same results as in Sec. IV.

VI. PERTURBATION THEORY

The criteria in Eqs. (5.3) and (5.4) for the coefficients in the canonical transformation cannot be used unless they are expanded in perturbation theory. In this section a finite temperature perturbation theory for bogolons is developed. Since the unperturbed Hamiltonian H_0 in Eq. (3.3) depends on the temperature, the perturbation H' in Eq. (3.4) also depends on the temperature. The usual finite-temperature perturbation theory[10] must therefore be somewhat modified.

The grand canonical density matrix,

$$\rho(\beta)=\exp(-\beta H), \tag{6.1}$$

satisfies the Bloch equation,

$$\frac{-\partial\rho(\beta)}{\partial\beta}=H\rho(\beta). \tag{6.2}$$

The Hamiltonian H in terms of some arbitrary inverse temperature α which will be determined later is

$$H=H_0(\alpha)+H'(\alpha), \tag{6.3}$$

from Eq. (3.4). The density matrix can be written in terms of the unperturbed Hamiltonian $H_0(\alpha)$ as

$$\rho(\beta)=\exp[-\beta H_0(\alpha)]W(\beta,\alpha). \tag{6.4}$$

By substituting Eq. (6.4) into Eq. (6.2) and using Eq. (6.3), the equation for the operator $W(\beta,\alpha)$ is

$$\frac{-\partial W(\beta,\alpha)}{\partial\beta}=H_I'(\alpha,\beta)W(\beta,\alpha), \tag{6.5}$$

where the operator in the interaction picture for finite temperature is

$$H_I'(\alpha,\beta)=\exp[\beta H_0(\alpha)]H'(\alpha)\exp[-\beta H_0(\alpha)]. \tag{6.6}$$

Equation (6.5) can be converted into an integral equation by integrating, and we obtain

$$W(\beta,\alpha)=1-\int_0^\beta d\beta_1 H_I'(\alpha,\beta_1)W(\beta_1,\alpha). \tag{6.7}$$

On iterating Eq. (6.7), the perturbation expansion

$$W(\beta,\alpha)=\sum_{n=0}^{\infty}(-1)^n\int_0^\beta d\beta_1\int_0^{\beta_1}d\beta_2\cdots\int_0^{\beta_{n-1}}d\beta_n \times H_I'(\beta_1,\alpha)H_I'(\beta_2,\alpha)\cdots H_I'(\beta_n,\alpha) \tag{6.8}$$

is obtained. Now the arbitrary inverse temperature α can be set equal to β and Eq. (6.4) becomes

$$\rho(\beta)=\exp[-\beta H_0(\beta)]W(\beta,\beta), \tag{6.9}$$

where $W(\beta,\beta)$ is obtained from Eq. (6.8) with $\alpha=\beta$. The result is exactly the same as if the temperature dependence of $H_0(\beta)$ and $H'(\beta)$ had been ignored.[10] In the following section this expansion is used to obtain a perturbation expansion of the PCDD.

VII. COMPENSATION OF THE LOWEST ORDER DANGEROUS DIAGRAMS

In this section the perturbation theory developed in the last section is used in first order in connection with the PCDD of Sec. V to obtain the *compensation of the lowest order dangerous diagrams* (CLODD). It is shown that the CLODD is completely equivalent to the results obtained from the free energy variational principle in Sec. IV.

The PCDD for the single-bogolon amplitude given in Eq. (5.3) can be written in first order of perturbation theory as

$$\langle H_{10}\gamma_0\rangle_0+\langle H_{21}\gamma_0\rangle_0=0, \tag{7.1}$$

on using Eqs. (5.2), (6.9), (3.4), and (2.4). The average is defined in Eq. (3.9). The other terms in H' do not contribute in Eq. (7.1) since the number of bogolons created must equal the number annihilated.

When Eq. (2.5) is used in Eq. (7.1), the result is

$$h_{10}\langle\gamma_0^\dagger\gamma_0\rangle_0+\sum_{pqr}h_{21}(pqr)\langle\gamma_p^\dagger\gamma_q^\dagger\gamma_r\gamma_0\rangle_0=0\,. \tag{7.2}$$

The number of bogolons in the zero-momentum state is given by Eq. (3.11) with $\mathbf{k}=0$, which is not infinity since $E_0\neq 0$ by Eq. (4.16). The average of the four operators in Eq. (7.2) is given by Eq. (B4) for $\mathbf{k}=0$. Therefore, Eq. (7.2) becomes

$$h_{10}+2\sum_p h_{21}(0pp)\overline{n}_p=0, \tag{7.3}$$

which is shown graphically in Fig. 4a. The first term in Fig. 4a describes the creation of a single bogolon, and the second term describes the creation of two bogolons and the annihilation of one.

From Table II of I the coefficient h_{10} is

$$h_{10}=(-\mu+\sum_k\langle k-k|V|00\rangle u_kv_k+2\sum_k\langle 0k|V|k0\rangle v_k^2 + \langle 00|V|00\rangle\phi_0^2)\phi_0(u_0+v_0), \tag{7.4}$$

and the coefficient $h_{21}(0,k,k)$ is

$$h_{21}(0,k,k)=\phi_0(u_0+v_0)[\langle 0k|V|k0\rangle(u_k^2+v_k^2) + \langle k-k|V|00\rangle u_kv_k]. \tag{7.5}$$

When these coefficients are substituted into Eq. (7.3), the result is

$$[-\mu+\langle 00|V|00\rangle\phi_0^2+g_0+f_0]2\phi_0=0, \tag{7.6}$$

which is the same as Eq. (4.14). Thus the PCDD to first order in Fig. 4a gives the same result as the minimization of the free energy with respect to ϕ_0 in Sec. IV.

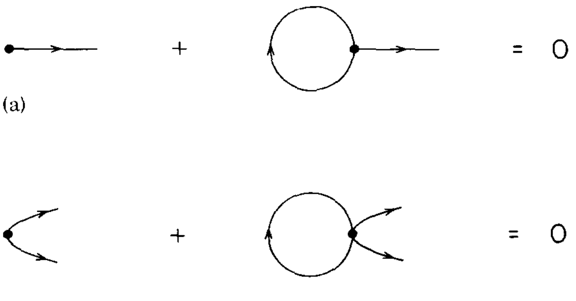


FIG. 4. Compensation of the lowest order dangerous diagrams (CLODD) at finite temperature: (a) creation of a single bogolon; (b) creation of a pair of bogolons.

The PCDD for the two-bogolon amplitude given in Eq. (5.4) can be written in first order of perturbation theory as

$$\langle H_{20}\gamma_k\gamma_{-k}\rangle_0+\langle H_{31}\gamma_k\gamma_{-k}\rangle_0=0, \tag{7.7}$$

on using Eqs. (5.2), (6.9), (3.4), and (2.4). The average is defined in Eq. (3.9). The other terms in H' do not contribute in Eq. (7.7) because the number of bogolons created must equal to the number annihilated. When Eq. (2.5) is used in Eq. (7.7) the result is

$$\sum_p h_{20}(p,-p)\langle\gamma_p^\dagger\gamma_{-p}^\dagger\gamma_k\gamma_{-k}\rangle_0+\sum_{pqrs} h_{31}(pqrs) \times\langle\gamma_p^\dagger\gamma_q^\dagger\gamma_r^\dagger\gamma_s\gamma_k\gamma_{-k}\rangle_0=0. \tag{7.8}$$

The average values in Eq. (7.8) are given in Eqs. (B5) and (B6) for $\mathbf{k}\neq 0$ and in Eqs. (B7) and (B8) for $\mathbf{k}=0$. Equation (7.8) for all $\mathbf{k}$ thus becomes

$$h_{20}(k,-k)+3\sum_p h_{31}(p,k,-k,p)\overline{n}_p=0, \tag{7.9}$$

which is shown graphically in Fig. 4b. The first term in Fig. 4b corresponds to the creation of a pair of bogolons, and the second to the creation of three and the annihilation of one.

From Table II of I the coefficient $h_{20}(k,-k)$ is

$$h_{20}(k,-k)=\Bigl(e_k-\mu+2\sum_p\langle kp|V|pk\rangle(v_p^2+\phi_0^2\delta_{p0})\Bigr) \times u_kv_k+\frac{1}{2}\sum_p\langle p-p|V|-kk\rangle(u_pv_p+\delta_{p0}\phi_0^2) \times(u_k^2+v_k^2), \tag{7.10}$$

and the coefficient h_{31} is

$$3h_{31}(p,k,-k,p)=\langle kp|V|pk\rangle(u_p^2+v_p^2)u_kv_k +\langle k-p|V|-pk\rangle(u_p^2+v_p^2)u_kv_k +\langle k-k|V|-pp\rangle u_pv_p(u_k^2+v_k^2). \tag{7.11}$$

When the coefficients in Eqs. (7.10) and (7.11) are substituted into Eq. (7.9) the result is

$$U_k2u_kv_k+\Delta_k(u_k^2+v_k^2)=0, \tag{7.12}$$

which is the same as Eq. (4.10). Thus the PCDD to first order in Fig. 4b gives the same result as the minimization of the free energy with respect to u_k and v_k in Sec. IV.

The bogolon energy E_k can be determined from the thermal average of $\gamma_kH_0\gamma_k^\dagger$ where H_0 is given in Eq. (3.3). In lowest order of perturbation theory, the average of $\gamma_kH_0\gamma_k^\dagger$ is its average for the unperturbed system in Eq. (3.9), and is

$$\langle\gamma_kH_0\gamma_k^\dagger\rangle_0/(\overline{n}_k+1)=\langle H_0\rangle_0+E_k, \tag{7.13}$$

where the average is divided by $\overline{n}_k+1$ for normalization. The average of H_0 in the unperturbed system

$$\langle H_0\rangle_0=U+\sum_p E_p\overline{n}_p, \tag{7.14}$$

is the unperturbed internal energy.

Equation (7.13) can be compared with the result obtained from calculating $\langle\gamma_kH\gamma_k^\dagger\rangle$. In lowest order of perturbation theory, we obtain

$$\langle\gamma_kH\gamma_k^\dagger\rangle_0=H_{00}\langle\gamma_k\gamma_k^\dagger\rangle_0+\sum_p h_{11}(p,p)\langle\gamma_k\gamma_p^\dagger\gamma_p\gamma_k^\dagger\rangle_0 +\sum_{pqrs} h_{22}(pqrs)\langle\gamma_k\gamma_p^\dagger\gamma_q^\dagger\gamma_r\gamma_s\gamma_k^\dagger\rangle_0, \tag{7.15}$$

on using Eqs. (3.9), (2.4), and (2.5). The average required are given in Eqs. (B9)–(B11), which gives the result

$$\langle\gamma_kH\gamma_k^\dagger\rangle_0/(\overline{n}_k+1)=\langle H\rangle_0+\xi_k, \tag{7.16}$$

where $\langle H\rangle_0$ is given in Eq. (3.14) and ξ_k is given in Eq. (4.2). The right-hand side of Eq. (7.16) is the internal energy plus one excitation of energy ξ_k.

The unperturbed Hamiltonian in Eq. (3.3) can be determined by equating Eqs. (7.13) and (7.16). The bogolon energy then becomes

$$E_k=\xi_k, \tag{7.17}$$

where ξ_k is given in Eqs. (4.2) and (4.5). The unperturbed Hamiltonian H_0 and the Hamiltonian H are chosen to have the same average value in the unperturbed ensemble, so that U is determined by

$$U=\langle H\rangle_0-\sum_p E_p\overline{n}_p \tag{7.18}$$

from Eq. (7.14).

The result in Eq. (7.17) is the same as Eq. (4.4), so the compensation of the lowest order dangerous diagrams (CLODD) in this section gives the same result as the free energy variational principle in Sec. IV.

VIII. GREEN'S FUNCTION THEORY

In this section the Green's function equations of motion for the bogolons obtained in II at zero temperature are generalized to finite temperature, and used to obtain the CLODD. The results obtained are identical to the results of the last section and Sec. IV.

The many-time causal propagator or Green's function describing the annihilation of n bogolons and the creation of m bogolons with all possible procesces allowed by the Hamiltonian taking place is

$$\mathcal{G}_{nm}(1,2,\ldots,n+1,\ldots,n+m) =i\langle T\{\gamma_1\gamma_2\cdots\gamma_n\gamma_{n+1}^\dagger\cdots\gamma_{n+m}^\dagger\}\rangle, \tag{8.1}$$

where $j=(\mathbf{k}_j,t_j)$ is the momentum $\mathbf{k}_j$ and time t_j associated with bogolon $j=1,2,\ldots,n+m$. The creation and annihilation operators are all in the Heisenberg picture, and the time-ordering operator T orders the creation and annihilation operators with the largest time on the left and the smallest on the right in descending order. The average in Eq. (8.1) is the average over the grand canonical ensemble defined in Eq. (5.2). The temperature β^{-1} is considered only as a parameter in this approach. The Fourier transform of $\mathcal{G}_{nm}$ in the sense of Eq. (7.2) of II is denoted as G_{nm}.

The equations of motion for $\mathcal{G}_{nm}$ may be obtained exactly as for the zero-temperature case in II. The method followed there was first used for fermion bogolons.[35] For particles a similar method was used at zero temperature,[36] and later extended to finite temperatures.[11]

The equations of motion for the finite temperature case have the same structure as the equations at zero temperature in Figs. 1, 2, and 3 of II. However, the

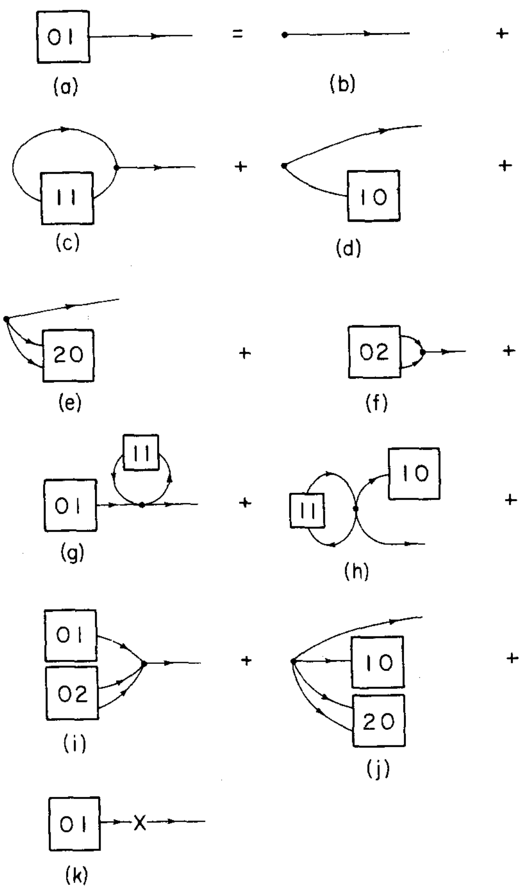


FIG. 5. The approximate equation of motion for the single-bogolon creation propagator. The propagators involving several bogolons in the exact equation have been factored.

zero-temperature unperturbed propagator in Eq. (7.5) of II is replaced by the finite-temperature unperturbed propagator

$$G^0(1,2)=(2\pi)^{-1}\delta_{12}\left(-\frac{(\bar{n}_1+1)}{\omega_1-E_1+i0}+\frac{\bar{n}_1}{\omega_1-E_1-i0}\right), \quad (8.2)$$

where $\bar{n}_1$ is the average bogolon occupation number in Eq. (3.11) and E_1 is the bogolon kinetic energy in Eq. (3.3). Equation (8.2) is the Fourier transform of the bogolon propagator

$$\mathcal{G}^0_{11}(1,2)=i\langle T\{\gamma_1\gamma_2^\dagger\}\rangle_0 \quad (8.3)$$

in the unperturbed system.[11]

Since the perturbation at nonzero temperatures is given by Eq. (3.4), there is an extra perturbation of the form

$$\sum_{12}[h_{11}(1,2)-E_1\delta_{12}]\gamma_1^\dagger\gamma_2 \quad (8.4)$$

added to the interaction Hamiltonian in II. The appropriate modifications must then be made in Sec. 7 of II.

The equation of motion for $G_{01}(1)$, the Fourier transform of $\mathcal{G}_{01}(1)$, is obtained from Fig. 3 of II for $n=0$, $m=1$. The Green's functions involving several bogolons can be factorized into Green's functions involving fewer bogolons in the usual way[35] [cf. Eq. (8.6)] and the resulting equation for G_{01} is given in Fig. 5. Figures 5b and 5c are irreducible in the sense that they cannot be split into two parts by cutting either one bogolon line or two bogolon lines pointing in the same direction. All the other diagrams in Fig. 5 are reducible. The compensation of these lowest order irreducible dangerous diagrams gives

$$2\pi i\int d1'h'_{01}(1')G^0(1'1)+4\pi\int d1'\,d2'\,d3'h'_{12}(3'2'1')$$
$$\times G_{11}(3'2')G^0(1'1)=0, \quad (8.5)$$

where the h'_{jk} is h_{jk} multiplied by a δ function for the conservation of energy. When the unperturbed propagator in Eq. (8.2) is substituted into Eq. (8.5) for $G_{11}(3',2')$ and the integral is closed in the upper-half ω'_3-plane, the complex conjugate of Eq. (7.3) is obtained from which the chemical potential is determined. Equation (8.5) in this approximation is shown graphically in Fig. 4a. When the irreducible dangerous diagrams are compensated, then Fig. 5 is a homogeneous equation for G_{01} involving only the functions G_{01}, G_{10}, G_{02}, and G_{20}.

The equation of motion for the propagator G_{02} can be obtained from Fig. 3 of II by setting $n=0$, $m=2$. When the propagators involving several bogolons are factorized into those for fewer, the equation shown in Fig. 6 is obtained. The diagrams of Figs. 6c and 6f show the irreducible dangerous diagrams that cannot be divided into two parts by cutting either one bogolon line or two bogolon lines going in the same direction. The diagram of Fig. 6f occurs because the two bogolon propagator G_{22} is factored,

$$G_{22}(1234)=-i[G_{11}(14)G_{11}(23)+G_{11}(13)G_{11}(24)]. \quad (8.6)$$

The lowest order irreducible dangerous diagrams can be

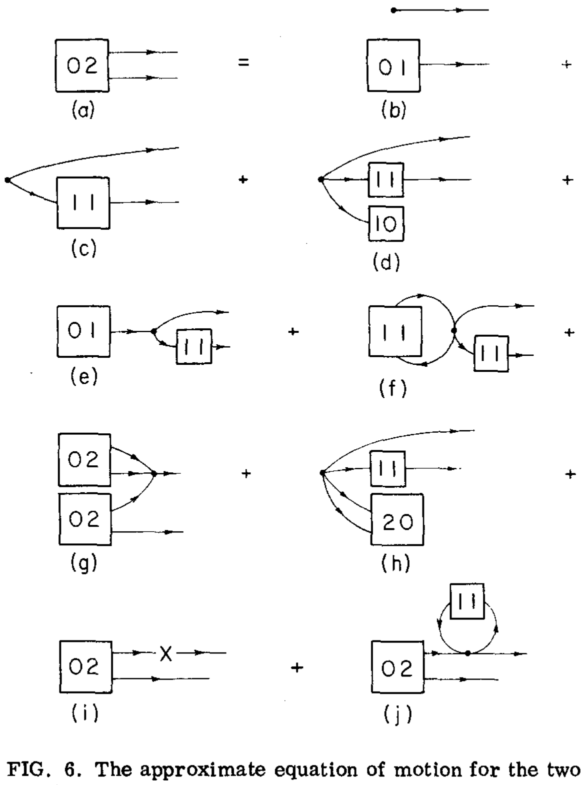


FIG. 6. The approximate equation of motion for the two-bogolon creation propagator. The propagators involving several bogolons in the exact equation have been factored.

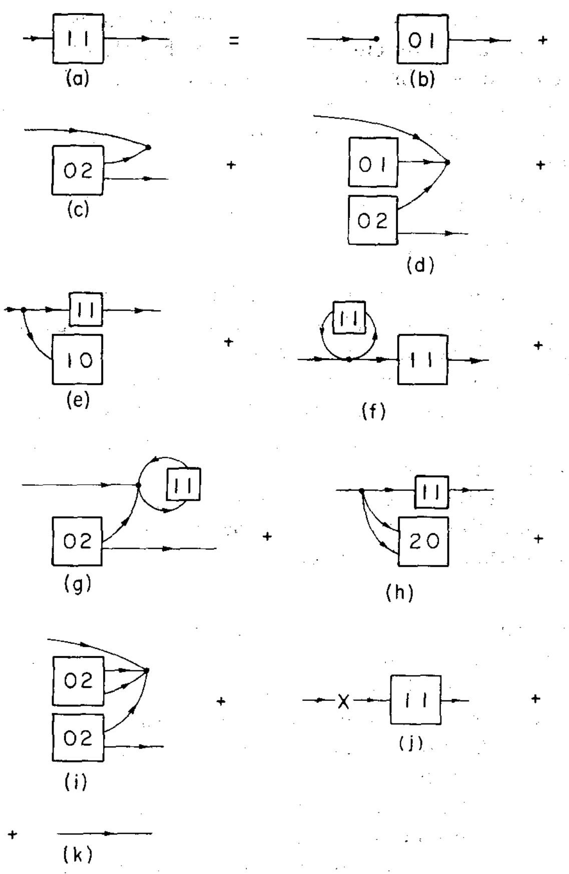


FIG. 7. The approximate equation of motion for the single-bogolon propagator. The propagators involving several bogolons in the exact equation have been factored.

compensated, which gives

$$4\pi \int d1'\,d2' h'_{02}(1'2')G^0(1'1)G_{11}(2'2) - 12\pi i \int d1'\,d2'\,d3'\,d4' h'_{13}(4'3'2'1') \times G_{11}(3'4')G^0(1'1)G_{11}(2'2) = 0. \qquad (8.7)$$

When Eq. (8.2) is substituted into Eq. (8.7) for $G_{11}(3'4')$, and the integral is closed in the upper-half ω'_3-plane, the complex conjugate of Eq. (7.9) is obtained which determines the coefficients u_k and v_k. Equation (8.7) in this approximation is shown graphically in Fig. 4b.

The equations in Fig. 6 for G_{02} and Fig. 5 for G_{01} are homogeneous after the compensation of the irreducible dangerous diagrams, and involve only the functions G_{01}, G_{10}, G_{02}, and G_{20}. Together with the corresponding homogeneous equations for G_{10} and G_{20} they have the trivial solution

$$G_{01} = G_{10} = G_{02} = G_{20} = 0. \qquad (8.8)$$

Therefore, in this order all the diagrams corresponding to the dangerous processes of creation or annihilation of a single bogolon or a pair of bogolons are zero.

The equation of motion for the single bogolon propagator can be obtained from Fig. 1 of II by setting $n=1$, $m=1$. When the higher order bogolon propagators are factorized, the result in Fig. 7 is obtained. All the diagrams involving G_{01}, G_{10}, G_{02}, or G_{20} are zero because of Eq. (8.8). The only surviving diagrams are those of Figs. 7f, 7j, and 7k. A further simplification occurs if the bogolon energy E_k is chosen such that Fig. 7j just cancels Fig. 7f. In other words, the bogolon energy is chosen such that the irreducible self-energy vanishes, which gives the condition

$$-2\pi \int d1'\,d2'[h'_{11}(1'2') - E_{1'}\delta_{1'2'}]G^0(11') \times G_{11}(2'2) + 8\pi i \int d1'\,d2'\,d3'\,d4' \times h'_{22}(1'3'4'2')G_{11}(4'3')G^0(11')G_{11}(2'2) = 0. \qquad (8.9)$$

When the propagator $G_{11}(4', 3')$ in the second term is replaced by the unperturbed propagator in Eq. (8.2) and the integrals performed, Eq. (4.2) for the bogolon energy is obtained. Equation (8.9) in this approximation is shown in Fig. 2.

Because of Eqs. (8.8) and (8.9), the bogolon propagator $G_{11}(1, 2)$ is

$$G_{11}(12) = G^0(12). \qquad (8.10)$$

In this order the replacement of the propagator G_{11} in Eqs. (8.5), (8.7), and (8.9) by G^0 involves no approximation. The bogolon kinetic energy, the chemical potential, and the coefficients u_k and v_k are all determined from these three equations.

Therefore, the finite temperature Green's function approach for the compensation of the lowest order dangerous diagrams gives the same result as obtained in Secs. VII and IV. It can be extended to higher-order dangerous diagrams, but this will not be done here.

IX. CONCLUSION

In this paper the principle of compensation of dangerous diagrams (PCDD) proposed by Bogoliubov[1,2] is generalized to nonzero temperature. The average number of bogolons in the system is minimized, which should make the free bogolon model a better approximation to the true system. The result of this variational principle is the vanishing of the amplitudes describing the creation or annihilation of a pair of bogolons or a single bogolon. In first order of finite-temperature perturbation theory,[10] the compensation of the lowest order dangerous diagrams (CLODD) is identical with the equations obtained from a free energy variational principle.[13] This result is the finite temperature generalization of the result that at zero temperature the CLODD was derived from the Rayleigh—Ritz energy variational principle.[7]

The compensation of dangerous diagrams to higher orders would give results differing from the free energy variational principle. Since the free energy variational principle gives only an upper bound, a free energy closer to the exact one could be obtained. In the case of the charged Bose gas at zero temperature, the compensation of dangerous diagrams to second order was shown to be important in eliminating divergences arising in the lower order approximations.[6] At finite temperature it would also be expected that the second order dangerous diagrams would be important.

The extension to finite temperatures of the PCDD I in which the overlap between the true ground state vector and the bogolon vacuum state was maximized does not seem to be generalizable to nonzero temperatures.

The methods used in this paper are of course applicable to fermion systems with only minor modifications. Since in fermion systems there is no zero-momentum condensate, all the terms involving ϕ_0 would be zero. There are no diagrams involving the creation (or annihilation) of only a single fermion bogolon. The criterion of minimizing the average number of bogolons in the system gives the condition that the amplitude for the creation (or annihilation) of a pair of bogolons is zero. When expanded to first order in perturbation theory to obtain the CLODD, the HFB equations[12] for a superconductor at finite temperature are obtained.[4,29,30]

ACKNOWLEDGMENTS

We would like to thank Professor Marvin D. Girardeau for a communication concerning the determination of the chemical potential in Sec. IV and for bringing Ref. 38 to our attention.

APPENDIX A: UPPER BOUND FOR THE FREE ENERGY

An inequality satisfied by the grand partition function Z in Eq. (3.2) was proved by Peierls,[28]

$$Z \geq \sum_n \exp[-\beta(\Phi_n, H\Phi_n)], \tag{A1}$$

where the set of functions $\{\Phi_n\}$ is orthonormal, but not necessarily complete.[37] The Hamiltonian H is broken into an unperturbed part H_0 in Eq. (3.3) and a perturbation H' in Eq. (3.4). The states $\{\Phi_n\}$ are taken to be the eigenstates of H_0 with eigenvalues $\{E_n^0\}$. Then Eq. (A1) can be written as

$$Z \geq Z_0 \sum_n w_n \exp[-\beta(\Phi_n, H'\Phi_n)] \tag{A2}$$

where $w_n = Z_0^{-1}\exp(-\beta E_n^0)$ and Z_0 is defined in Eq. (3.7). On using the lemma proved by Huang[37] that the average of a function with positive second derivative is greater than the function of the average, we have

$$Z \geq Z_0 \exp(-\beta\langle H'\rangle_0). \tag{A3}$$

When this inequality is substituted into Eq. (3.1), the inequality in Eq. (3.5) is obtained.

A very thorough review of variational methods in classical and quantum statistical mechanics has recently been published by Girardeau and Mazo.[38] They refer to the inequality of Eq. (3.5) as the Gibbs—Bogoliubov inequality, since its classical form was originally proved by Gibbs.

APPENDIX B: AVERAGES OF OPERATORS

The averages of various operators in the unperturbed ensemble that are required throughout the paper are given here. The average of the bogolon number operator $\gamma_k^\dagger\gamma_k$ is obtained by using the average defined in Eq. (3.9). The result given in Eq. (3.11) is obtained from

$$\bar{n}_k = \langle\gamma_k^\dagger\gamma_k\rangle_0 = -\frac{\partial \ln Z_0}{\partial(\beta E_k)}\,. \tag{B1}$$

The average of the number operator squared is

$$\langle n_k^2\rangle_0 = \langle\gamma_k^\dagger\gamma_k\gamma_k^\dagger\gamma_k\rangle_0 = \frac{Z_0^{-1}\partial^2 Z_0}{\partial(\beta E_k)^2}\,, \tag{B2}$$

which reduces to

$$\langle n_k^2\rangle_0 = 2\bar{n}_k^2 + \bar{n}_k. \tag{B3}$$

The relative variance in the number of bogolons in the state k is thus of order one.

By using the above method and the boson commutation relations, it can be shown that

$$\langle\gamma_k^\dagger\gamma_p^\dagger\gamma_p\gamma_k\rangle_0 = \begin{cases} \bar{n}_k\bar{n}_p & \text{for } \mathbf{k}\neq\mathbf{p}, \\ 2\bar{n}_k^2 & \text{for } \mathbf{k}=\mathbf{p}, \end{cases} \tag{B4}$$

which is used in Eq. (3.14). In Eq. (7.8) the averages for $\mathbf{k}\neq 0$,

$$\langle\gamma_p^\dagger\gamma_{-p}^\dagger\gamma_k\gamma_{-k}\rangle_0 = \bar{n}_k\bar{n}_{-k}(\delta_{pk} + \delta_{p,-k}) \tag{B5}$$

and

$$\langle\gamma_p^\dagger\gamma_k^\dagger\gamma_{-k}^\dagger\gamma_p\gamma_k\gamma_{-k}\rangle_0 = \begin{cases} \bar{n}_p\bar{n}_k\bar{n}_{-k} & \text{for } \mathbf{p}\neq\mathbf{k}, -\mathbf{k}, \\ 2\bar{n}_k^2\bar{n}_{-k} & \text{for } \mathbf{p}=\mathbf{k}, \\ 2\bar{n}_k\bar{n}_{-k}^2 & \text{for } \mathbf{p}=-\mathbf{k}, \end{cases} \tag{B6}$$

are needed, with the others being zero by momentum conservation. For $\mathbf{k}=0$ the averages needed in Eq. (7.8) are

$$\langle\gamma_0^\dagger\gamma_0^\dagger\gamma_0\gamma_0\rangle = 2\bar{n}_0^2 \tag{B7}$$

and

$$\langle\gamma_p^\dagger\gamma_0^\dagger\gamma_0^\dagger\gamma_p\gamma_0\gamma_0\rangle_0 = \begin{cases} 2\bar{n}_p\bar{n}_0^2 & \text{for } \mathbf{p}\neq 0, \\ 6\bar{n}_0^3 & \text{for } \mathbf{p}=0. \end{cases} \tag{B8}$$

In calculating the bogolon energy in Eq. (7.15) the average of the four operators

$$\langle\gamma_k\gamma_p^\dagger\gamma_p\gamma_k^\dagger\rangle_0 = \begin{cases} \bar{n}_p(\bar{n}_k + 1) & \text{for } \mathbf{k}\neq\mathbf{p}, \\ (2\bar{n}_k + 1)(\bar{n}_k + 1) & \text{for } \mathbf{k}=\mathbf{p} \end{cases} \tag{B9}$$

is required. It is also necessary to have the average of the six operators for $\mathbf{k}\neq\mathbf{p}, \mathbf{q}$,

$$\langle\gamma_k\gamma_p^\dagger\gamma_q^\dagger\gamma_q\gamma_p\gamma_k^\dagger\rangle_0 = \begin{cases} \bar{n}_q\bar{n}_p(\bar{n}_k + 1) & \text{for } \mathbf{p}\neq\mathbf{q}, \\ 2\bar{n}_p(\bar{n}_k + 1) & \text{for } \mathbf{p}=\mathbf{q}, \end{cases} \tag{B10}$$

and for $\mathbf{k}=\mathbf{p}$,

$$\langle\gamma_k\gamma_k^\dagger\gamma_q^\dagger\gamma_q\gamma_k\gamma_k^\dagger\rangle_0 = \begin{cases} \bar{n}_q(2\bar{n}_k + 1)(\bar{n}_k + 1) & \text{for } \mathbf{k}\neq\mathbf{q}, \\ 2\bar{n}_k(3\bar{n}_k + 2)(\bar{n}_k + 1) & \text{for } \mathbf{k}=\mathbf{q}, \end{cases} \tag{B11}$$

in Eq. (7.15).

APPENDIX C: INTERNAL ENERGY

The expression for the internal energy in Eq. (3.14) is evaluated here. From Table II of I the internal energy at absolute zero is

$$\begin{aligned} H_{00} = {} & -\mu\phi_0^2 + \sum_k (e_k - \mu)v_k^2 + \tfrac{1}{2}\langle 00|V|00\rangle\phi_0^4 \\ & + \phi_0^2\sum_k \langle 00|V|-kk\rangle u_k v_k + 2\phi_0^2\sum_k \langle 0k|V|k0\rangle v_k^2 \\ & + \tfrac{1}{2}\sum_{kp} \langle k,-k|V|-p,p\rangle u_k v_k u_p v_p \\ & + \sum_{kp} \langle kp|V|pk\rangle v_k^2 v_p^2. \end{aligned} \tag{C1}$$

The coefficient $h_{11}(k,k)$ is the bogolon kinetic energy at absolute zero, and is

$$\begin{aligned} h_{11}(k,k) = {} & \Bigl(e_k - \mu + 2\sum_p \langle kp|V|pk\rangle(v_p^2 + \delta_{p0}\phi_0^2)\Bigr) \\ & \times(u_k^2 + v_k^2) + \Bigl(\sum_p \langle p,-p|V|-k,k\rangle \end{aligned}$$

$$\times(u_p v_p + \delta_{p0}\phi_0^2)\Big)\, 2u_k v_k, \tag{C2}$$

from Table II of I. The coefficient h_{22} in Eq. (3.14) describes the scattering of bogolons and can also be obtained from Table II of I:

$$\begin{aligned}2h_{22}(kppk) = {}&\langle kp\,|\,V\,|\,pk\rangle(u_k^2u_p^2 + v_k^2v_p^2)\\ &+\langle k-p\,|\,V\,|-pk\rangle(u_k^2v_p^2+u_p^2v_k^2)\\ &+2\langle k-k\,|\,V\,|-pp\rangle u_pv_pu_kv_k.\end{aligned} \tag{C3}$$

When Eqs. (C1)—(C3) are substituted into Eq. (3.14) and combined, the expression for $\langle H\rangle_0$ in Eq. (3.15) is obtained.

The effect of the last two terms in Eq. (3.14) is to make the replacements

$$u_kv_k \to u_kv_k(1+2\overline{n}_k) \tag{C4}$$

and

$$v_k^2 \to v_k^2 + (u_k^2+v_k^2)\overline{n}_k \tag{C5}$$

in the ground-state energy H_{00} in Eq. (C1). These replacements are equivalent to replacing the unperturbed ground-state averages of a_ka_{-k} and $a_k^\dagger a_k$ by their unperturbed thermal averages in Eq. (3.9) for $\mathbf{k}\neq 0$.

[*] Supported in part by the North Texas State University Faculty Research Fund.

[1] N.N. Bogoliubov, Zh. Eksp. Teor. Fiz. **34**, 58 (1958) [Sov. Phys.-JETP **7**, 41 (1958)]; Nuovo Cimento **7**, 794 (1958).

[2] N.N. Bogoliubov, V.V. Tolmachev, and D.V. Shirkov, *A New Method in the Theory of Superconductivity* (Academy of Sciences of the USSR Press, Moscow, 1958), Chap. 1. (English translation: Consultants Bureau, New York, 1959); Fortshr. Physik **6**, 605 (1958).

[3] N.N. Bogoliubov, J. Phys. (USSR) **11**, 23 (1947).

[4] J.G. Valatin, Nuovo Cimento **7**, 843 (1958).

[5] E.M. Henley and L. Wilets, Phys. Rev. **133**, B1118 (1964); Phys. Rev. Lett. **11**, 326 (1963).

[6] C.-W. Woo and S.K. Ma, Phys. Rev. **159**, 176 (1967).

[7] D.H. Kobe, J. Math. Phys. **9**, 1779 (1968), henceforth called I.

[8] D.H. Kobe, J. Math. Phys. **9**, 1795 (1968), henceforth called II.

[9] D.H. Kobe, J. Math. Phys. **8**, 1200 (1967); Ann. Phys. (N.Y.) **40**, 395 (1966); Phys. Rev. **140**, A825 (1965).

[10] T. Matsubara, Prog. Theor. Phys. (Kyoto) **14**, 351 (1955).

[11] D.H. Kobe, Ann. Phys. (N.Y.) **75**, 9 (1973).

[12] N.N. Bogoliubov, Usp. Fiz. Nauk **67**, 549 (1959) [Sov. Phys.-Usp. **2**, 236 (1959)].

[13] V.V. Tolmachev, Doklady Akad. Nauk SSSR **134**, 1324 (1960) [Sov. Phys.-Doklady **5**, 984 (1960)]. The coefficient of the first summation in his Eqs. (8) and (9) should be $1/V$, not $1/2V$.

[14] M. Girardeau and R. Arnowitt, Phys. Rev. **113**, 755 (1959).

[15] G. Wentzel, Phys. Rev. **120**, 1572 (1960).

[16] M. Girardeau, J. Math. Phys. **3**, 131 (1962). In footnote 13 he discusses the free-energy variational principle.

[17] M. Luban, Phys. Rev. **128**, 965 (1962).

[18] J.G. Valatin and D. Butler, Nuovo Cimento **10**, 37 (1958).

[19] J.G. Valatin in *Lectures in Theoretical Physics*, edited by W.E. Brittin and W.R. Chappel (University of Colorado Press, Boulder, 1963), Vol. 6, pp. 345—372. On p. 357 he sets $\alpha_0=0$, where α_0 is the condensate amplitude, in order to eliminate the gap.

[20] D.H. Kobe, Ann. Phys. (N.Y.) **47**, 15 (1968).

[21] P.C. Hohenberg and P.C. Martin, Ann. Phys. (N.Y.) **34**, 291 (1965); V.K. Wong, University of California Radiation Laboratory Report Number 17159 (Berkeley, California, 1966).

[22] N.N. Bogoliubov, D.N. Zubarev, and Iu. A. Tserkovnikov, Doklady Akad. Nauk SSSR **117**, 5 (1957) [Sov. Phys.-Doklady **2**, 535 (1957)].

[23] D.N. Zubarev and Iu. A. Tserkovnikov, Doklady Akad. Nauk SSSR **120**, 991 (1958) [Sov. Phys.-Doklady **3**, 603 (1958)].

[24] V.V. Tolmachev, Doklady Akad. Nauk SSSR **135**, 41 (1960) [Sov. Phys.-Doklady **5**, 1190 (1961)].

[25] B.T. Gelikman, Doklady Akad. Nauk SSSR **123**, 430 (1958) [Sov. Phys.-Doklady **3**, 1168 (1958)].

[26] See, e.g., D.H. Kobe, Am. J. Phys. **34**, 1150 (1966) for a discussion of second quantization.

[27] E.P. Gross, Ann. Phys. (N.Y.) **9**, 292 (1960) [cf. Eq. (2.7)]; H. Ezawa, J. Math. Phys. **6**, 380 (1965); H. Umezawa, Acta Phys. Hung. **19**, 9 (1965); A. Coniglio and M. Marinaro, Nuovo Cimento **48**, 249 (1967).

[28] R.E. Peierls, Phys. Rev. **54**, 918 (1938).

[29a] I.A. Kvasnikov, Doklady Akad. Nauk SSSR **110**, 755 (1956); Doklady Akad. Nauk SSSR **119**, 675 (1958) [Sov. Phys.-Doklady 3, 329 (1958)]; Doklady Akad. Nauk SSSR **119**, 475 (1958) [Sov. Phys.-Doklady **3**, 318 (1958)]; I.A. Kvasnikov and V.V. Tolmachev, Doklady Akad. Nauk SSSR **120**, 273 (1958) [Sov. Phys.-Doklady **3**, 553 (1958)].

[29b] T.D. Schultz, Nuovo Cimento **8**, 943 (1958).

[30] H. Koppe and B. Mühlschlegel, Z. Physik **151**, 613 (1958); B. Mühlschlegel, Bayer. Akad. Wiss., München, Sitzber. Math.-naturw. Kl. **1960**, 123 (1960).

[31] This is the form used by Valatin in Ref. 19, p. 367.

[32] Except for the treatment of the condensate, this section is similar to Ref. 19, pp. 365—369.

[33] To obtain results equivalent to Valatin in Ref. 19, pp. 365—369, we would choose $\phi_0=0$ and choose the condensate density to be $v_0^2\cong u_0v_0$. Then the chemical potential could be chosen to eliminate the gap. However, this approach would lead to difficulties in Eq. (3.15) for the zero-momentum contribution, and for a strongly repulsive potential $u_0v_0<0$. See Ref. 20 for a discussion of these points.

[34] N.M. Hugenholtz and D. Pines, Phys. Rev. **116**, 489 (1959). J. Gavoret and P. Nozières, Ann. Phys. (N.Y.) **28**, 349 (1964).

[35] D.H. Kobe and W.B. Cheston, Ann. Phys. (N.Y.) **20**, 279 (1962).

[36] D.H. Kobe, J. Math. Phys. **7**, 1806 (1966).

[37] A simple proof of this inequality is given by K. Huang, *Statistical Mechanics* (Wiley, New York, 1963), pp. 220—222.

[38] M.D. Girardeau and R.M. Mazo, in *Advances in Chemical Physics*, edited by I. Prigogine and S.A. Rice (Wiley, New York, 1973), Vol. XXIV, pp. 187—255.

Exact next nearest neighbor degeneracy

R. B. McQuistan

Department of Physics and Laboratory for Surface Studies, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201

(Received 7 January 1974)

It is shown that $A[n_{11}, n_{101}, n_{111}, q, N]$, the arrangement degeneracy arising when q indistinguishable particles are placed on a one-dimensional lattice of N equivalent compartments so that n_{11} occupied nearest neighbors, n_{101} next nearest neighbors of the 101-type, and n_{111} next nearest neighbors of the 111-type are created, is given by $A[n_{11}, n_{101}, n_{111}, q, N] = \binom{N-2q+n_{11}+2}{q-n_{11}-n_{101}}\binom{q-n_{11}-1}{n_{101}}\binom{q-n_{11}}{n_{11}-n_{111}}\binom{n_{11}-1}{n_{111}}$. The normalization and first moment of the next nearest neighbor density are determined. Similar results for the vacant next nearest neighbor degeneracy are also presented.

I. INTRODUCTION

The present paper is concerned with the development of an expression which will describe, for simple particles on a one-dimensional lattice, the degeneracy of those arrangements containing a prescribed number of occupied nearest and next nearest pairs. We will couch the following discussion in terms of the vacancy and occupation of lattice sites. Obviously, the results are applicable to any binary variable such as magnetic spin or $A-B$ atoms in a binary alloy.

For purposes of the present discussion we will consider that there are two types of occupied next nearest neighbors: those with no intervening particle, which we will refer to as the 101-type (see Fig. 1A) and those in which an intervening particle is present, designated the 111-type (see Fig. 1B). The number of 101 next nearest neighbor pairs in an arrangement is n_{101} and n_{111} denotes the number of 111-type occupied next nearest neighbors in a single arrangement. Thus we consider the situation in which E_i, the interaction energy, can be written

$$E_i = n_{11}V_{11} + n_{101}V_{101} + n_{111}V_{111}, \tag{1}$$

where n_{11} is the number of occupied nearest neighbor pairs on the arrangement and V_{11}, V_{101}, and V_{111} are the appropriate energies of interaction. This expression for the interaction energy does not preclude the special case where $V_{101} = V_{111}$.

Specifically, in the present paper we seek to determine the multiplicity of those states characterized by n_{11}, n_{101}, and n_{111}, that is, we wish to calculate $A[n_{11}, n_{101}, n_{111}, q, N]$, the number of independent ways of arranging q indistinguishable particles on a one-dimensional space of N equivalent sites so that n_{11} occupied nearest neighbor pairs are created together with n_{101} next nearest neighbor pairs of the 101-type and n_{111} next nearest neighbor pairs of the 111-type.

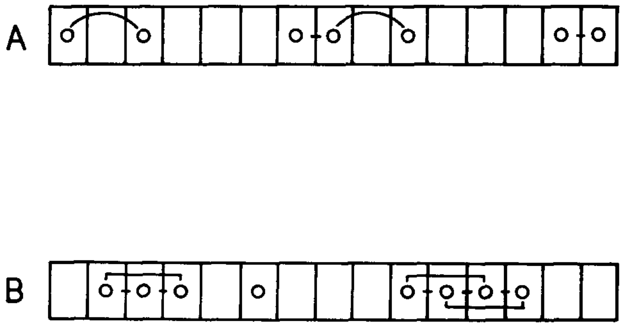


FIG. 1. A. This figure shows two next nearest neighbors of the 101-type. Two nearest neighbor pairs are also shown. B. Three next nearest neighbors of the 111-type are created from five occupied nearest neighbor pairs.

2. DETERMINATION OF $A[n_{11}, n_{101}, n_{111}, q, N]$

When q indistinguishable particles are arranged on a one-dimensional lattice of N equivalent sites to form n_{11}, n_{101}, and n_{111} pairs of nearest and next nearest neighbors there are always $q - n_{11} - n_{101}$ "units" formed (see Fig. 2). Here, we define such "units" to consist of

(1) a sequence of one or more occupied sites in which each occupied site is separated from its occupied neighbors by at most a single vacancy;

(2) two vacant sites (if needed) at one end of the sequence to separate the "unit" from other "units" on a particular arrangement.

Thus a "unit" is a contiguous sequence of nearest neighbor and/or next nearest neighbor pairs separated

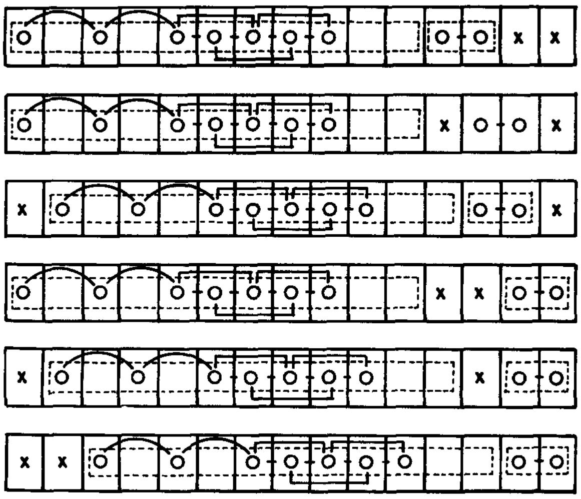


FIG. 2. For the situation in this figure, $N=15$, $q=9$, $n_{11}=5$, $n_{101}=2$, and $n_{111}=3$, we see that

$$\binom{N-2q+n_{11}+2}{q-n_{11}-n_{101}} = \binom{4}{2} = 6$$

reflects the fact that there are two "units", which we assume initially to be indistinguishable, and two indistinguishable, permutable vacancies (indicated by x's); these "units" and vacancies may be made to form six independent arrangements.

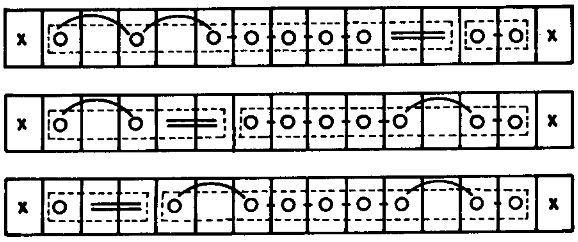


FIG. 3. In this figure we show that the separations within and between "units" may be interchanged to form new arrangements while conserving the number of "units" as well as q, n_{11} and n_{101}. We choose the third arrangement shown in Fig. 2. The single vacancies, of which there are two within the "units," may be permuted with the double vacancy within the "units" in

$$\binom{q-n_{11}-1}{n_{101}}=\binom{3}{2}=3$$

ways. Next nearest neighbors of the 101-type representing the number of separations consisting of a single vacancy are indicated by ⌒ and the other kind(s) of separations consisting of two or more contiguous vacancies are represented by ═.

from other "units" by two vacant sites at one end which serve to terminate the "unit" and isolate it from the rest of the particles on the array.

The reason there are $q-n_{11}-n_{101}$ "units" is that there are $q-1$ separations between the q particles on an array; n_{11} of these separations are between occupied nearest neighbors and n_{101} separations are between next nearest neighbor pairs of the 101-type. Consequently, there are $q-1-n_{11}-n_{101}$ separations which are neither between nearest neighbor nor next nearest neighbor pairs. Such separations are therefore composed of two or more vacancies and therefore between "units." If there are $q-1-n_{11}-n_{101}$ of these separations between "units" there must be $q-n_{11}-n_{101}$ "units." We shall assume initially that these units are identical (indistinguishable from one another) regardless of their composition and/or configuration.

These "units" can be permuted with some of the vacancies to form additional independent arrangements (see Fig. 2). Not all of the vacancies can be so used, however; some must be utilized to form n_{101} pairs and $2[q-n_{11}-n_{101}-1]$ vacancies are required to separate the "units" from other "units" (One of the "units" at the end of the array does not need two vacancies to separate it from the other "units"). There are available for permuting with the "units" $N-q-n_{101}-2[q-n_{11}-n_{101}-1]$ $=N-3q+2n_{11}+n_{101}+2$ indistinguishable vacancies. Thus there are a total of $[q-n_{11}-n_{101}]+[N-3q+2n_{11}+n_{101}+2]=N-2q+n_{11}+2$ objects which may be permuted to form

$$\binom{N-2q+n_{11}+2}{q-n_{11}-n_{101}}=\binom{N-2q+n_{11}+2}{N-3q+2n_{11}+n_{101}+2} \tag{2}$$

independent arrangements.

Initially we have assumed the "units" to be indistinguishable; obviously this is not the case. To determine $A[n_{11}, n_{101}, n_{111}, q, N]$ we must determine the number of ways that the "units" can be created from the particles and arrangements present on the array. Such a determination must be consistent with the constraints imposed by the requirement of a prescribed number of nearest and next nearest neighbor pairs.

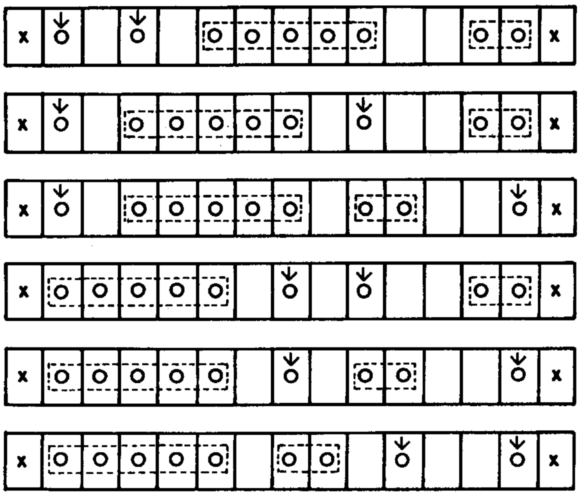


FIG. 4. In this figure we have selected the third arrangement in Fig. 2. There are $n_{11}-n_{111}=2$ groups consisting of two or more contiguous particles. We initially consider these groups to be indistinguishable from one another. There are $q-2n_{11}+n_{111}=2$ indistinguishable groups consisting of exactly of a single particle. These two groups may be arranged in

$$\binom{q-n_{11}}{n_{11}-n_{111}}=\binom{4}{2}=6$$

independent ways shown in this figure. The single particles are indicated by arrows and the groups consisting of two or more contiguous particles are surrounded by dashed boxes.

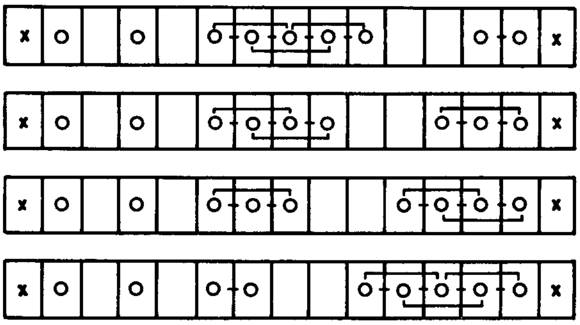


FIG. 5. In this figure we again consider the third arrangement shown in Fig. 2. We had assumed, in connection with Eq. 4 and Fig. 4, that the groups containing two or more contiguous particles were indistinguishable from each other. Such an assumption must be corrected by recognizing that the nearest neighbor pairs may be moved around within and between "units" to form independent arrangements while preserving the number of nearest neighbors and next nearest neighbors of the 111-type. This may be done in

$$\binom{n_{11}-1}{n_{111}}=\binom{4}{3}=4$$

ways for the situation depicted in the third drawing of the Fig. 2. The nearest neighbor pairs are shown as a short horizontal line and the next nearest neighbors of the 111-type are represented by ┌┐ or └┘.

We must now examine the number of ways in which the "units" may be constituted from the particles and the pairs of neighbors. The $q - n_{11} - n_{101}$ "units" may be created in different ways by the following processes:

(1) permuting the separations (between contiguous groups of particles) which consist of exactly one vacancy with those separations consisting of two or more contiguous vacancies (Fig. 3);

(2) permuting the single particles (separated from other particles by at least one vacancy) with the groups (indistinguishable) consisting of two or more contiguous particles (Fig. 4);

(3) constructing all the possible configurations involving the prescribed number of next nearest neighbors of the 111-type from the prescribed number of nearest neighbor pairs (Fig. 5).

Each of the three processes described above can lead to factors reflecting the creation of arrangements which conserve the number of "units" as required by Eq. 2 but in which the composition and/or configuration of the "units" is different. We now discuss each process in more detail.

(1) In Fig. 3 we observe that because the number of "units" is conserved, the number of separations between the units, $q - 1 - n_{11} - n_{101}$ is also conserved. These separations may be permuted with the prescribed number of next nearest neighbor pairs, of the 101-type thereby constructing new kinds of "units" without violating the constraints imposed by the stipulation of n_{11}, n_{101}, n_{111}, q, N or the number of "units." There are a total of $[q - 1 - n_{11} - n_{101}] + n_{101} = [q - 1 - n_{11}]$ of such separations and because the 101-type separations are indistinguishable from one another, as are the separations between the "units," these kinds of separations may be arranged in

$$\binom{q - n_{11} - 1}{n_{101}} = \binom{q - n_{11} - 1}{q - 1 - n_{11} - n_{101}} \tag{3}$$

independent ways.

Equation 3 indicates that the number of next nearest neighbor pairs of the 101-type cannot exceed the number of separations (between the particles) which are not involved in nearest neighbor pairs.

(2) In Fig. 4 we note that there are always $q - n_{11}$ groups composed of one or more contiguous particles. Each group is separated from other groups by one or more contiguous vacancies. There are always $q - n_{11}$ of such groups because there are a total of $q - n_{11} - 1$ separations between the particles which are *not* involved in nearest neighbor pairs.

Now the $q - n_{11}$ groups may be considered to be composed of groups consisting of two or more contiguous particles and another group consisting of single particles. Of the former kind there are $n_{11} - n_{111}$ and of the latter $[q - n_{11}] - [n_{11} - n_{111}] = [q - 2n_{11} + n_{111}]$. Each member of these two groups is indistinguishable from other members in the same group. Thus the $q - n_{11}$ groups may be permuted in

$$\binom{q - n_{11}}{n_{11} - n_{111}} = \binom{q - n_{11}}{q - 2n_{11} + n_{111}} \tag{4}$$

independent ways without changing the specified variables of the situation.

(3) Of course the groups consisting of two or more particles discussed in (2) above are not really indistinguishable by virtue of the fact that some of the groups may contain two contiguous particles, some may contain three contiguous particles, etc. It is possible to shift a particle from one such group to another to form additional independent arrangements without altering the total number of nearest neighbors and/or the number of 111-type next nearest neighbors (see Fig. 5). To determine the factor describing these changes we observe that between the n_{11} nearest neighbor pairs there are $n_{11} - 1$ separations. Of these n_{111} constitute indistinguishable next nearest neighbor pairs of the 111-type and $n_{11} - n_{111} - 1$ do not. These may be permuted in

$$\binom{n_{11} - 1}{n_{111}} = \binom{n_{11} - 1}{n_{11} - n_{111} - 1} \tag{5}$$

ways to form independent arrangements which satisfy the stipulated constraints on the enumeration process. Equation 5 describes the fact that the number of 111-type next nearest neighbor pairs on an arrangement cannot exceed the number of separations between nearest neighbor pairs.

Each one of the factors represented by Eqs. 3, 4, and 5 increase the multiplicity of the arrangement of the "units" with the vacancies described in Eq. 2. Thus

$$A[n_{11}, n_{101}, n_{111}, q, N] = \binom{N - 2q + n_{11} + 2}{q - n_{11} - n_{101}} \binom{q - n_{11} - 1}{n_{101}} \binom{q - n_{11}}{n_{11} - n_{111}} \binom{n_{11} - 1}{n_{111}}. \tag{6}$$

3. NORMALIZATION

If Eq. 6 is summed over all possible values of n_{101} and n_{111} the result should agree with the results[1] of a previous determination of the degeneracy of nearest neighbor pairs. As we have discussed in connection with Eqs. 3 and 5 the maximum number of next nearest neighbor pairs of the 101-type on an arrangement cannot exceed the number of separations not associated with nearest neighbor pairs, i.e.,

$$0 \leq n_{101} \leq q - n_{11} - 1, \tag{7}$$

and the maximum number of next nearest neighbor pairs of the 111-type on an arrangement cannot be greater than one less than the number of nearest neighbor pairs, i.e.,

$$0 \leq n_{111} \leq n_{11} - 1. \tag{8}$$

By the Vandermonde theorem[2] the sum

$$\sum_{n_{101}=0}^{q-n_{11}-1} \sum_{n_{111}=0}^{n_{11}-1} \binom{N - 2q + n_{11} + 2}{q - n_{11} - n_{101}} \binom{q - n_{11} - 1}{n_{101}} \binom{q - n_{11}}{n_{11} - n_{111}} \binom{n_{11} - 1}{n_{111}} = \binom{N - q + 1}{q - n_{111}} \binom{q - 1}{n_{11}}. \tag{9}$$

If Eq. 9 is summed over all possible values of the number of next nearest neighbors the result is just $\binom{N}{q}$.

4. DETERMINATION OF $A[n_{00}, n_{010}, n_{000}, q\ N]$

By reasoning similar to that employed in the previous section, one can determine the degeneracy of those states specified by n_{00} vacant nearest neighbor pairs, n_{010} vacant next nearest neighbor pairs of the 010-type and n_{000} vacant next nearest neighbor pairs of the 000-type. However, the desired result can more readily be obtained by means of the following transformations:

$$q \rightarrow N-q,$$

$$n_{11} \rightarrow n_{00},$$

$$n_{101} \rightarrow n_{010},$$

$$n_{111} \rightarrow n_{000}.$$

Then Eq. 6 becomes

$$A[n_{00}, n_{010}, n_{000}, q, N] = \binom{2q-N+2+n_{00}}{N-q-n_{00}-n_{010}}\binom{N-q-n_{00}-1}{n_{010}}\binom{N-q-n_{00}}{n_{00}-n_{000}}\binom{n_{00}-1}{n_{000}}. \tag{10}$$

5. FIRST MOMENT

The ensemble average number of next nearest neighbor pairs $\langle n_2\rangle = \langle n_{101}+n_{111}\rangle = \langle n_{101}\rangle + \langle n_{111}\rangle$ can be determined as follows:

$$\langle n_{101}\rangle = \binom{N}{q}^{-1} \sum_{n_{101}=0}^{q-1} \sum_{n_{101}=0}^{q-n_{11}-1} \sum_{n_{111}=0}^{n_{11}-1} n_{101}\, A[n_{11}, n_{101}, n_{111}, q, N] \tag{11}$$

$$= \frac{q(q-1)(N-q)}{N(N-1)}$$

and

$$\langle n_{11}\rangle = \binom{N}{q}^{-1} \sum_{n_{11}=0}^{q-1} \sum_{n_{101}=0}^{q-n_{11}-1} \sum_{n_{111}=0}^{n_{11}-1} n_{111}\, A[n_{11}, n_{101}, n_{111}, q, N] = \frac{q(q-1)(q-2)}{N(N-1)}. \tag{12}$$

Thus, in the limit $N\rightarrow\infty$, the ensemble average probability that a site is occupied by a particle which has an occupied next nearest neighbor is, according to Eqs. 11 and 12,

$$\frac{\langle n_2\rangle}{N} = \theta^2(1-\theta) + \theta^3 = \theta^2, \tag{13}$$

where $\theta \equiv \mathrm{Lim}_{N\rightarrow\infty} q/N$.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to Professor David Lichtman and Professor Masashi Tachiki for reading the manuscript of this paper and for useful discussions of it.

[1] R.B. McQuistan, J. Math. Phys. 13, 1317 (1972).
[2] J. Riordan, *Combinational Methods* (Wiley, New York 1968).

Schrödinger equation and quantum state codons in discrete transform space

Jean I. F. King

Air Force Cambridge Research Laboratories, Bedford, Massachusetts 01730

(Received 6 March 1974)

Quantum states of simple systems are shown to have composite root-pole structure in the complex transform plane. The Schrödinger condition becomes the inverse of a continuity equation expressing invariance of the ψ-transform codon under discrete displacement. Four distinct quotient polynomial solutions model the Legendre, Hermite, Laguerre, and Jacobi polynomial families. Schrödinger coefficients are identified with pole strengths of quotient polynomials and are geometrically interpreted in terms of universal root and pole interactions.

LEGENDRE FAMILY

For the hydrogenic atom the separated Schrödinger equation defining the spherical harmonic gives rise to the familiar equation

$$(1-x^2)P''-2(m+1)xP'+[\beta-m(m+1)]P=0. \qquad (1)$$

The eigenvalues and eigensolutions of this system can be determined by application of the finite Mellin transform

$$p(\kappa)\equiv\int_1^0 P(x)\,x^{-\kappa-1}dx, \qquad (2)$$

which yields, after collecting terms,

$$[(\kappa+m)(\kappa+m+1)-\beta]p(\kappa)-(\kappa+1)(\kappa+2)p(\kappa+2)=2mP(1). \qquad (3)$$

Clearly, the solution of the *homogeneous* equation $\mathcal{P}(\kappa)$ must satisfy the functional equation

$$\frac{\mathcal{P}(\kappa)}{\mathcal{P}(\kappa+2)}=\frac{(\kappa+1)(\kappa+2)}{(\kappa+m)(\kappa+m+1)-\beta}=\frac{(\kappa+1)(\kappa+2)}{(\kappa-l+m)(\kappa+l+m+1)}, \qquad (4)$$

where the eigenvalue spectrum $\beta=l(l+1)$ follows from the superposition requirement of integral roots for the quadratic denominator polynomial. The transform statement of the Schrödinger equation requires a $\mathcal{P}$-function structure such that its displacement two units toward positive κ is equivalent to the materialization of a pole at $\kappa=l-m$ and root suppression at $-(l+m+1)$, together with a pole annihilation and root creation at $\kappa=-1,\ -2$.

We verify by direct substitution that the functional equation is satisfied for $l=0,\ 1,\ 2,\dots,\ |m|\leqslant l$, by the *associated Legendre transform*[1]

$$\mathcal{P}_l^m(\kappa)=\prod_{n=1}^{[(l+m)/2]}(\kappa+l+m-2n+1)\Big/\prod_{n=0}^{[(l-m)/2}(\kappa-l+m+2n). \qquad (5)$$

The quotient polynomial character specifies the residues at the poles $l-m-2\lambda$ as

$$\begin{aligned}R_{l-m-2\lambda}&=\lim_{\kappa\to l-m-2\lambda}(\kappa-l+m+2\lambda)\mathcal{P}_l^m(\kappa)\\&=\prod_{n=1}^{[(l+m)/2]}(2l-2\lambda-2n+1)\Big/\prod_{n=0}^{[(l-m)/2]}(2n-2\lambda)\\&\qquad(\lambda=0,\ 1,\dots,[(l-m)/2]). \qquad (6)\end{aligned}$$

These pole strengths are identified for $m\geqslant 0$ as coefficients of corresponding associated Legendre polynomials[2]

$$\Theta(x)=(1-x^2)^{m/2}P_l^m(x)=(1-x^2)^{m/2}\sum_{\lambda=0}^{[(l-m)/2]}R_{l-m-2\lambda}x^{l-m-2\lambda}. \qquad (7)$$

We note that in transform space the $\mathcal{P}_l^m$ function possesses a root-pole structure for $m<0$ as well as for positive m. This is not true in Schrödinger space for which P_l^m has no polynomial representation for negative m.

Returning to the functional equation, it can be shown using the asymptotic behavior of the transform that for negative m the inhomogeneous term vanishes. In these cases it follows that $\mathcal{P}_l^m(\kappa)=p_l^m(\kappa)$. For positive m the quotient polynomial becomes "improper," with more roots than poles. Although this relation no longer holds, the Schrödinger polynomial coefficients nevertheless remain identified with the residues $R_{l-m-2\lambda}$ at the poles of the homogeneous solution $\mathcal{P}_l^m$, enabling us to ignore the inhomogeneous terms in this and subsequent coefficient specifications.

Figure 1 shows the root-pole patterns of the Associated Legendre transform $\mathcal{P}_6^2$ and $\mathcal{P}_6^1$ as well as the manner in which structural displacement arises from the creation and annihilation of root and pole form factors.

HERMITE FAMILY

For the linear harmonic oscillator the Schrödinger solution factors into $\psi=H(x)\exp(-x^2/2)$, with H the eigensolution of the Hermite equation

$$H''-2xH'+(2E/\hbar\omega-1)H=0. \qquad (8)$$

By operating on the equation with the finite Mellin transform we obtain, after collecting terms,

$$\begin{aligned}&2[\kappa-(E/\hbar\omega-\tfrac{1}{2})]h(\kappa)-(\kappa+1)(\kappa+2)h(\kappa+2)\\&=-(\kappa-1)H(1)-H'(1). \qquad (9)\end{aligned}$$

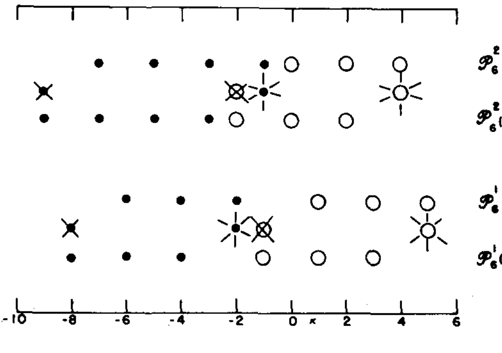

$\mathcal{P}_6^2(\kappa)$
$\mathcal{P}_6^2(\kappa+2)$
$\mathcal{P}_6^1(\kappa)$
$\mathcal{P}_6^1(\kappa+2)$
-10
-8
-6
-4
-2
0
κ
2
4
6

FIG. 1. Legendre quotient polynomial root-pole pattern.

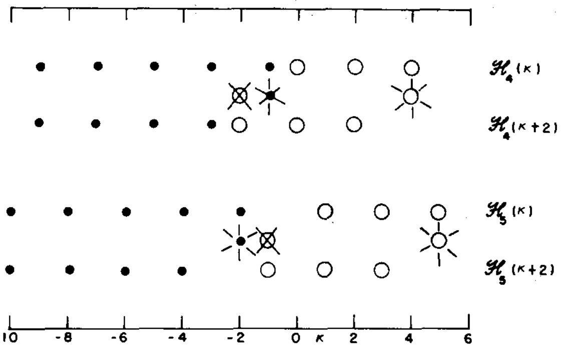


FIG. 2. Hermite quotient polynomial root-pole pattern.

The solution of the *homogeneous* equation $\mathcal{H}(\kappa)$ must satisfy the functional equation

$$\frac{\mathcal{H}(\kappa)}{\mathcal{H}(\kappa+2)}=\frac{(\kappa+1)(\kappa+2)}{2[\kappa-(E/\hbar\omega-\frac{1}{2})]}=\frac{(\kappa+1)(\kappa+2)}{2(\kappa-n)}. \tag{10}$$

where the eigenvalue spectrum, $E_n=(n+\frac{1}{2})\hbar\omega$, has been determined as before by the superposition requirement of integral roots for the denominator. This statement requires that a displacement two units to the right be equivalent to the materialization of a pole at $\kappa=n$ together with a pole suppression and root appearance at $\kappa=-1$, -2 (n odd) or $\kappa=-2$, -1 (n even), respectively.

A new feature is the absence of root destruction, implied by the *linear* rather than quadratic denominator form which characterizes the Legendre family. The lack of suppression is seen in the infinite root pattern of the reciprocal Γ function entering into the improper quotient polynomial, the *Hermite transform* (Fig. 2)

$$\mathcal{H}_n(\kappa)=\frac{C_n/\Gamma((\kappa+n+1)/2-[n/2])}{\Pi_{m=0}^{[n/2]}(\kappa-n+2m)}, \tag{11}$$

which satisfies the functional equation.

By choosing the arbitrary constant $C_n=2^{[n/2]}\sqrt{\pi}n!$, the residues at the poles $n-2\lambda$,

$$\begin{aligned}R_{n-2\lambda}&=\lim_{\kappa\to n-2\lambda}(\kappa-n-2\lambda)\mathcal{H}_n(\kappa)\\&=\frac{\sqrt{\pi}n!/\Gamma(n-[n/2]+\frac{1}{2}-\lambda)}{\Pi_{m=0,\, \neq\lambda}^{[n/2]}(m-\lambda)}\\&=(-)^{\lambda}\frac{2^{n-2\lambda}n!}{\lambda!(n-2\lambda)!}\ (\lambda=0,\ 1,\ldots,\ [n/2]),\end{aligned} \tag{12}$$

are readily identifiable as the coefficients of the Hermite polynomials $H_n(x)$.[3]

LAGUERRE FAMILY

The radial equation for the hydrogenic atom in spherical polar coordinates involves the associated Laguerre equation

$$\rho L''+(2l+2-\rho)L'+\left(\frac{Ze^2}{\hbar(-2E/\mu)^{1/2}}-l-1\right)L=0. \tag{13}$$

By operating with the finite Mellin transform, we obtain, after collecting terms,

$$\begin{aligned}&[\kappa-Ze^2/\hbar(-2E/\mu)^{1/2}+l+1]\,l(\kappa)-(\kappa+1)(\kappa+2l+2)\,l(\kappa+1)\\&=-(\kappa+2l+1)L(1)-L'(1).\end{aligned} \tag{14}$$

The solution of the homogeneous equation $\mathcal{L}(\kappa)$ must satisfy the functional equation

$$\frac{\mathcal{L}(\kappa)}{\mathcal{L}(\kappa+1)}=\frac{(\kappa+1)(\kappa+2l+2)}{\kappa-[Ze^2/\hbar(-2E/\mu)^{1/2}-l-1]}=\frac{(\kappa+1)(\kappa+2l+2)}{\kappa-(n-l-1)}, \tag{15}$$

where the superposition requirement of integral poles establishes the principal quantum number n and energy levels $E_n=-\mu Z^2e^4/(2n^2\hbar^2)$. This equation demands a configuration such that the creation and suppression of poles at $\kappa=n-l-1$ and -1, together with a root appearance at $\kappa=-(2l+2)$ be equivalent to a pattern shift one unit toward higher κ. These stipulations are met by the *associated Laguerre transform*

$$\mathcal{L}_{n+l}^{2l+1}(\kappa)=(-)^{n-l}\frac{C_{nl}/\Gamma(\kappa+2l+2)}{\Pi_{m=0}^{n-l-1}(\kappa-m)}. \tag{16}$$

Like the Hermite, this transform structure is associated through the reciprocal Γ function with an infinite sequence of roots toward negative κ. In contrast, however, to the Legendre and Hermite families in which the roots and poles are separately at even or odd integers, the Laguerre roots and poles are integral and singly spaced [see Fig. 3(a)].

By choosing the constant $C_{nl}=[(n+l)!]^2$, the residues at the poles $\kappa=\lambda$,

$$\begin{aligned}R_\lambda&=\lim_{\kappa\to\lambda}(\kappa-\lambda)\mathcal{L}_{n+l}^{2l+1}(\kappa)\\&=(-)^{n-l}\frac{[(n+l)!]^2/\Gamma(\lambda+2l+2)}{\Pi_{m=0,\,\neq\lambda}^{n-l-1}(\lambda-m)}\quad(\lambda=0,\ 1,\ldots,\ n-l-1),\end{aligned} \tag{17}$$

are identified as the coefficients of the associated Laguerre polynomials $L_{n+l}^{2l+1}(\rho)$.[4]

JACOBI FAMILY

As the final system we consider the Jacobi differential equation which arises in connection with the symmetrical top[5]

$$(x-x^2)J''+[q-(p+1)x]J'+n(n+p)J=0. \tag{18}$$

A transform development analogous to that applied to the other families yields as the solution to the homogeneous transform equation

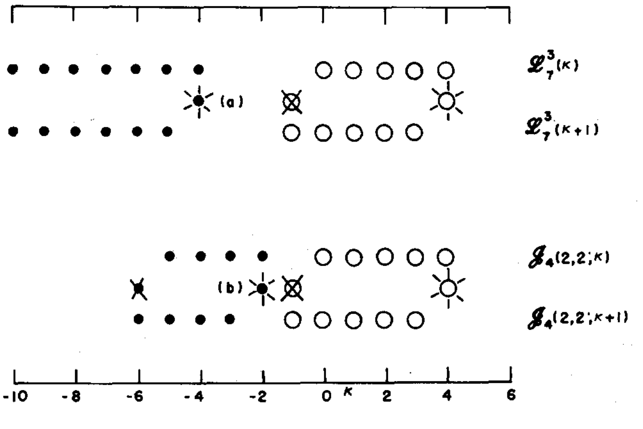

$\mathcal{L}_7^3(\kappa)$
(a)
$\mathcal{L}_7^3(\kappa+1)$
$\mathcal{J}_4(2,2;\kappa)$
(b)
$\mathcal{J}_4(2,2;\kappa+1)$
-10
-8
-6
-4
-2
0
κ
2
4
6

FIG. 3. Quotient polynomial root-pole patterns. (a) Laguerre codon; (b) Jacobi codon.

TABLE I. Root-pole structure of Schrödinger polynomial families.

Spacing	Finite no. of roots	Infinite no. of roots
Double	Legendre	Hermite
Single	Jacobi	Laguerre

$$\mathcal{J}_n(p,q;\kappa)=c_{npq}\frac{\Pi_{m=1}^{n+p-q}(\kappa+p+n-m)}{\Pi_{m=0}^{n}(\kappa-n+m)}\,. \tag{19}$$

By appropriate choice of the constant c_{npq}, the residues at the poles $n-\lambda$,

$$R_{n-\lambda}=(-)^{\lambda}\binom{n}{\lambda}\frac{\Gamma(q+n)}{\Gamma(p+2n)}\,\frac{\Gamma(p+2n+\lambda)}{\Gamma(q+n-\lambda)}\quad(\lambda=0,\ 1,\ldots,n), \tag{20}$$

identify with the coefficients of the Jacobi polynomials $J_n(p,q;x)$.[6]

Like the Legendre, the Jacobi transform has a finite number of roots [Fig. 3(b)]. On the other hand, the Jacobi exhibits an *integral*, rather than the even—odd pattern of the Lagendre roots and poles. It would appear that the Jacobi polynomial completes the quartet of finite pole structures associated with finite orthogonal polynomial solutions of the Schrödinger equation, with all four combinations of finite or infinite root patterns and odd—even or integral root-pole structure represented in Table I.

The disposition of quotient polynomial roots and poles in transform space thus specifies a unique, space-invariant configuration representing each electron state in simple quantum systems. These highly ordered codons provide all the information—in readily retrievable geometric form—which is contained within the physically more opaque Schrödinger ψ functions. The eigenvalue spectrum, for example, follows from the superposition requirement of integrally valued roots and poles. The eigenfunction coefficients are identified with pole strengths and are thus easily calculable from the root-pole patterns. Since the residue of a pole is directly proportional to the product of the separations of the pole from the roots and inversely proportional to the product of the separations from other poles, the Schrödinger coefficients can be understood in terms of a universal *attractive* interaction between root and pole and a *repulsive* interaction between two poles.

Finally, we see that the wave equation is the inverse transform statement in Schrödinger configuration space of the requirement that the root-pole quantum state configuration be invariant under integral displacement. We may consequently view the Schrödinger equation then as following from a conservation law of electron morphology in a discrete Mellin transform space.

[1] $[n/2]$ denotes the largest integer less than or equal to $n/2$.

[2] E. Jahnke, F. Emde, and F. Lösch, *Tables of Higher Functions* (McGraw-Hill, New York, 1960), 6th ed., p. 115.

[3] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I.A. Stegun, NBS Applied Mathematics Series 55 (U.S. Govt. Printing Office, Washington, 1964), par. 22.3.10.

[4] L. Pauling and E.B. Wilson, *Introduction to Quantum Mechanics* (McGraw-Hill, New York, 1935), p. 131.

[5] H. Margenau and G.M. Murphy, *Mathematics of Physics and Chemistry* (Van Nostrand, Princeton, N.J., 1956), 2nd ed., pp. 368—71.

[6] Reference 3, par. 22.3.3.

Coherent pulse propagation, a dispersive, irreversible phenomenon

Mark J. Ablowitz, David J. Kaup, and Alan C. Newell*

Clarkson College of Technology, Potsdam, New York

(Received 12 April 1974)

The initial value problem for the propagation of a pulse through a resonant two-level optical medium is solved by the inverse scattering method. In general, an incident pulse decomposes not only into a special class of pulses to which the medium is transparent but also yields radiation which is absorbed by the medium. In this respect "this problem" has properties markedly different from other dispersive and reversible wave phenomena some of which are tractable by the inverse scattering method. Indeed, it is remarkable that in the present case the method still applies. In particular, we show that, while there are an infinite number of local conservation laws, the integrated densities, and in particular the energy, are only conserved for a very special class of initial conditions. The theoretical results obtained are in close agreement with all the qualitative features observed in the experiments on coherent pulse propagation. Finally, we also show that causality is preserved. Two new and novel features are introduced and briefly discussed. First, we show that if the homogeneous broadening effect is a function of position in the medium, the pulses may speed up and slow down accordingly, without losing their permanent identities. Second, we have found a new kind of solution mode corresponding to a proper eigenvalue of the scattering problem which is not a bound state.

I. INTRODUCTION

Self-induced transparency (SIT), † the effect of a coherent medium response (acting as an attenuator) to an incident electric field, was first discovered by McCall and Hahn.[1,2] More recently, G. Lamb, *et al.*[3-6] have been able to obtain a whole class of special solutions by the inverse scattering method. By assuming both that the eigenvalues of the appropriate scattering problem remain invariant and that there is no continuous spectrum, permanent localized solutions (analogous to the solitons of the Korteweg—deVries equation; in the context of the sine-Gordon equation they have been termed kinks and breathers; colleagues in nonlinear optics refer to them as 2π and 0π pulses) of the relevant Maxwell—Bloch equations are obtained. Propagation heights and speeds are approximated by using the conservation equations.

In short, Lamb has treated only the case of an incident pulse to which the medium is totally transparent and which undergoes pure lossless propagation. In this situation, the incident pulse decomposes only into a sequence of "solitons" which interact with the medium in a very special way so that no net energy is exchanged. In general, however, only a certain portion of the incident pulse forms these special solitons to which the medium is transparent. The rest of the energy, which is mathematically characterized as the continuous spectrum of the appropriate eigenvalue problem (to be introduced in succeeding paragraphs), is "radiation" and is eventually transferred irreversibly to the medium leaving the portion of the medium in which the decomposition of the incident pulse occurs in a permanently excited state. (The eventual decay of these excited atoms through spontaneous emission occurs over a longer time scale and is not incorporated in this mathematical model.)

In this paper we present the procedure for solving the general initial value problem by the inverse scattering technique. We follow closely the ideas laid out in our recent articles.[7,8] Significantly, it is found that many of the aspects of SIT are remarkably different from all of the nonlinear evolution equations solved previously by this method. Most particularly, while there is a sequence of *local* conservation laws,

$$\frac{\partial T_n}{\partial \tau} + \frac{\partial f_n}{\partial x} = 0, \tag{1}$$

the integrated desities $\int f_n d\tau$, including the positive definite norm corresponding to energy, *are not necessarily conserved*. This is a consequence of the irreversible losses to the medium. Indeed, for an *arbitrary* incident pulse, the total energy of the electromagnetic field is a monotonically decreasing function of time, decaying to a constant that depends on the number and amplitudes of the permanent localized pulses which emerge from the decomposition of the incident pulse.

Simply stated, SIT has properties in common with known dispersive and reversible wave phenomena, and still others which are essentially irreversible. By irreversible we mean that for any particular initial condition, energy is transferred to the medium. This results in a population inversion which, due to dephasing effects, is exponentially decaying in the direction of propagation. Thus, integration in the reverse direction would be accompanied by exponential growth. [This is not to say that a sequential pulse in the same direction cannot synchronize (rephase) the system and lead to a coherent photon echo, an effect discussed by Hahn[9] and Abella, Kurnit, and Hartman[10]]. Only if the continuous spectrum is absent, is the problem purely dispersive and reversible. It is indeed remarkable, then, that when the irreversible effects are included, the inverse scattering method can still be applied.

In Sec. II, we give the eigenvalue problem, derive the evolution equations for the scattering data of this eigenvalue problem, explicitly solve them, and also give the equations necessary for solving the inverse problem. In Sec. III, we first give a brief review of the typical results obtained by the inverse scattering method. Then we compare and contrast the solutions from SIT with the typical case, and discuss the agreement of these solutions with what is experimentally known about ultrashort coherent pulse propagation. Finally, in Sec. IV, we discuss the unique feature of SIT wherein the "transmission coefficient" is *not* time invariant, and its implication for the conservation laws. Also, by using the evolution equations for the scattering data, a closed form solution for the "conserved" quantities can be obtained.

II. EVOLUTION OF THE SCATTERING DATA

We consider the following initial value problem. An incident electromagnetic wavetrain within the confines of a spatially modulated envelope impinges on a medium at $x=0$. Measuring time τ in a frame moving with the phase speed of the incident wave pulse, the SIT equations (following Ref. 6) are, in nondimensional form,

$$\epsilon_x = \langle\lambda\rangle, \tag{2}$$

$$\lambda_\tau + 2i\alpha\lambda = \epsilon N, \tag{3a}$$

$$N_\tau = -\tfrac{1}{2}(\epsilon^*\lambda + \epsilon\lambda^*). \tag{3b}$$

Here ϵ is the complex electric field envelope, λ is the out-of-phase and in-phase components of the induced polarization (also complex), N is the normalized population inversion, and $\langle\lambda\rangle \equiv \int_{-\infty}^{\infty} g(\alpha)\lambda(\alpha,x,\tau)d\alpha$, where $g(\alpha)$ characterizes the inhomogeneous broadening of the medium and is normalized to unit area. The initial conditions are the values of $\epsilon(x=0,\tau)$ (which is assumed to decay sufficiently rapidly as $\tau\to\pm\infty$), $\lambda(\tau\to-\infty)\to 0$, and $N(\tau\to-\infty)\to -1$. We remark that given $\epsilon(x=0,\tau)$, only one set of boundary conditions $(\tau\to-\infty)$ can be prescribed for the "Bloch equations" (3).

Following Ref. 6, consider the eigenvalue problem

$$v_{1\tau} + i\zeta v_1 = \tfrac{1}{2}\epsilon v_2, \tag{4a}$$

$$v_{2\tau} - i\zeta v_2 = -\tfrac{1}{2}\epsilon^* v_1, \tag{4b}$$

on the interval $-\infty<\tau<\infty$ (subscripts in τ and x denote partial differentiation). Using the ideas in Refs. 7 and 8, we now show how the x dependencies of v_1 and v_2,

$$v_{1x} = A(\zeta,x,\tau)v_1 + B(\zeta,x,\tau)v_2, \tag{5a}$$

$$v_{2x} = C(\zeta,x,\tau)v_1 - A(\zeta,x,\tau)v_2, \tag{5b}$$

can be used to construct $\epsilon(x,\tau)$ with the above initial and boundary conditions.

Equations (4), (5) require the integrability conditions

$$A_\tau = \tfrac{1}{2}\epsilon C + \tfrac{1}{2}\epsilon^* B, \tag{6a}$$

$$B_\tau + 2i\zeta B = \tfrac{1}{2}\epsilon_x - A\epsilon, \tag{6b}$$

$$C_\tau - 2i\zeta C = -\tfrac{1}{2}\epsilon_x^* - A\epsilon^*, \tag{6c}$$

which ensure that the eigenvalue ζ is independent of x. With ζ real, it is straightforward to show that the choices

$$A(\zeta,x,\tau) = \frac{i}{4}\left\langle\frac{N}{\zeta-\alpha}\right\rangle = \frac{i}{4}P\int_{-\infty}^{\infty}\frac{N(\alpha,x,\tau)g(\alpha)}{\zeta-\alpha}\,d\alpha, \tag{7a}$$

$$B(\zeta,x,\tau) = -\frac{i}{4}\left\langle\frac{\lambda}{\zeta-\alpha}\right\rangle = -\frac{i}{4}P\int_{-\infty}^{\infty}\frac{\lambda(\alpha,x,\tau)g(\alpha)}{\zeta-\alpha}\,d\alpha, \tag{7b}$$

$$C(\zeta,x,\tau) = -\frac{i}{4}\left\langle\frac{\lambda^*}{\zeta-\alpha}\right\rangle = -\frac{i}{4}P\int_{-\infty}^{\infty}\frac{\lambda^*(\alpha,x,\tau)g(\alpha)}{\zeta-\alpha}\,d\alpha, \tag{7c}$$

where $P\int_{-\infty}^{\infty}$ denotes the Cauchy principal value integral, satisfy (6) because of (2) and (3). [As might be expected, the consistent choice of principal value or indenting the contour under (over) the singularity $\alpha=\zeta$ leads to the same final results.] The unique features of this problem are manifested in the mathematical fact that, as $\tau\to+\infty$, A,B,C need not be equal to their respective values as $\tau\to-\infty$ (unlike all other nonlinear evolution equations previously solved by the inverse scattering method). These results are simply seen by noting that (3), given ϵ at any x, constitute ordinary linear (in λ, N) differential equations in τ, the solutions of which are uniquely determined by the conditions $N(\tau\to-\infty)\to-1$ and $\lambda(\tau\to-\infty)\to 0$. Naturally, N and λ do not, in general, take on these values as $\tau\to+\infty$.

Indeed, the quantities N, λ, and λ^*, as shown by Lamb,[6] are related to the fundamental solutions of (4). We define

$$\phi = \begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix}, \quad \bar\phi = \begin{bmatrix}\bar\phi_1\\ \bar\phi_2\end{bmatrix}$$

to be independent solutions of (2), which satisfy the boundary conditions

$$\left.\begin{aligned}\bar\phi &\to \begin{bmatrix}1\\0\end{bmatrix}\exp(-i\zeta\tau) \qquad (8a)\\ \bar\phi &\to \begin{bmatrix}0\\-1\end{bmatrix}\exp(i\zeta\tau) \qquad (8b)\end{aligned}\right\}\text{as } \tau\to-\infty \text{ for all } x.$$

Then we can identify N and λ with ϕ and $\bar\phi$ as follows:

$$N = [\phi_1(\zeta,x,\tau)\bar\phi_2(\zeta,x,\tau) + \bar\phi_1(\zeta,x,\tau)\phi_2(\zeta,x,\tau)]\,\big|_{\zeta=\alpha}, \tag{9a}$$

$$\lambda = 2\phi_1(\zeta,x,\tau)\bar\phi_1(\zeta,x,\tau)\,\big|_{\zeta=\alpha}. \tag{9b}$$

Note that (9) gives N and λ in terms of ϕ and $\bar\phi$ at $\zeta=\alpha$. When ζ is real.

$$\bar\phi = \begin{bmatrix}\phi_2^*\\ -\phi_1^*\end{bmatrix}, \tag{10}$$

and it is the second independent solution of (4) with the above boundary condition (8b). In accordance with the usual scattering procedure[11] let, as $\tau\to+\infty$,

$$\phi \to \begin{bmatrix}a(\zeta,x)\exp(-i\zeta\tau)\\ b(\zeta,x)\exp(i\zeta\tau)\end{bmatrix}, \tag{11a}$$

$$\bar\phi \to \begin{bmatrix}\bar b(\zeta,x)\exp(-i\zeta\tau)\\ -\bar a(\zeta,x)\exp(i\zeta\tau)\end{bmatrix}, \tag{11b}$$

where for ζ real, $a\bar a + b\bar b = 1$, $\bar a = a^*$, $\bar b = b^*$. By using these results, as $\tau\to+\infty$, N can be concisely written as,

$$N(\alpha,x,\tau\to+\infty)\to -1 + 2bb^*(\alpha,x), \tag{12}$$

and

$$\lambda(\alpha,x,\tau\to+\infty)\to 2ab^*(\alpha,x)\exp(-2i\alpha\tau). \tag{13}$$

Notice if $b\equiv 0$ (no continuous spectrum), then $N(\tau\to+\infty)\to -1$, and $\lambda(\tau\to+\infty)\to 0$. (12) and (13) indicate that in general the medium is left in an excited state. The τ dependency of the polarization (λ) is a reflection of the fact that the oscillators return to their natural frequency; 2α is a measure of the difference between the carrier wave frequency of the incident pulse and the natural frequency corresponding to the difference in energy levels of the broadened two level medium.

Since it is the quantities $\phi\exp(A_{-}x)$ and $\bar\phi\exp(-A_{-}x)$ which satisfy (5), then

$$\phi_x=\begin{bmatrix} A-A_- & B \\ C & -(A+A_-) \end{bmatrix}\phi, \tag{14a}$$

$$\overline{\phi}_x=\begin{bmatrix} A+A_- & B \\ C & -A+A_- \end{bmatrix}\overline{\phi}, \tag{14b}$$

where

$$A_-(\zeta,x)\equiv\lim_{\tau\to-\infty} A(\zeta,x,\tau). \tag{15}$$

Then, in the limit of $\tau\to+\infty$, by using (7), (11), (12), (13), (14) and (15), the evolution equations for $a(\zeta,x)$ and $b(\zeta,x)$ are

$$a_x=\frac{i}{4}\left[\left\langle\frac{-1+2b^*b(\alpha,x)}{\zeta-\alpha}\right\rangle-\left\langle\frac{-1}{\zeta-\alpha}\right\rangle\right] -\frac{i}{4}\left[\lim_{\tau\to+\infty}\left\langle\exp[2i(\zeta-\alpha)]\tau\ \frac{2b^*a(\alpha,x)}{\zeta-\alpha}\right\rangle\right]b, \tag{16a}$$

$$b_x=-\frac{i}{4}\left[\lim_{\tau\to+\infty}\left\langle\exp[2i(\alpha-\zeta)\tau\ \ \frac{2a^*b(\alpha,x)}{\zeta-\alpha}\right\rangle\right]a -\frac{i}{4}\left[\left\langle\frac{-1+2b^*b(\alpha,x)}{\zeta-\alpha}\right\rangle+\left\langle\frac{-1}{\zeta-\alpha}\right\rangle\right]b, \tag{16b}$$

where, for ζ real,

$$\left\langle\frac{(\circ\bullet\circ)}{\zeta-\alpha}\right\rangle\equiv P\int_{-\infty}^{\infty}\frac{(\bullet\bullet\circ)g(\alpha)}{\zeta-\alpha}\,d\alpha.$$

Note that the singular point $\alpha=\zeta$ is removable and therefore any choice for $\langle(\circ\circ\circ)/(\zeta-\alpha)\rangle$, applied consistently, yields the same analytic function.

Using well-known results when ζ is real, we find

$$\lim_{\tau\to\infty}\left\langle\frac{ab^*\exp[2i(\zeta-\alpha)\tau]}{\zeta-\alpha}\right\rangle=i\pi a(\zeta,x)b^*(\zeta,x)g(\zeta), \tag{17a}$$

$$\lim_{\tau\to\infty}\left\langle\frac{a^*b\exp[2i(\alpha-\zeta)\tau]}{\zeta-\alpha}\right\rangle=-i\pi a^*(\zeta,x)b(\zeta,x)g(\zeta). \tag{17b}$$

Thus, (16) reduces to

$$a_x=\frac{i}{2}a\left(\left\langle\frac{bb^*}{\zeta-\alpha}\right\rangle-i\pi bb^*g\right) =\frac{i}{2}a\int_{C_u}\frac{bb^*}{\zeta-\alpha}g(\alpha)\,d\alpha, \tag{18a}$$

$$b_x=\frac{i}{2}b\left(\left\langle\frac{aa^*}{\zeta-\alpha}\right\rangle+i\pi aa^*g\right) =\frac{i}{2}b\int_{C_a}\frac{aa^*}{\zeta-\alpha}g(\alpha)\,d\alpha. \tag{18b}$$

In (15), $\int_{C_u}(\int_{C_a})$ refer to the contours along the real axis indenting under (over) the pole at $\alpha=\zeta$.

To complete the solution of (18), we need the x dependency of aa^* for real ζ $(bb^*=1-aa^*)$. This follows directly from (18a). Defining

$$\mathscr{A}\equiv aa^*, \tag{19}$$

then (18a) gives

$$\mathscr{A}_x=\mathscr{A}(1-\quad)\pi g, \tag{20}$$

or

$$\mathscr{A}(\alpha,x)=\frac{\mathscr{A}_0}{\mathscr{A}_0+(1-\mathscr{A}_0)\exp(-\pi gx)}, \tag{21}$$

where $\mathscr{A}_0=\mathscr{A}(\alpha,0)$. Consequently, the solution of (18) is

$$a(\zeta,x)=a(\zeta,0)\exp[-i\Omega(\zeta,x)], \tag{22a}$$

$$b(\zeta,x)=\frac{b(\zeta,0)\exp(i\Omega)\exp(-\pi gx)}{\mathscr{A}_0+(1-\mathscr{A}_0)\exp(-\pi gx)}\exp\left(i\,\frac{x}{2}\int_{C_u}\frac{gd\alpha}{\zeta-\alpha}\right), \tag{22b}$$

where

$$\Omega(\zeta,x)\equiv\frac{1}{2\pi}\int_{C_u}\frac{d\alpha}{\zeta-\alpha}\ln[\mathcal{A}_0+(1-\mathcal{A}_0)\exp(-\pi gx)]. \tag{23}$$

In order to determine ϵ, N, and λ for $x>0$ via the inverse scattering method, we do not need the general result given by (22) and (23), but only the x dependence of (i) b^*/a for real ζ, (ii) the bound state eigenvalues (ζ_k) in the upper half ζ-plane [which are found from the eigenvalue problem (4) and are the zeros of a], and (iii) $\overline{C}_k$. (When b^* is analytically extendable into the upper half ζ-plane, $\overline{C}_k$ is simply the residue of b^*/a at the eigenvalue $\zeta=\zeta_k$.) First, the x independence of the eigenvalues [assumed by Lamb[6] and required by (6)] can immediately be seen from (22a) and (23). Since, in the upper half ζ-plane, Ω is analytic, the zeros of a do not move (furthermore, new zeros do not appear), and the eigenvalues will therefore remain independent of x. From (22) and (23) [or also from (18)] we have

$$\frac{b^*}{a}(\xi,x)=\frac{b^*}{a}(\xi,0)\exp\left[-\frac{i}{2}x\int_{C_u}\frac{g(\alpha)}{\xi-\alpha}\,d\alpha\right] \tag{24}$$

and

$$\overline{C}_k(x)=\overline{C}_k(0)\exp\left[-\frac{i}{2}x\int_{C_u}\frac{g(\alpha)}{\zeta_k-\alpha}\,d\alpha\right]. \tag{25}$$

To complete the solution, one continues as given in Ref. 11. First, solve the eigenvalue problem (4) for the bound state eigenvalues (ζ_k), and $\overline{C_k}$, and also for b^*/a $(\zeta=\text{real})$, all at $x=0$. Then, using (24) and (25), construct

$$F(y)=-i\sum_k\overline{C}_k(x)\exp(-i\zeta_k y)+\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{b^*}{a}(\zeta,x) \times\exp(-i\zeta y)\,d\zeta. \tag{26}$$

Solve the inhomogeneous linear integral equation

$$K(\tau,\theta) =F(\tau+\theta)-\int_{-\infty}^{\tau}\int_{-\infty}^{\tau}F(\theta+\beta)F^*(\beta+\gamma)K(\tau,\gamma)\,d\beta\,d\gamma; \tag{27}$$

then ϵ is given by

$$\epsilon(x,\tau)=-4K(\tau,\tau;x). \tag{28}$$

Once K is found, N and λ can also be determined.[11]

In concluding this section, we note an alternative form for (24) is

$$\frac{b^*}{a}(\xi,x)=\frac{b^*}{a}(\xi,0)\exp\left[-\frac{\pi}{2}g(\xi)x-\frac{i}{2}xP\int_{-\infty}^{\infty}\frac{g(\alpha)}{\xi-\alpha}\,d\alpha\right], \tag{29}$$

which explicitly shows that b^*/a decays exponentially as $x\to\infty$ at a rate proportional to the inhomogeneous broadening.

III. GENERAL FEATURES OF THE SOLUTION

Of all the steps required for the application of the inverse scattering method, the most important and crucial step is to be able to solve for the time dependence (in this case, the x-dependence) of the scattering data for arbitrary initial scattering data. Once this is done, everything else follows, allowing one to construct the solution at any later time (in this case, x) from the initial data and to determine the form and structure of the general solution. For those familiar with Hamilton—Jacobi theory, the power of the inverse scattering method can best be appreciated as follows: The inverse scattering method is simply a canonical transformation which yields the Hamilton—Jacobi functional differential equation completely separable. Naturally once separation has been achieved, the solution for the resulting "action-angle" variables is trivial. Although complete separation is not achieved in the case of SIT, still the separation is sufficient to allow a solution to be found, as we have just seen. For the rest of this section, we want to give a review of the typical forms and features of solutions obtained via the inverse scattering method, discuss the analogies and distinct differences of the solutions for SIT compared to other inverse scattering solutions, and show the remarkable qualitative agreement between these solutions with what is known experimentally about ultrashort coherent pulse propagation.

Throughout all applications of the inverse scattering method,[7,8,11-13] there are two distinct features of the general solution which have remained invariant. The first is the concept of the "soliton,"[14] which is a stable, localized, permanent waveform which evolves in time by a simple translation. The second is the concept of "radiation" which is not in general localized, does not have a permanent shape, and in general does decay algebraically in time. Any general solution of the evolution equations will always contain a mixture of these two fundamental solutions, and in general, it is impossible to separate (by inspection) a general solution into these two fundamental modes since the mixing is nonlinear. However, when a general solution is "mapped" by the direct scattering problem (which is a nonlinear mapping) into the scattering data, these fundamental modes are then separated. [This is simply a generalization of the well-known technique for solving linear evolution equations by Fourier transformations, whereby one "maps" a function into its Fourier transform. In this case, the evolution equation for the Fourier transform is also separable. One should also note that Eq. (26) is in effect a Fourier transform!] In terms of the scattering data, each "soliton" corresponds to exactly one bound state of the eigenvalue problem and vise versa, while the "radiation" corresponds to the continuous spectra of the eigenvalue problem. These modes are easily seen from the form of F [Eq. (26)]. In (26), each soliton is specified by giving ζ_k and $\overline{C}_k$, where ζ_k gives the velocities of the soliton, while $\overline{C}_k$ essentially specifies the initial position and phase of the soliton. Consequently, the number of solitons is exactly equal to the number of bound states. For the radiation mode, this is represented in (26) by the integral along the real axis over the continuous spectrum, and is specified by giving (b^*/a).

The simplest solution to find via the inverse scattering method is the solution for a single soliton with no radiation present. In this case, when we set $b(\zeta)\equiv 0$, the kernel of (26) becomes completely separable, allowing an explicit solution. Inserting the x-dependence given by (25), then, from (26), (27), and (28), we find

$$\epsilon(x,\tau)=4\eta\exp(-i\phi)\operatorname{sech}\theta, \tag{30}$$

where

$$\zeta_1=\xi+i\eta, \tag{31a}$$

$$\overline{C_1}=-2i\eta\exp(-\theta_0)\exp(+i\phi_0), \tag{31b}$$

$$\theta=\theta_0+\omega_2x-2\eta\tau, \tag{32a}$$

$$\phi=\phi_0+\omega_1x-2\xi\tau, \tag{32b}$$

$$\omega_1+i\omega_2=-\frac{1}{2}\int_{-\infty}^{\infty}\frac{g(\alpha)\,d\alpha}{\zeta_1-\alpha}. \tag{33}$$

To relate these variables to physical quantities, we first note that x and τ are not the usual space—time coordinates. Letting the usual space—time coordinates be X and T, then when $c=1$ ($c=$speed of light)

$$x=X, \tag{34a}$$

$$\tau=T-X. \tag{34b}$$

Thus by (30), (32a), and (34) this soliton has a velocity of

$$v=(1+\omega_2/2\eta)^{-1}, \tag{35}$$

which is less than unity. Before proceeding further, it becomes necessary to choose a model for the inhomogeneous broadening term in (33). A physical and simple model is the Lorentzian line shape

$$g(\alpha)=\frac{1}{\pi}\frac{\Gamma}{\alpha^2+\Gamma^2} \tag{36}$$

where 2Γ is the width at half-height. From (33) and (36) we have

$$\omega_1=-\frac{1}{2}\frac{\xi}{\xi^2+(\eta+\Gamma)^2}, \tag{37a}$$

$$\omega_2=+\frac{1}{2}\frac{\eta+\Gamma}{\xi^2+(\eta+\Gamma)^2}, \tag{37b}$$

and consequently, when $\Gamma\ll\eta$, the velocity is essentially dependent only on the magnitude of ζ_1. On the other hand, the width of the soliton (in time, T) is inversely proportional to the imaginary part of the eigenvalue η, while the amplitude is proportional to η. This is a well-known result for nonlinear waveforms, in that the height, width, and velocity are interrelated.

In addition to single soliton solutions, multiple soliton solutions can also be explicitly given.[11,13] A special type of a multiple soliton solution occurs when more than one soliton have the same velocity, and these have been called "multi-soliton bound states." These solutions in general have a very complicated and oscillating waveform. A simple example of a multisoliton bound state for the SIT equations is the analogy of the "breather" (called a 0π pulse by Lamb) solution of the sine-Gordon equation.[7] In the special case where Reλ, N, and g are even in α, Imλ is odd in α, and $\epsilon(0,\tau)$ is real, then one can show that $\epsilon(x,\tau)$ remains real for all x and the discrete eigenvalues must occur either on the imaginary axis $\xi=0$, whence we have either a simple soliton (kink)

$$\epsilon(x,\tau)=4\eta\,\mathrm{sech}\,\theta \tag{38}$$

or in complex conjugate pairs (ζ and $-\zeta^*$) whence we have a breather

$$\epsilon(x,\tau)=8\,\frac{\eta}{\xi}\,\frac{\xi\cosh\theta\sin\phi+\eta\sinh\theta\cos\phi}{\cosh^2\theta+\eta^2/\xi^2\cos^2\phi}$$

$$=4\,\frac{\partial}{\partial\tau}\tan^{-1}\left(\frac{\eta}{\xi}\quad\frac{\cos\phi}{\cosh\theta}\right), \tag{39}$$

with θ and ϕ as given by (32).

Usually, the computation of solutions when the radiation (the continuous spectrum) is present is very difficult.[15] The exact manner in which this part of the spectrum evolves in time depends on the specific problem being solved, but one can still make some general statements concerning it. This fundamental mode of the solution is invariably characterized by a series of oscillations which propagate away from the initial disturbance (whence the name "radiation"). In all other cases (except SIT), these oscillations decay only algebraically in time, usually approaching some special decaying nonlinear oscillating state. Consequently, all of these systems evolve toward a general final state consisting of free solitons, multisoliton bound states, and decaying radiation, with the soliton states eventually ordering themselves according to their velocities.

Much more could be said about the inverse scattering solutions, but it is now perhaps best to refer the reader to the literature in this area,[11-15] and instead go on to discuss some of the specifics of the solutions for SIT.

Many of the features of SIT are very similar to the general case discussed above in that we have these two fundamental modes consisting of solitons and radiation. However, SIT is distinctly different from all other previous systems solved by the inverse scattering method in that the x dependence of the continuous spectrum [Eq. (29)] is not simply oscillatory, but is damped! This has the physical consequence that the medium will act as a "filter," and will only allow the discrete spectrum (the solitons) to be propagated through. Of course, this is exactly what is observed experimentally. To see what has happened to the continuous spectrum, let us consider an arbitrary initial pulse incident on a medium at $x=0$. Knowing the shape of the initial pulse, we can solve the eigenvalue problem (4) for the bound state parameters (ζ_k, $\overline{C_k}$, $k=1,2,\ldots,N$), the "transmission coefficient", a, and the "reflection coefficient", b, for real ζ. Let us now look at N and λ in the limit of $\tau\to+\infty$, which corresponds to the respective values after the initial pulse has passed. Directly from (12), (13), (22), and (23), we find

$$N(\alpha,x)\to-1+\frac{2(1+N_0)\exp(-\pi gx)}{1-N_0+(1+N_0)\exp(-\pi gx)}, \tag{40}$$

$$\lambda(\alpha,x,\tau)\to\frac{2\lambda(\alpha,0,\tau)\exp(-\pi gx/2)\exp(-i\chi)}{1-N_0+(1+N_0)\exp(-\pi gx)}, \tag{41}$$

where N_0 is N at $x=0$ as $\tau\to+\infty$ and χ is a real phase given by

$$\chi=\Omega+\Omega^*+\frac{x}{2}\,P\int_{-\infty}^{\infty}\frac{g(\alpha)\,d\alpha}{\zeta-\alpha}, \tag{42}$$

with Ω given by (23). Equations (40) and (41) exhibit two more well-known but related phenomena: the excitation of the medium and its consequent "ringing" after the initial pulse has passed.[1,2] Since (40) shows that, in general, $N+1$ is not zero as $\tau\to+\infty$, a certain fraction of the atoms remain excited after the initial pulse has passed. In order to do this, energy must be extracted from the initial pulse, and it is then shared coherently between the atoms, causing the ringing as given by (41). Since the solitons will eventually be propagated through, they cannot lose energy, so that the energy must come from the continuous spectrum. Further, the absorption of the continuous spectrum continues until it becomes exponentially small as $x\to\infty$, with both $N+1$ and $\lambda\to0$ in this limit.

Inspection of (41) reveals a very interesting feature of the ringing. For certain initial pulse profiles, the maximum amplitude of the ringing will not occur at $x=0$, but can occur well inside the medium, at $x=x_r(\alpha)$, given by

$$x_r(\alpha)=\frac{1}{\pi g(\alpha)}\ln\left(\frac{1+N_0(\alpha)}{1-N_0(\alpha)}\right). \tag{43}$$

Naturally, to be physical, x_r must be greater than zero, requiring $N_0(\alpha)>0$, and if $N_0(\alpha)<0$, then the maximum in the physical region occurs at $x=0$. Of course, this is not totally unexpected since as a consequence of (3) and the boundary conditions, we have

$$N^2+\lambda^*\lambda=1, \tag{44}$$

showing that $|\lambda|$ is a maximum when $N=0$. Thus, if the initial pulse gives $N>0$ for a range of α, due to the following absorption of the continuous spectrum, N will monotonically decrease in x, giving the maximum in λ at some $x>0$. What is new about (43) is by solving for the complete x dependence of the scattering data we have an explicit expression for x_r.

Of course, the rate at which the continuous spectrum is absorbed depends only on the inhomogeneous broadening factor $g(\alpha)$. Since g is normalized to have a unit area, the effective absorption rate depends mostly on the width of the level and the width and centering of the incident pulse. If the central frequency of the incident pulse is centered on the resonant frequency and if its width is smaller than the level width, then a maximum filtering effect is achieved. For the model (36), the decay length in this case for the continuous spectrum [see Eq. (29)] is simply $\frac{1}{2}\pi\Gamma$. When the central frequency of the incident pulse is not centered on the resonant frequency by a significant amount, then, in terms of (36), the decay length increases significantly to α^2/Γ, giving inefficient filtering.

In concluding this section, we want to look at the form of the solution as $x\to\infty$, and will direct our attention to the function F in (26). In this limit, the contribution of the radiation term to F becomes exponentially small while the soliton contribution becomes exponentially large, forcing F to approach the form for pure solitons (i.e., no radiation). If one now neglects the radiation contribution, a closed form solution for $\epsilon(x,\tau)$ is possible.[11] As is well known, as $x \to \infty$, this solution approaches a linear sum of the simple soliton solutions, (30), and multisoliton bound states. This illustrates another well-known property of ultrashort coherent pulse propagation called "pulse-reshaping" whereby the incident pulse is "reshaped" into those pulses capable of undergoing lossless propagation (solitons).

Let us now return and consider the radiation contribution to F in this limit. If one uses the method of steepest descent, one finds that the radiation contribution to F does vanish exponentially everywhere, except near the light cone ($\tau = 0$). Here, when b^*/a approaches zero only algebraically as $|\zeta| \to \infty$, the radiation field is merely a small "blip." Otherwise, it gives no contribution.

Now, let the initial conditions be such that $\epsilon = 0$ if $\tau < 0$. Then since (2) is causal, ϵ must remain zero for all x when $\tau < 0$. The radiation contribution to F guarantees this, because if $\epsilon = 0$ when $\tau < 0$ at $x = 0$, one can show that b^*/a is analytically extenable into the upper half ζ-plane and that $\overline{C_k}$ is then simply the residue of b^*/a at $\zeta = \zeta_k$. Then, by contour integration, one can show that F, and hence $\epsilon(x,\tau)$, are identically zero for all x when $\tau < 0$. In other words, in this case the radiation field is necessary to ensure that the forward tail of the leading soliton does not extend beyond the light cone.

Finally, we point out that the pulse heights and shapes are dependent on the medium parameters, but not on the inhomogeneous broadening $g(\alpha)$. The pulse speeds do depend on this factor. But, returning to the derivations in Sec. II, one sees that, without loss of generality, we could allow $g(\alpha)$ to be also a function of x and still obtain the x dependence of the scattering data. In this case, the solitons would still retain the same heights and shape while changing their velocities as they propagate.

IV. MATHEMATICAL ASPECTS OF SIT

In all other previous examples using the inverse scattering method to solve nonlinear evolution equations, the x dependency of the scattering data was always given by

$$a_x = 0, \quad b_x = -2A_0(\zeta)\, b,$$

where $A_0(\zeta) = A_-(\zeta) = A_+(\zeta)$ and was independent of x. The simplicity of these expressions, and in particular the x invariance of a (the "transmission coefficient"), was related to the existence of globally conserved quantities. (For a further and fuller discussion see Ref. 12.) The present problem has this property only when the incident pulse is so special as to decompose into only kinks and breathers with no "radiation", i.e., $b(\zeta) \equiv 0$. The fact that the initial value problem is still tractable when Eq. (16) are fairly complicated leads us to conjecture that the inverse method may be applicable to a wider class of problems than heretofore believed.

In general, $b(\zeta) \neq 0$, and although one still has *local* conservation laws, the *global* quantities are not conserved. As examples, the first two local conservation laws are given by

$$f_1 = \tfrac{1}{2}\epsilon^*\epsilon, \tag{45a}$$

$$f_2 = \tfrac{1}{4}(\epsilon^*\epsilon)^2 - \epsilon_\tau^*\epsilon_\tau, \tag{45b}$$

$$T_1 = \langle 1 + N\rangle, \tag{46a}$$

$$T_2 = \epsilon^*\epsilon\langle N\rangle + 2i\epsilon\langle\alpha\lambda^*\rangle - 2i\epsilon^*\langle\alpha\lambda\rangle - 8\langle\alpha^2(N+1)\rangle, \tag{46b}$$

where (45) and (46) satisfy (1). Defining

$$F_n \equiv \int_{-\infty}^{\infty} f_n\, d\tau, \tag{47}$$

then from (1) we have

$$\frac{dF_n}{dx} = T_n \Big|_{\tau=-\infty}^{+\infty} \tag{48}$$

As shown by Schnack and Lamb,[16] when ϵ vanishes sufficiently rapid as $\tau \to +\infty$, (48) becomes

$$\frac{dF_n}{dx} = K_n\langle\alpha^{2n-2}(N+1)\rangle \Big|_{\tau=-\infty}^{+\infty}, \tag{49}$$

where K_n is a set of numerical coefficients. By using (40), (49) is integrable, and this gives

$$F_n(x) = F_n(0) - \frac{2K_n}{\pi}\int_{-\infty}^{\infty} d\alpha\, \alpha^{2n-2} \times \ln\left(\frac{1-N_0}{2} + \frac{1+N_0}{2}\exp(-\pi g x)\right). \tag{50}$$

We note that, as $x \to \infty$, F_n becomes *independent* of the inhomogeneous broadening factor, a result which is contrary to that suggested by Ref. 16. Still, one can use the conservation laws in certain cases to obtain reasonable values for the eigenvalues, although one can easily devise many examples where this technique will fail. For example, let $\epsilon(\tau,0)$ be zero if $\tau < 0$ or $\tau > \tau_1$, and a constant value of ϵ_0 between these limits. Then, from (4), it is easy to show that $b(\alpha) \to O(1/\alpha)$ as $|\alpha| \to \infty$, and, by (12), $N_0 \to -1 + O(1/\alpha)$. Then inspection of (43) shows that $F_n(x \to +\infty)$ is undefined if $n \geq 2$. Thus, in this example, one has only *one* conservation law which can be used, and if the initial profile contains more than one soliton, a unique determination of the eigenvalues is impossible.

In any case, whenever $|\epsilon| \to 0$ faster than $|\tau|^{-1}$ as $\tau \to \pm\infty$, one can *always* determine the eigenvalues by simply solving the eigenvalue problem, Eq. (4). Even in the most complicated cases, numerical determination of the eigenvalues is quite practical with present high speed computers.

Finally, one interesting feature of the eigenvalue problem (4) is the possibility of having $a = 0$ ("bound states") on the real axis[12]! For the KdV equation,[13] bound states on the real axis are strictly forbidden, but are allowed by (4) as can be shown easily by specific examples. One can now ask whether or not these modes give anything new for SIT. First, if $a = 0$ on the real axis, F as given by (26) has a pole in the integral on the real axis. If one retraces the derivation of (27), one finds that this integral is to be replaced by the Cauchy principle value plus $(-i)\cdot\mathrm{Res}[(b^*/a)\exp(-i\zeta y)]$ at the pole (i.e., when b^* is sufficiently analytic to be extended a certain amount into the upper half ζ-plane, F is always a contour integral above all zeros of a). Taking the limit of large x and using the method of steepest descent, one finds that the contribution to F from a zero on the real axis vanishes exponentially in

x like the radiation does. Meanwhile, the τ dependence of F is in between that of a soliton and radiation, since for small τ it gives zero and for large τ it simply oscillates like $\exp(-i\xi_0\tau)$, where ξ_0 is the zero of a. (Solitons grow exponentially in τ while the radiation decays algebraically.) Due to this x and τ dependence, a zero on the real axis corresponds more to a particular form of radiation than to a soliton. From (35) and (37) we see that if we did consider it to be a soliton, it would have a zero velocity; consequently it will never "detach" itself from the radiation, in agreement with the x and τ dependence of F.

Finally, for a zero of a on the real axis, we note the form of a and b as $x\rightarrow\infty$. From (22), in this limit, $|a|\rightarrow 1$ and $|b|\rightarrow 0$ exponentially everywhere on the real axis except at the zero of a. Here, $|a|\rightarrow 0$ and $|b|\rightarrow 1$. Consequently, in this limit a and b do not possess a first derivative with respect to ζ, which implies the integral $\int_{-\infty}^{\infty}|\epsilon|(1+|\tau|)\,d\tau$ does not exist as $x\rightarrow\infty$.[12] A zero on the real axis also has a consequence for the ringing, since at $\alpha=\xi_0$, x_r [Eq. (43)] is infinity. However, the width of this ringing about $\alpha=\xi_0$ vanishes exponentially in x, causing the stored energy to also vanish exponentially.

*Supported in part under NSF Grants Nos. GP-32839X2 and GP-43653.

[†] *Note added in proof*: While the term "Self-induced transparency" literally connotes only lossless propagation, we use the term in the wider context as referring to general coherent pulse propagation.

[1] S. L. McCall and E. L. Hahn, Phys. Rev. Lett. **18**, 908 (1967).

[2] S. L. McCall and E. L. Hahn, Phys. Rev. **103**, 183 (1969).

[3] G. L. Lamb, Jr., Phys. Lett. A **25**, 181 (1967).

[4] G. L. Lamb, Jr., Rev. Mod. Phys. **43**, 99 (1971).

[5] G. L. Lamb, Mo. O. Scully, and F. A. Hopf, Appl. Opt. **11**, 2572 (1972).

[6] G. L. Lamb, Phys. Rev. Lett. **31**, 196 (1973).

[7] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Phys. Rev. Lett. **30**, 1262 (1973).

[8] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Phys. Rev. Lett. **31**, 125 (1973).

[9] E. L. Hahn, Phys. Rev. **80**, 580 (1950).

[10] I. D. Abella, N. A. Kurnit, and S. R. Hartmann, Phys. **141**, 391 (1966).

[11] V. E. Zakharov and A. B. Shabat, Zh. Eksp. Teor. Fiz. **61**, 118 (1971) [Sov. Phys. JETP **34**, 62 (1972)].

[12] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, to be published.

[13] C. S. Gardner, J. Greene, M. D. Kruskal, and R. M. Mirua, Phys. Rev. Lett. **19**, 1095 (1967); M. D. Kruskal, R. M. Miura, C. S. Gardner, and N. J. Zabusky, J. Math. Phys. **11**, 952 (1970).

[14] A. C. Scott, F. Y. F. Chu, and D. W. McLaughlin, Proc. IEEE **61**, 1443 (1973).

[15] M. J. Ablowitz and A. C. Newell, J. Math. Phys. **14**, 1277 (1973).

[16] D. D. Schnack and G. L. Lamb, Jr., in *Coherence and Quantum Optics*, edited by L. Mandel and E. Wolf (Plenum, New York, 1973), pp. 23—33.

Factorizability of resonance poles in multiparticle amplitudes

J. Denmead Smith

College of the Resurrection, Mirfield, Yorkshire, England

(Received 25 April 1973)

Using the energy-analytic representation of Green's functions and relying on certain explicitly stated properties of the off-shell scattering elements, it is shown that resonance poles in the S matrix contribute poles to the off-shell scattering amplitude, and the residues there have the same factorizable form as that associated with bound state particles.

1. INTRODUCTION

It is a well-known deduction from the LSZ formulation of quantum field theory that bound state particles produce direct channel poles in the off-shell scattering amplitude, and these poles have factorizable residues. This behavior is usually assumed to hold for unstable particles as well, both in quantum field theory and S-matrix theory. In the latter case, where particles are actually identified with their poles, the factorization of the residues into on-shell vertex functions is a very strong condition and is taken as an essential axiom of the theory. For quantum field theory the condition is even more stringent, since it involves the appearance, in the *off-shell* scattering elements, of poles which depend only on the total energy and have residues which factorize into "in" and "out" *off-shell* wavefunctions.

In this paper we shall not be concerned with the difficult problem of whether quantum fields may be ascribed to unstable particles: all fields that appear will relate to stable particles (having spin 0, for simplicity). Rather, our aim is to demonstrate, within the LSZ framework, that unstable particles which result from resonances of *stable* particles and appear in the S matrix as second-sheet poles, necessarily contribute poles to the *off-shell* scattering elements; these poles depend only on the total energy and have factorizable residues in the sense described above.

Except for the case of 2—2 scattering, the analytic structure of multiparticle S-matrix elements is very imperfectly understood, and present-day knowledge of the analytic structure of off-shell amplitudes is even less complete. Our treatment therefore demands that certain very plausible assumptions are made concerning the analyticity of the off-shell amplitudes. Modulo these assumptions, which have been considered in the literature in connection with Feynman diagrams,[1] the problem is successfully solved.

In the course of the explanation we define an on-shell phase space. This describes asymptotic particle states of a fixed total 4-momentum, and so allows the analytic continuation of the total energy to be considered independently of the "relative" variables; it is also shown that "Hermitian symmetry" of the S matrix holds with this particular definition.

The following notation is used for 4-vectors and the Lorentz product:

$$x \equiv (x_0, \mathbf{x}), \quad xy \equiv x_0 y_0 - \mathbf{x}\cdot\mathbf{y}.$$

2. KINEMATICS

Consider an n-particle asymptotic channel "c", either of the "in" or "out" type, defined by particles with masses $m_1, \cdots, m_n$. For simplicity it will be assumed that these constituent particles are all different. The purpose of this section is to define a parametrization of the physical states in "c" which have the same 4-momentum $E \equiv (E_0, \mathbf{E})$.

If $|\mathbf{p}_1, \ldots, \mathbf{p}_n\rangle$ is such a state in which the particles have 4-momenta $p_j \equiv (p_{j0}, \mathbf{p}_j)$, where $p_{j0} = (\mathbf{p}_j^2 + m_j^2)^{1/2}$, then

$$E_0 = \sum_{j=1}^{n} (\mathbf{p}_j^2 + m_j^2)^{1/2}, \tag{2.1}$$

$$\mathbf{E} = \sum_{j=1}^{n} \mathbf{p}_j. \tag{2.2}$$

The relative 4-momenta $\bar{p}_1, \ldots, \bar{p}_n$ are defined by $\bar{p}_j = \bar{p}_j - E/n$, and we shall denote by $\bar{\mathbf{P}}$ the $3n$-vector $(\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_n)$. Equation (2.2) thus restricts $\bar{\mathbf{P}}$ to the $(3n-3)$-dimensional subspace

$$\mathcal{L} \equiv \sum_{j=1}^{n} \bar{\mathbf{p}}_j = \mathbf{0},$$

while Eq. (2.1) restricts $\bar{\mathbf{P}}$ to the convex hypersurface

$$\mathcal{K} \equiv \sum_{j=1}^{n} \left\{ \left[\left(\frac{\mathbf{E}}{n} \right) + \bar{\mathbf{p}}_j \right]^2 + m_j^2 \right\}^{1/2} = E_0.$$

Now provided E is physical, i.e., $E_0^2 - \mathbf{E}^2 > (m_1 + \ldots + m_n)^2$, a half-ray from the origin $\bar{\mathbf{P}} = \mathbf{0}$ intersects $\mathcal{K}$ in just one point, so that the possible on-shell momenta of fixed total energy can be represented by points of the unit sphere $\mathcal{S}^c$ in $\mathcal{L}$, i.e., the $3n$ vectors $\mathbf{R}^c \equiv (\mathbf{r}_1^c, \ldots, \mathbf{r}_n^c)$ satisfying $\mathbf{R}^c \in \mathcal{L}$, $(\mathbf{R}^c)^2 \equiv (\mathbf{r}_1^c)^2 + \ldots + (\mathbf{r}_n^c)^2 = 1$.

With this representation the physical on-shell momenta are

$$\mathbf{p}_j = \mathbf{E}/n + \bar{\mathbf{p}}_j = \mathbf{E}/n + \lambda \mathbf{r}_j^c,$$

where λ satisfies

$$E_0 = \sum_{j=1}^{n} \omega_j(\mathbf{p}_j) \equiv \sum_{j=1}^{n} \left\{ \left[\left(\frac{\mathbf{E}}{n} \right) + \lambda \mathbf{r}_j^c \right]^2 + m_j^2 \right\}^{1/2}.$$

The transformation from the variables $\mathbf{p}_1, \ldots, \mathbf{p}_n$ to the variables E_0, $\mathbf{E}$, $\mathbf{R}^c$ will enable us to talk about states of fixed 4-momentum, and the Jacobian of the transformation is

$$\frac{\partial(\mathbf{p}_1, \ldots, \mathbf{p}_n)}{\partial(E_0, \mathbf{E}, \mathbf{R}^c)} = \frac{\lambda^{3n-5}}{\sum_{j=1}^{n} (\mathbf{r}_j^c)^2 / \{[(\mathbf{E}/n) + \lambda \mathbf{r}_j^c]^2 + m_j^2\}^{1/2}}.$$

Identifying $|\mathbf{p}_1, \ldots, \mathbf{p}_n\rangle$ with $|E, \mathbf{R}^c\rangle$, it therefore follows that the inner products of basis vectors are given by

$$\langle E_1, \mathbf{R}_1^c | E, \mathbf{R}^c \rangle = \left(\prod_{j=1}^{n} 2\omega_j(\mathbf{p}_j) \right) \frac{\partial(E, \mathbf{R}^c)}{\partial(\mathbf{p}_1, \ldots, \mathbf{p}_n)} \delta(\mathbf{R}_1^c, \mathbf{R}^c)\, \delta^4(E - E_1),$$

where $\delta(\mathbf{R}_1^c, \mathbf{R}^c)$ is the δ function on $\mathcal{S}^c$.

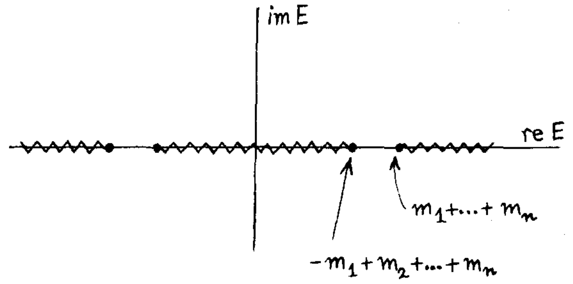


FIG. 1. Analytic structure of the function $\lambda(E)$.

It will be sufficient (and more convenient) to use a center of mass frame where, with a slightly ambiguous notation we write $E \equiv (E, 0)$. Apart from a trivial rotation in 3-space, we then have a *Lorentz-invariant* description of the states of fixed energy. Omitting the momentum-preserving δ function,

$$\langle E_1, \mathbf{R}^c | E, \mathbf{R}^c\rangle = \left(\prod_{j=1}^{n} 2[\lambda^2(\mathbf{r}_j^c)^2 + m_j^2]^{1/2}\right) \times \left(\sum_{j=1}^{n} \frac{(\mathbf{r}_j^c)^2}{[\lambda^2(\mathbf{r}_j^c)^2 + m_j^2]^{1/2}}\right)\lambda^{5-3n}\,\delta(\mathbf{R}_1^c, \mathbf{R}^c),$$

$$\equiv J_E^c(\mathbf{R}^c)\,\delta(\mathbf{R}_1^c, \mathbf{R}^c), \tag{2.3}$$

where $\mathbf{p}_j = \lambda \mathbf{r}_j^c$ and

$$E = \sum_{j=1}^{n} [\lambda^2(\mathbf{r}_j^c)^2 + m_j^2]^{1/2}. \tag{2.4}$$

Now all states in "c" with 4-momentum $E \equiv (E, 0)$ can be represented as a superposition of the base states $|E, \mathbf{R}^c\rangle$. If f is a complex-valued function on $\mathcal{S}^c$, the state

$$X^c(f) = \int_{\mathcal{S}^c} d\mathbf{R}^c f(\mathbf{R}^c)\,|E, \mathbf{R}^c\rangle$$

is in channel "c", and all such states may be expressed in this form. Further,

$$\langle X^c(g)\,|\,X^c(f)\rangle = \int_{\mathcal{S}^c} d\mathbf{R}^c\, J_E^c(\mathbf{R}^c)\, g^*(\mathbf{R}^c) f(\mathbf{R}^c),$$

so that under the identification $f \longleftrightarrow X^c(f)$, the states of channel "c" may be identified with $L^2(\mathcal{S}^c, J_E^c)$, the space of square-integrable functions on $\mathcal{S}^c$ with a weighting function J_E^c. Under this identification, the asymptotic state space is therefore

$$\oplus_c' L^2(\mathcal{S}^c, J_E^c),$$

where the prime denotes that summation is only to be taken over channels that are open at the energy E. (The states which are realized physically have a nonzero component in just *one* channel.)

For fixed $\mathbf{R}^c$ and E satisfying $E > m_1 + \ldots + m_n$ there is, as we said, precisely one positive solution of equation (2.4). $\lambda(E)$ can clearly be analytically continued away from this region, and it is not difficult to show that there are branch points of the square-root type at $E = \pm m_1 \pm m_2 \pm \ldots \pm m_n$, where $\pm(m_1 + \ldots + m_n)$ are the principal thresholds of the channel and the other 2^{n-2} points are pseudo-thresholds. We define the principal branch of λ to be the function with the following cuts:

$$-\infty < \lambda < -(m_1 + \cdots + m_n), \quad m_1 + \cdots + m_n < \lambda < \infty,$$

$$m_1 - m_2 - \ldots - m_n < \lambda < -m_1 + m_2 + \ldots + m_n,$$

where m_1 is the smallest mass (see Fig. 1).

When E describes the path labeled Γ in Fig. 2, the path of λ is as shown. Thus for continuation round the threshold cut λ changes sign or, more generally, λ satisfies the equation $\lambda(E^*) = -\lambda(E)$. It can be seen that the points $im_j/|\mathbf{r}_j|$ are not circled in the λ plane, and hence $[\lambda^2(\mathbf{r}_j^c)^2 + m_j^2]^{1/2}$ are single-valued functions of E. Thus the Jacobian function $J_E^c(\mathbf{R}^c)$ appearing in Eq. (2.3) is multiplied by $(-1)^{3n-5}$ when λ is continued round the threshold cut.

Thus far we have been considering asymptotic states, either of the "in" or "out" type, with a fixed 4-momentum $E \equiv (E, 0)$. The scattering operator S_E links the "in" and "out" states, and under the identification described above provides a unitary map of $\oplus_c' L^2(\mathcal{S}^c, J_E^c)$ into itself (the prime again denotes open channels). S_E^{dc} is the restriction of S_E to channels "c" and "d", and maps the "in" space $L^2(\mathcal{S}^c, J_E^c)$ into the "out" space $L^2(\mathcal{S}^d, J_E^d)$; *a priori*, it is only defined if "c" and "d" are open at the energy E. The matrix elements of S_E are thus

$$S_E^{dc}(\mathbf{R}^d | \mathbf{R}^c) = {}_{\text{out}}\langle E, \mathbf{R}^d | E, \mathbf{R}^c\rangle_{\text{in}}.$$

The on-shell transition matrix T_E is given in operator form by

$$S_E = I + iT_E$$

and in matrix form by

$$S_E^{dc}(\mathbf{R}^d|\mathbf{R}^c) = \begin{cases} T_E^{dc}(\mathbf{R}^d|\mathbf{R}^c) & \text{if "}d\text{"} \neq \text{"}c\text{"}, \\ \delta(\mathbf{R}^d, \mathbf{R}^c) + iT_E^{dc}(\mathbf{R}^d|\mathbf{R}^c) & \text{if "}d\text{"} = \text{"}c\text{"}. \end{cases}$$

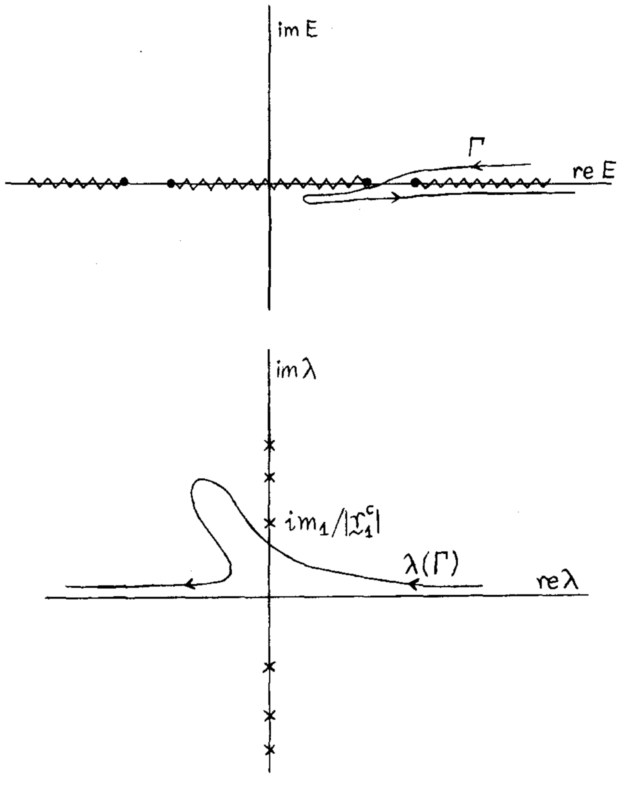


FIG. 2. The image $\lambda(\Gamma)$ in the λ plane of a curve Γ which passes round the threshold cut in the E plane.

It is now possible to speak (where appropriate) of the analytic continuation of the transition matrix elements when E varies and the "relative" variables $\mathbf{R}^c, \mathbf{R}^d$ are fixed; that such continuations do in fact exist is the basis of dispersion relations in S-matrix theory.

3. DYNAMICS

The off-shell Green's function for a scattering process involving N neutral particles with fields $\phi_1, \ldots, \phi_N$ and masses $m_1, \ldots m_N$ is defined as $G(p_1, \ldots, p_N)$, where

$$G(p_1, \ldots, p_N)\,\delta^4(p_1 + \cdots + p_N)$$
$$= (-i)^N \int\left(\prod_{j=1}^N d^4 x_j\right) \exp(-i\sum_j p_j x_j)\langle 0|T[\phi_1(x_1)\cdots\phi_N(x_N)]|0\rangle$$

and T is the time-ordering operator. The off-shell transition function is

$$T(p_1, \ldots, p_N) = -iG(p_1, \ldots, p_N)\prod_{j=1}^N (p_j^2 - m_j^2)$$

and by the LSZ reduction formulas the physical transition matrix for the process in which the first n particles leave with on-shell 4-momenta $-p_1, \ldots, -p_n$ and the last $N-n$ particles enter with on-shell 4-momenta $p_{n+1}, \ldots, p_n$ is equal to $T(p_1, \ldots, p_N)$.

Now

$$G(p_1, \ldots, p_N)\,\delta^4(p_1 + \cdots + p_N)$$
$$= (-i)^N \sum_\pi \int\left(\prod_{j=1}^N d^4 x_j\right)\exp(-i\sum_j p_j x_j) \qquad (3.1)$$
$$\times \langle 0|\phi_{\pi(1)}(x_{\pi(1)})\cdots\phi_{\pi(N)}(x_{\pi(N)})|0\rangle\theta(x_{\pi(j)0} - x_{\pi(j+1)0}),$$

where

$$\theta(t) = \begin{Bmatrix} 0 & t<0 \\ 1: & t>0 \end{Bmatrix}$$

and π denotes a permutation of (1, 2, ..., N). Thus

$$G(p_1, \ldots, p_N)\,\delta^4(p_1 + \cdots + p_N)$$
$$= (-i)^N \sum_\pi \int\left(\prod_{j=1}^{N-1} d^4(x_{\pi(j)} - x_{\pi(j+1)})\right)$$
$$\times \exp\left(-i\sum_{j=1}^{N-1}(x_{\pi(j)} - x_{\pi(j+1)})(p_{\pi(1)} + \cdots + p_{\pi(j)})\right)$$
$$\times d^4 x_{\pi(N)} \exp[-ix_{\pi(N)}(p_{\pi(1)} + \cdots + p_{\pi(N)})]$$
$$\times\langle 0|\phi_{\pi(1)}(x_{\pi(1)})\cdots\phi_{\pi(N)}(x_{\pi(N)})|0\rangle\theta(x_{\pi(j)0} - x_{\pi(j+1)0})$$

and hence, writing $y_j = x_{\pi(j)} - x_{\pi(j+1)}$, we have

$$G(p_1, \ldots, p_N) = (2\pi)^4(-i)^N \sum_\pi \int\left(\prod_{j=1}^{N-1} d^4 y_j\,\theta(y_{j0})\right)$$
$$\times\exp\left(-i\sum_{j=1}^{N-1} y_j P_j^\pi\right)\omega_\pi(y_1, \ldots, y_{N-1}),$$

where $P_j^\pi = p_{\pi(1)} + \cdots + p_{\pi(j)}$ and ω_π is defined by

$$\omega_\pi(y_1, \ldots, y_{N-1}) = \langle 0|\phi_{\pi(1)}(x_{\pi(1)})\cdots\phi_{\pi(N)}(x_{\pi(N)})|0\rangle.$$

Now using the fact that products become convolution products under Fourier transformation and

$$\int_{-\infty}^{\infty} dx_0 \exp(-ix_0 p_0)\,\theta(x_0) = -i/(p_0 - i\epsilon),$$

we have

$$G(p_1, \ldots, p_N)$$
$$= -i(2\pi)^4 \sum_\pi \int dt_1 \cdots dt_{N-1}\,\frac{\hat\omega_\pi(t_1, \mathbf{P}_1^\pi, \ldots, t_{N-1}, \mathbf{P}_{N-1}^\pi)}{\prod_{j=1}^{N-1}(P_{j0}^\pi - t_j + i\epsilon)}, \quad (3.2)$$

where

$$\hat\omega_\pi(q_1, \ldots, q_{N-1}) =$$
$$\frac{1}{(2\pi)^{N-1}}\int\left(\prod_{j=1}^{N-1} d^4 y_j\right)\exp(-i\sum_j y_j q_j)\omega_\pi(y_1, \ldots, y_{N-1}).$$

In Eq. (3.2) the integration is taken over $t_j \geqslant [(\mathbf{P}_j^\pi)^2 + M_j^\pi]^{1/2}$, where M_j^π is the least mass occurring in the mass spectrum of the j-particle states

$$\phi_{\pi(1)}(x_{\pi(1)})\cdots\phi_{\pi(j)}(x_{\pi(j)})|0\rangle.$$

Equation (3.2) exposes the analyticity of G in its energy variables $p_{10}, \ldots, p_{N0}$ by expressing G as a sum of $N!$ multiple Cauchy integrals in the "nested" sequence of energy variables $P_{10}^\pi, \ldots, P_{N0}^\pi$, and the result is due to Taylor (Ref. 2). G thus has cut surfaces which are functions of the energy sums P_{j0}^π. For the scattering channel in which particles 1, 2, ..., n leave and particles $n+1, \ldots, N$ enter, the cut surfaces which are functions of the direct channel energy *only*, i.e., $P_{j0}^\pi = \pm(p_{10} + \cdots + p_{n0})$, occur in the parts of G which result from integration over the following domains in Eq. (3.1):

(i) the direct channel region (DCR)

$$x_{10}, \ldots, x_{n0} > x_{(n+1)0}, \ldots, x_{N0},$$

(ii) the antidirect channel region (anti-DCR)

$$x_{10}, \ldots, x_{n0} < x_{(n+1)0}, \ldots, x_{N0}.$$

To fix ideas we now consider a scattering process in which m neutral scalar particles with fields $\phi_1, \ldots, \phi_m$ enter with 4-momenta $p_1, \ldots, p_m$ and n neutral scalar particles with fields $\psi_1, \ldots, \psi_n$ leave with 4-momenta $q_1, \ldots, q_n$. If the total 4-momentum is $E \equiv (E_0, \mathbf{E})$ the relative 4-momenta are defined as $\bar p_j = p_j - E/m$, $\bar q_k = q_k - E/n$, and although they are not independent since $\sum \bar p_j = 0 = \sum \bar q_k$, it will pay to preserve the symmetrical description and retain all of the relative momentum variables. The Green's function

$G(-q_1, \ldots, -q_n, p_1, \ldots, p_m)$ will be denoted by

$G(q_1, \ldots, q_n | p_1, \ldots, p_m)$.

The energy-analytic representation (EAR) in Eq. (3.2) gives the analytic structure of G in the variables E, $\bar p_{j0}$, $\bar q_{k0}$, and provided the functions $\hat\omega_\pi$ are boundary values of analytic functions, the relative energy variables may be kept constant and G considered as an analytic function of E in a cut plane. The cuts which do not depend on $\bar p_j$, $\bar q_k$ (and are therefore "fixed") are the kinematic cuts, and as stated above they arise from contributions from the DCR and anti-DCR.

Now

$$G(q_1, \ldots, q_n | p_1, \ldots, p_m)\,\delta^4(E - F)$$
$$= (-i)^{m+n}\int\left(\prod_k d^4 y_k\right)\exp(i\sum_k q_k y_k)\langle 0|T\left[\prod_k \psi_k(y_k)\prod_j \phi_j(x_j)\right]|0\rangle$$
$$\times\exp(-i\sum_j p_j x_j)\left(\prod_j d^4 x_j\right),$$

where $E=\sum p_j$, $F=\sum q_k$, and j and k range from 1 to m and from 1 to n, respectively. For the incoming particles the centroid is defined by $X=(x_1+\cdots+x_m)/m$ and the relative coordinates by $\bar{x}_j=x_j-X$. Thus $\sum p_j x_j = EX+\sum \bar{p}_j \bar{x}_j$. Y and $\bar{y}_k$ are similarly defined for the outgoing particles and satisfy $\sum q_k y_k = FY+\sum \bar{q}_k \bar{y}_k$. If

$$\max_j(\bar{x}_{j0})=u$$

and

$$\min_k(\bar{y}_{k0})=v,$$

the DCR is defined by $\theta(Y_0-X_0+v-u)=1$,

where

$$\theta(t)=\begin{Bmatrix}0:t<0\\1:t>0\end{Bmatrix}.$$

Thus, if $|P\alpha\rangle$ is a complete set of states of 4-momentum P and quantum numbers $|\alpha\rangle$,

$$G^{\mathrm{DCR}}(q_1,\ldots,q_n|p_1,\ldots,p_m)\,\delta^4(E-F)$$
$$=(-i)^{m+n}\sum_{P,\alpha}\int\Big(\prod_k d^4y_k\Big)\exp[i(FY+\sum_k\bar{q}_k\bar{y}_k)]\langle 0|T\Big(\prod_k\psi_k(y_k)\Big)|P\alpha\rangle$$
$$\times\langle P\alpha|T\Big(\prod_j\phi_j(x_j)\Big)|0\rangle\exp[-i(EX+\sum_j p_j\bar{x}_j)]\Big(\prod_j d^4x_j\Big).$$

Using the translation operators,

$$\langle P\alpha|T\Big(\prod_j\phi_j(x_j)\Big)|0\rangle=\langle P\alpha|T\Big(\prod_j\phi_j(\bar{x}_j)\Big)|0\rangle\exp(iPX),$$

and replacing the measure $\Pi_j d^4x_j$ by $d^4X\,d\bar{x}$, where

$$d\bar{x}=\prod_j d^4\bar{x}_j\,\delta^4(x_1+\cdots+x_m),$$

we can now write

$$G^{\mathrm{DCR}}(q_1,\ldots,q_n|p_1,\ldots,p_m)\,d^4(E-F)$$
$$=(-i)^{m+n}\sum_{P,\alpha}\int d\bar{y}\,d^4Y$$
$$\times\exp\{i[(F-Q)Y+\sum\bar{q}_k\bar{y}_k]\}\langle 0|T[\prod_k\psi_k(\bar{y}_k)]|P\alpha\rangle$$
$$\times\langle P\alpha|T[\prod_j\phi_j(\bar{x}_j)]|0\rangle$$
$$\times\exp\{-i[(E-P)X+\sum_j\bar{p}_j\bar{x}_j]\}\theta(Y_0-X_0+v-u)\,d\bar{x}\,dX.$$

Now inserting

$$\theta(t)=\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{\exp(izt)}{z-i\epsilon}\,dz$$

and performing the X and Y integrals, we obtain

$$G^{\mathrm{DCR}}(q_1,\ldots|p_1,\ldots)=$$
$$-\frac{(2\pi)^8}{2\pi i}(-i)^{m+n}\sum_{P,\alpha}\int d\bar{y}\exp(i\sum_k\bar{q}_k\bar{y}_k)\langle 0|T[\prod_k\psi_k(\bar{y}_k)]|P\alpha\rangle$$
$$\times\langle P\alpha|T[\prod_j\phi_j(\bar{x}_j)|0\rangle$$
$$\times\exp(-i\sum_j\bar{p}_j\bar{x}_j)\,d\bar{x}\,\frac{\exp[i(E_0-P_0)(v-u)]}{E_0-P_0+i\epsilon}\,\delta^3(\mathbf{E}-\mathbf{P}). \tag{3.3}$$

The bound-state poles appear in expression (3.3), and if $|P\alpha\rangle$ is a bound state $|b\rangle$ of mass m_b, its contribution is obtained when $\sum_{P,\alpha}$ is replaced by $\int d^4P\,\delta^+(p^2-m_b^2)$, to give

$$-\frac{(2\pi)^8}{2\pi i}(-i)^{m+n}\int d\bar{y}\exp(i\sum_k\bar{q}_k\bar{y}_k)\langle 0|T[\prod_k\psi_k(\bar{y}_k)]|b\rangle$$
$$\times\langle b|T[\prod_j\phi_j(\bar{x}_j)]|0\rangle$$
$$\times\exp(-i\sum\bar{p}_j\bar{x}_j)\frac{\exp\{i[E_0-(\mathbf{E}^2+m_b^2)^{1/2}][v-u]\}}{[E_0-(\mathbf{E}^2+m_b^2)^{1/2}][\mathbf{E}^2+m_b^2]^{1/2}}.$$

The pole is thus generated on the upper sheet of the mass hyperboloid $E^2=m_b^2$ and its residue is

$$-[4\pi i(\mathbf{E}^2+m_b^2)^{1/2}]^{-1}\,G(\bar{q}_1,\ldots,\bar{q}_n|b)\,G(b|\bar{p}_1,\ldots,\bar{p}_m),$$

where

$$G(\bar{q}_1,\ldots,\bar{q}_n|b)=$$
$$(2\pi)^4(-i)^n\int d\bar{y}\exp(i\sum_k\bar{q}_k\bar{y}_k)\langle 0|T[\prod_k\psi_k(\bar{y}_k)|b\rangle$$

and

$$G(b|\bar{p}_1,\ldots,\bar{p}_m)=$$
$$(2\pi)^4(-i)^m\int\langle b|T[\prod_j\phi_j(\bar{x}_j)]|0\rangle\exp(-i\sum_k\bar{p}_j\bar{x}_j)\,d\bar{x}.$$

A similar pole is generated on the lower sheet of the hyperboloid when $|b\rangle$ makes its contribution to the anti-DCR and the pole structure of G near $E^2=m_b^2$ is given by the sum of these:

$$G(q_1,\ldots|p_1,\ldots)\sim-\frac{1}{2\pi i}\cdot\frac{G(\bar{q}_1,\ldots,\bar{q}_n|b)\,G(b|\bar{p}_1,\ldots,\bar{p}_m)}{E^2-m_b^2}.$$

The pole thus has the characteristic form for a stable particle in the direct channel, i.e., it is a function of E and its residue factorizes into two wavefunctions which are functions of the in- and out-relative momenta, respectively.

It will now be assumed that there are no bound states in the theory, so that the Hilbert space is spanned by the many particle states of the elementary particles, both in the in- and out-asymptotic representations; as shown by Zimmerman[3] this does not involve a loss of generality. We take a center of mass frame where the total 4-momentum is $E\equiv(E,0)$, and use the following notation for the on- and off-shell transition functions for scattering between channels "c" and "d":

(i) $T_E^{dc}(\bar{q}_1,\ldots|\bar{p}_1,\ldots)=T(q_1,\ldots|p_1,\ldots)$,

(ii) $T_E^{dc}(\bar{q}_1,\ldots|\mathbf{R}^c)=T(q_1,\ldots|p_1(E,\mathbf{R}^c),\ldots)$,

where $p_j(E,\mathbf{R}^c)$ are the on-shell 4-momenta for channel "c", associated with $\mathbf{R}^c\in\mathcal{S}^c$ and E above the physical threshold of "c", as described in Sec. 2.

(iii) The on-shell transition matrix $T_E^{dc}(\mathbf{R}^d|\mathbf{R}^c)$ is similarly defined for channels "c" and "d".

These definitions will enable us to vary the total energy E while the remaining variables are kept fixed.

In order to obtain analagous results to the above in the case of *unstable* particles (rather than bound states), it will be necessary to make the following assumptions about the analytic structure of the transition functions.

Assumption 1: $T_E^{dc}(\bar{q}_1,\ldots|\bar{p}_1,\ldots)$ is the boundary value of an analytic function of E, given by the prescription $\epsilon\to+0$ in the EAR, Eq. (3.2). The spectral functions

ω_r are analytic, so that continuation through the cuts is possible.

The threshold cuts depending on E only and starting at the fixed branch points $E=\pm\sum m_j$, where m_1, $m_2, \ldots$ are the masses of the particles making up a certain channel, together constitute the kinematic cuts, and the continuation of T_E^{dc} around the kinematic cuts (avoiding the moving singularities) is defined as the physical sheet.

Assumption 2: Similar remarks may be made about $T_E^{dc}(\bar{q}_1, \ldots | \mathbf{R}^c)$ and $T_E^{dc}(\mathbf{R}^d | \mathbf{R}^c)$, in which $\mathbf{R}^d$ and $\mathbf{R}^c$ represent the relative momenta of the on-shell particles as in Sec. 2.

This assumption is rather different from Assumption 1, because when some of the particles are on-shell their relative 4-momenta are $\bar{p}_j = \bar{p}_j(E, \mathbf{R}^c)$, etc., and when $\mathbf{R}^c$ is kept fixed they are functions of E and vary as E is continued around the cut. Hence the path of continuation must also avoid the movable real singularities in $\lambda(E)$, the function defined in Sec. 2.

Now in the EAR it was shown that the cuts in G which are functions of E alone arise from the DCR and the anti-DCR. Expression (3.3) can be decomposed into the sum of

$$-\frac{(2\pi)^8}{2\pi i}(-i)^{m+n}\sum_{P,\alpha}\int d\bar{y}\exp(i\sum_k \bar{q}_k\bar{y}_k)\langle 0|T\prod_k\psi_k(\bar{y}_k)\;|P\alpha\rangle$$
$$\times\langle P\alpha|T\Big[\prod_j\phi_j(\bar{x}_j)\Big]|0\rangle\exp(-i\sum_j\bar{p}_j\bar{x}_j)\,d\bar{x}\times\frac{\delta^3(\mathbf{P})}{E-P_0+i\epsilon}$$

and a term that is nonsingular on the kinematic cut. ΔG, the discontinuity in G across the kinematic cut, is obtained from this expression by inserting a complete set of "in" states. Since the states belong to continuums for each asymptotic channel, $\sum_{P,\alpha}$ is replaced by $\int d^4P\sum_a'$, where $\sum_a'$ denotes a sum of integrals $\int_{\mathcal{S}^a} d\mathbf{R}^a$ over channels "a" which are open at the energy E, to give

$$\Delta G_E^{dc}(\bar{q}_1, \ldots |\bar{p}_1, \ldots)$$
$$\equiv\sum_b{}'\sum_a{}' G_E^{db}(q_1, \ldots |\mathbf{R}^b)J_E^b(\mathbf{R}^b)$$
$$\times S_E^{ab*}(\mathbf{R}^b|\mathbf{R}^a)J_E^a(\mathbf{R}^a)G_E^{ac}(\mathbf{R}^a|p_1, \ldots),$$

where

$$G_E^{db}(\bar{q}_1, \ldots |\mathbf{R}^b)$$
$$=(2\pi)^4(-i)^n\int d\bar{y}\exp(i\sum_k\bar{q}_k\bar{y}_k)\langle 0|T\Big[\prod_k\psi_k(\bar{y}_k)\Big]|E,\mathbf{R}^b\rangle_{\text{in}}$$

and

$$G_E^{ac}(\mathbf{R}^a|p_1, \ldots)$$
$$=(2\pi)^4(-i)^m\int_{\text{out}}\langle E,\mathbf{R}^a|T\Big[\prod_j\phi_j(\bar{x}_j)\Big]|0\rangle\exp(-i\sum_j\bar{p}_j\bar{x}_j)\,d\bar{x}.$$

For the T matrix we can thus write

$$T_{E*}^{dc}(\bar{q}_1, \ldots |\bar{p}_1, \ldots)=T_E^{dc}(\bar{q}_1, \ldots |\bar{p}_1, \ldots) \tag{3.4}$$
$$-i\sum_b{}'\sum_a{}' T_E^{db}(\bar{q}_1, \ldots |\mathbf{R}^b)J_E^b(\mathbf{R}^b)$$
$$\times S_E^{ab*}(\mathbf{R}^b|\mathbf{R}^a)J_E^a(\mathbf{R}^a)T_E^{ac}(\mathbf{R}^a|\bar{p}_1, \ldots).$$

T_{E*}^{dc} represents the value of T_E^{dc} at a point on the lower lip of the kinematic cut opposite the energy E on the upper lip, and is the continuation of T_E^{dc} along a path which avoids the moving singularities.

Consider the case where $\bar{p}_j$ are the on-shell relative momenta of an m-particle state of total energy E described by $\mathbf{R}^c\in\mathcal{S}^c$, with $\bar{p}_j=\bar{p}_j(E,\mathbf{R}^c)$. Then, as proved in Sec. 2, $\bar{p}_j(E^*,\mathbf{R}^c)=I_s\bar{p}_j(E,\mathbf{R}^c)$, where I_s is the space-inverting operator. Hence, *for l-wave angular momentum states* in the direct channel,

$$T_{E*}^{dc}(\bar{q}_1, \ldots |\bar{p}_1(\mathbf{R}^c, E^*)\ldots)=(-)^l T_{E*}(\bar{q}_1, \ldots |\mathbf{R}^c).$$

Thus

$$(-)^l T_{E*}^{dc}(\bar{q}_1, \ldots |\mathbf{R}^c)=T_E^{dc}(\bar{q}_1, \ldots |\mathbf{R}^c)$$
$$-i\sum_b{}'\sum_a{}' T_E^{db}(\bar{q}_1, \ldots |\mathbf{R}^b)J_E^b(\mathbf{R}^b)S_E^{ab*}(\mathbf{R}^b|\mathbf{R}^a)J_E^a(\mathbf{R}^a)T_E^{ac}(\mathbf{R}^a|\mathbf{R}^c).$$

Remembering that the on-shell T and S matrices satisfy $S_E=I+iT_E$, and using the unitarity of S_E, $S_E{}^*S_E=I$, this gives

$$T_E^{dc}(\bar{q}_1, \ldots |\mathbf{R}^c)$$
$$=(-)^l\sum_b{}' T_{E*}^{db}(\bar{q}_1, \ldots |\mathbf{R}^b)J_E^b(\mathbf{R}^b)S_E^{cb}(\mathbf{R}^b|\mathbf{R}^c). \tag{3.5}$$

Substituting this expression for T_E in (3.4),

$$T_{E*}^{dc}(\bar{q}_1, \ldots |\bar{p}_1, \ldots)-T_E^{dc}(\bar{q}_1, \ldots |\bar{p}_1, \ldots)$$
$$=(-)^{l+1}i\sum_a{}' T_{E*}^{da}(\bar{q}_1, \ldots |\mathbf{R}^a)J_E^a(\mathbf{R}^a)T_E^{ac}(\mathbf{R}^a|\bar{p}_1, \ldots). \tag{3.6}$$

Putting $\bar{q}_1, \ldots \bar{q}_n$ on-shell in (3.4), we obtain a similar equation to (3.5)

$$(-)^l T_E^{dc}(\mathbf{R}^d|\bar{p}_1, \ldots)=\sum_b{}' S_E^{db}(\mathbf{R}^d|\mathbf{R}^b)J_E^b(\mathbf{R}^b)T_E^{bc}(\mathbf{R}^b|\bar{p}_1, \ldots).$$

Thus, substituting this expression for T_E in (3.6),

$$T_E^{dc}(\bar{q}_1, \ldots |\bar{p}_1, \ldots)-T_{E*}^{dc}(\bar{q}_1, \ldots |\bar{p}_1, \ldots)$$
$$=i\sum_a{}'\sum_b{}' T_{E*}^{da}(\bar{q}_1, \ldots |\mathbf{R}^a)J_E^a(\mathbf{R}^a)$$
$$\times S_E^{ab}(\mathbf{R}^a|\mathbf{R}^b)J_E^b(\mathbf{R}^b)T_{E*}^{bc}(\mathbf{R}^b|\bar{p}_1, \ldots).$$

This formula will be the basis of our discussion of off-shell second-sheet structure.

The connection between the off-sheel unitarity relation (3.6) and Hermitian analyticity is seen by putting all of the particles on-shell:

$$T_{E*}^{dc}(\mathbf{R}^d|\mathbf{R}^c)-T_E^{dc}(\mathbf{R}^d|\mathbf{R}^c)=$$
$$-i\sum_a{}' T_{E*}^{da}(\mathbf{R}^d|\mathbf{R}^a)J_E^a(\mathbf{R}^a)T_E^{ac}(\mathbf{R}^a|\mathbf{R}^c).$$

Unitarity gives

$$T_E^{cd}(\mathbf{R}^c|\mathbf{R}^d)^*-T_E^{dc}(\mathbf{R}^d|\mathbf{R}^c)=$$
$$-i\sum_a{}' T_E^{ad}(\mathbf{R}^a|\mathbf{R}^d)^*J_E^a(\mathbf{R}^a)T_E^{ac}(\mathbf{R}^a|\mathbf{R}^c).$$

Thus $T_{E*}^{dc}(\mathbf{R}^d|\mathbf{R}^c)=T_E^{cd}(\mathbf{R}^c|\mathbf{R}^d)^*$, which is the formula for Hermitian analyticity.

4. SECOND-SHEET STRUCTURE AND RESONANCES

In the previous section a basic minimal analyticity was assumed which is sufficient to define the continuation of the transition matrix elements around the kinematic cut in the physical sheet. We shall now discuss the continuation of the scattering elements downwards *through* the kinematic cut into the second Riemann sheet, which is the *unphysical* sheet.

A priori the scattering operator S_E^{dc} is defined on the space $\oplus_c' L^2(\mathcal{S}^c, J_E^c)$. For this to remain a Hilbert space when E is complex we extend the definition of the inner product, and for $\phi^c, \psi^c \in L^2(\mathcal{S}^c, J_E^c)$, where E is complex, put

$$\langle \phi^c | \psi^c \rangle = \int d\mathbf{R}^c |J_E^c| \phi^c(\mathbf{R}^c)^* \psi^c(\mathbf{R}^c).$$

Let us now suppose that ϕ_E^c is a vector in $L^2(\mathcal{S}^c, J_E^c)$ for different values of E. Technically, the vector ϕ_E^c belongs to a different Hilbert space for each value of E. However, in practice this presents no problem, since for two complex energies E and E', $L^2(\mathcal{S}^c, J_E^c)$ and $L^2(\mathcal{S}^c, J_{E'}^c)$ are isomorphic under the map $\phi^c \to |J_{E'}^c / J_E^c|^{1/2} \phi^c$. Using this identification we may therefore define continuity, analyticity, etc. of ϕ_E^c with respect to E.

At a real physical energy E the S-matrix operator satisfies $S_E^* S_E = I$. Assuming that S_E possesses an analytic continuation into $\mathrm{im}E > 0$ and remains a bounded and invertible operator, $(S_{E^*}^*)^{-1}$ is defined and analytic in a region of $\mathrm{im}E < 0$ which is the mirror image of the physical-sheet domain. This operator agrees with S_E for real values of E, and is therefore a downward continuation of S_E, called the second or unphysical sheet of S_E and denoted S_E^{II}.

If the continuation is carried out from $E^a < E < E^b$ where E^a and E^b are consecutive threshold energies for channels a and b, the second-sheet scattering operator is defined on the space $\oplus_c' L^2(\mathcal{S}^c, J_E^c)$ and has components $(S_E^{dc})^{\mathrm{II}}$, where c and d range over channels which are open when $E^a < E < E^b$. By choosing different consecutive thresholds E^a and E^b we can therefore define $(S_E^{dc})^{\mathrm{II}}$ in several ways. Unlike the first-sheet structure these definitions do not all agree, because every threshold energy is a branch-point and comparison can only be made between different values of $(S_E^{dc})^{\mathrm{II}}$ by circling one or more of these branch points. To prevent such ambiguities, $(S_E^{dc})^{\mathrm{II}}$ will denote the continuation of S_E^{dc} from the interval (E^a, E^b), where E^a and E^b are consecutive thresholds and $E^a < \mathrm{re}E < E^b$. This is clearly the most direct path of continuation, and physically the most significant. The unphysical sheet so defined has its threshold cuts pointing vertically downwards.

The reason why the second-sheet structure of S_E is important is, of course, because for small values of ϵ, energies $E - i\epsilon$ on the second sheet are topologically near to the real scattering region, just as energies $E + i\epsilon$ on the first sheet are also near. Suppose that S_E^{II} has a pole $A/(E - E_r)$, where the residue A is an *operator*, and $\mathrm{im}E_r < 0$. If $\mathcal{N}$ is the null space of A and $\mathcal{R}$ is its range, then the restriction $A : \mathcal{N}^\perp \to \mathcal{R}$ is one to one and onto. Now G, the product of the rotation and internal symmetry groups of the theory, has in the Hilbert space a representation U that commutes with A and thus leaves $\mathcal{N}$ and $\mathcal{N}^\perp$ invariant. Assuming that the pole is not accidentally degenerate, the representation induced by U in $\mathcal{N}^\perp$ is irreducible and therefore finite-dimensional (since G is compact). Thus A can be written

$$\sum_{r=1}^{n} |\Psi_r\rangle\langle\Phi_r|,$$

where $\{\Phi_r\}$ is an orthonormal basis for $\mathcal{N}^\perp$, $\Psi_r = A\Phi_r$ and n is the degeneracy. When these conditions are satisfied the pole is called a *resonance* of complex mass E_r.

Since Φ_r can be obtained from the orbit (under U) of a single vector Φ, and Ψ_r can similarly be obtained from $\Psi \equiv A\Phi$, the pole will be denoted by $|\Psi\rangle\langle\Phi|/(E - E_r)$. (If the channels are assumed to have fixed quantum numbers, this representation is literally correct because the degeneracy is 'factored away'.) Putting $\Phi = \sum_c' \phi^c$, $\Psi = \sum_c' \psi^c$ we have $(S_E^{dc})^{\mathrm{II}} \sim |\psi^d\rangle\langle\phi^c|/(E - E_r)$ near the pole, and when ϕ^c is nonzero the resonance is then coupled to channel "c". The S matrix elements satisfy $S_E^{dc}(\mathbf{R}^d | \mathbf{R}^c)^{\mathrm{II}} \sim \psi^d(\mathbf{R}^d)^* \phi^c(\mathbf{R}^c)/(E - E_r)$ and have factorizable residues characteristic of a resonance. (The factorizability is verifiable for a resonance provided it is possible to determine the second-sheet residues for the different processes in which the resonance appears.)

Let us now consider the presence of resonances in off-shell amplitudes. The equation

$$\begin{aligned} T_E^{dc}(\bar{q}_1, \ldots | \bar{p}_1, \ldots) &= T_{E^*}^{dc}(\bar{q}_1, \ldots | \bar{p}_1, \ldots) \\ &+ i \sum_a{}' \sum_b{}' T_{E^*}^{da}(\bar{q}_1, \ldots | \mathbf{R}^a) J_E^a(\mathbf{R}^a) \\ &\times S_E^{ab}(\mathbf{R}^a | \mathbf{R}^b) J_E^b(\mathbf{R}^b) T_{E^*}^{bc}(\mathbf{R}^b | \bar{p}_1, \ldots) \end{aligned} \qquad (4.1)$$

can be written, when E is continued downwards into the second sheet,

$$\begin{aligned} T_E^{dc}(\bar{q}_1, \ldots | \bar{p}_1, \ldots)^{\mathrm{II}} &= T_E^{dc}(\bar{q}_1, \ldots | \bar{p}_1, \ldots)^{\mathrm{I}} \\ &+ i \sum_a{}' \sum_b{}' T_E^{da}(\bar{q}_1, \ldots | \mathbf{R}^a)^{\mathrm{I}} J_E^a(\mathbf{R}^a) S_E^{ab}(\mathbf{R}^a | \mathbf{R}^b)^{\mathrm{II}} \\ &\times J_E^b(\mathbf{R}^b) T_E^{bc}(\mathbf{R}^b | \bar{p}_1, \ldots)^{\mathrm{I}}, \end{aligned}$$

where the superscript I emphasizes first- (physical-) sheet values. The highly interesting thing about this equation is that at a resonance pole in S_E^{II}, where

$$S_E^{dc}(\mathbf{R}^d | \mathbf{R}^c)^{\mathrm{II}} \sim \psi^d(\mathbf{R}^d)^* \phi^c(\mathbf{R}^c)/(E - E_r)$$

the off-shell unphysical-sheet amplitude $T_E^{dc}(\bar{q}_1, \ldots | \bar{p}_1, \ldots)^{\mathrm{II}}$ has a pole with a factorizable residue:

$$T_E^{dc}(\bar{q}_1, \ldots | \bar{p}_1, \ldots)^{\mathrm{II}} \sim \chi_{\mathrm{out}}(\bar{q}_1, \ldots) \chi_{\mathrm{in}}(\bar{p}_1, \ldots)/(E - E_r),$$

where

$$\chi_{\mathrm{out}}(\bar{q}_1, \ldots) = \sum_a{}' T_{E_r}^{da}(\bar{q}_1, \ldots | \mathbf{R}^a)^{\mathrm{I}} J_{E_r}^a(\mathbf{R}^a) \psi^a(\mathbf{R}^a)^*,$$

and

$$\chi_{\mathrm{in}}(\bar{p}_1, \ldots) = \sum_b{}' \phi^b(\mathbf{R}^b) J_{E_r}^b(\mathbf{R}^b) T_{E_r}^{bc}(\mathbf{R}^b | \bar{p}_1, \ldots)^{\mathrm{I}}.$$

This demonstrates that the second-sheet structure of the off-shell T matrix is such that at a resonance energy the components of T have poles in E with residues which factorize into wavefunctions of the incoming and outgoing relative momenta. This is completely analogous to the first-sheet behavior of the T matrix in the neighborhood of a bound state energy and in agreement with the alternative description of resonances as unstable particles. Although it would be convenient to identify Φ or Ψ as the *state* of the resonance, it must be remembered that at real physical values of E the Hilbert space is not $\oplus' L^2(\mathcal{S}^c, J_E^c)$, but an isomorphic copy of this. The isomorphism depends on the total energy E and Φ and Ψ belong to the space when E assumes an unphysical *complex* value: Thus no direct physical meaning can be attached to these vectors.

5. CONCLUSION

The extension to the case of charged particles is trivial, nor do we expect the occurrence of spin in the scattered particles to present any difficulty.

It is interesting to compare these results with previous work on the scalar Bethe–Salpeter equation with an exchange potential.[4] On performing the Wick rotation[5] by which the relative energy variables are continued to the imaginary axis, a scattering equation may be obtained that is analytic in a subset of the total energy plane, and this subset includes the real elastic scattering region.[6] The effect of the Wick rotation is thus to remove altogether the moving singularities. The double-sheet structure of the off-shell amplitude, with imaginary relative energies, may then be exhibited for the continuation of the total energy through the elastic cut, although the situation when higher energies, and therefore inelastic thresholds, are admitted, is not so clear.

ACKNOWLEDGMENT

I am grateful to Dr. M.M. Broido of Magdalen College, Oxford for helpful advice.

[1] R.J. Eden, *et al.*, *The Analytic S-Matrix* (Cambridge U.P., Cambridge, 1966).

[2] J.G. Taylor, *Lectures in Theoretical Physics*, edited by W.E. Britten, A.O. Barut, and M. Guenin (Gordon and Breach, New York, 1967), Vol. IXA, pp. 353–400.

[3] W. Zimmerman, Nuovo Cimento 10, 597 (1958).

[4] E.E. Salpeter and H.A. Bethe, Phys. Rev. 84, 1232 (1951).

[5] J.C. Wick, Phys. Rev. 96, 1124 (1954).

[6] M.J. Levine, *et al.*, Phys. Rev. 154, 1433 (1967); 157, 1416 (1967).

The renormalization group and the large n limit

Shang-keng Ma*

Department of Physics and Institute for Pure and Applied Physical Sciences, University of California, San Diego, La Jolla, California 92037

(Received 21 March 1973)

The basic concepts and formulation of the renormalization group are explained beginning at an elementary level. Discussion is in the framework of classical statistical mechanics with emphasis on applications to the theory of critical phenomena. The details are worked out in the large n limit for $2 < d < 4$, where n is the number of components of the fluctuating field of interest and d is the dimension of the thermodynamical system. In the large n limit, the infinite sum of "tree graphs" offers an exact and analytically tractable description of the renormalization group. It illustrates many concepts including the fixed point, the critical surface in the space of coupling parameters, and critical exponents. Most important, it illustrates the origin and the limitation of the scaling hypothesis. The critical behavior of various correlation functions and the free energy is examined. Attention is paid to terms often ignored in qualitative scaling arguments. We have attempted to make this paper self-contained and of pedagogical value to a wide audience.

I. INTRODUCTION

The notion of a renormalization group appeared decades ago in relativistic field theories.[1] It appeared in the study of the relationship between the momentum cutoff and coupling constants. Over the past several years, Wilson has made important advances in bringing the ideas of renormalization group into concrete and useful concepts and has successfully applied them to different areas of physics.[1] So far the most successful application has been to the theory of critical phenomena.[2] On the other hand, existing knowledge in critical phenomena has been very helpful in understanding the renormalization group idea as well. In this paper, we shall explain the idea of renormalization group beginning at an elementary level. Our discussion will be within the framework of statistical mechanics of an n-component classical field in a d-dimensional space. If $n=3$, $d=3$, this classical field would describe the fluctuation of magnetization in a ferromagnetic material, for example. We expect also that the amplitude of ^{4}He atoms, which becomes large near the λ point of liquid HeII, can be adequately described by a classical field with $n=2$, $d=3$.

Although the basic principles are established and numerical investigations have begun, the complexities of the renormalization group machinery makes idealized model calculations highly desirable for illustrating the general features. The first simple analytical illustration of the renormalization group was found by Wilson and Fisher,[3] who demonstrated that for small $\epsilon \equiv 4-d$, the mathematical complication disappeared. Once the structure of the renormalization group was understood for small ϵ, perturbation theory calculations of critical exponents as expansions in powers of ϵ followed.[4] Simplicity was expected also in the limit of large n. The limit of large n first appeared in the "spherical model."[5] More recently, much work has been done in computing critical exponents as power series in $1/n$ and in studying field theory models with large n.[6–8] The renormalization group in the large n limit, which is fundamental to the understanding of the results pertaining to large n, was expected to be tractable analytically, but so far no comprehensive and reasonably complete information has been available in the literature. This paper is to present this information. We illustrate the full details of the renormalization group in the large n limit. This illustration is more complicated than the small ϵ limit, but it demonstrates many important features which are difficult to visualize in the small ϵ limit.

It is hoped that this paper will serve pedagogical purposes. We shall include discussions at a very elementary level so that this paper is self-contained as well for those readers who are not familiar with the theory of critical phenomena or some jargons of field theory. These elementary discussions have been included in a recent review article.[9]

An introduction to the use of graph representation is included. Graphs will be used for studying the large n limit. However, we want to emphasize that the renormalization group idea is valuable partly because *it is free from any perturbation theory*, i.e., it is a nonperturbative concept. The graph representation, which is a perturbation expansion, is *not essential* to the study of renormalization group. It is nevertheless useful as a tool of simple calculation, and make some ideas easier to visualize. In analyzing the large n limits, we shall sum over an infinite set of graphs and our results demonstrate well nonperturbative features of the renormalization group. The analysis can be done without introducing graphs at all.[9]

Before giving the outline of this paper, it should be helpful to review a few basic ideas in the theory of critical phenomena.[10] Let us imagine a sample of isotropic ferromagnetic material. If the temperature T is below its critical temperature T_c, there is a spontaneous magnetization. Right above T_c, there is not. There are large fluctuations in magnetization for T near T_c. As the temperature T approaches T_c the magnetic susceptibility and some other measurable quantities diverge. For example, the susceptibility diverges like $(T-T_c)^{-\gamma}$, for $T > T_c$, where γ, one of the *critical exponents*, is observed to be near 1.3 for many materials exhibiting a critical point. The theory of critical phenomena has the task of explaining these divergences.

These divergences are believed to be consequences of the large fluctuations of magnetization. Also the observed universal (i.e., independent of materials) character of these divergences suggests that only the large scale behavior, not the detail microscopic interactions, is relevant in a correct explanation.

A useful concept is the correlation length ξ, which may be thought of as measuring the average distance over which the fluctuations of magnetization are correlated. The *scaling hypothesis* says that ξ should be the longest and the only relevant length in explaining critical phenomena. It says also that ξ, diverging like $|T-T_c|^{-\nu}$, $\nu > 0$, counts for the dominating temperature dependence near T_c of all quantities. In other words, physical quantities depend on $T-T_c$ only through their dependence on ξ. For example, it leads to the following very important consequence. If we increase the unit of length by a factor s, then in the new unit, the system appears *shrunk* by a factor s. The correlation length now becomes ξ/s under this scale change. Since the correlation length is proportional to $|T-T_c|^{-\nu}$, a decrease in correlation length corresponds to an increase in $|T-T_c|$. Therefore, near T_c, the temperature dependence of a physical quantity can be deduced from the way it behaves under a change of scale. The simplest example of applying this idea is the following. The free energy per unit volume $F(\xi)$ becomes $s^dF(\xi)$ when the volume of the system is shrunk; d is the dimension. Therefore $F(\xi/s)=s^dF(\xi)$. Since s is arbitrary, we set $s=\xi$. We then get

$$F(\xi)=s^{-d}F(\xi/s)=\xi^{-d}F(1) \propto |T-T_c|^{\nu d}, \tag{1.1}$$

since $\xi \propto |T-T_c|^{-\nu}$. Later we shall examine the validity of such arguments. Another important consequence is that in the limit $T=T_c$, ξ becomes infinite and there is no longer any length parameter. Thus the system would look the same if a change in length scale is made. There are many important consequences of the scaling hypothesis and also ambiguities. It is clear that this hypothesis is very powerful but its origin is not clear. A more fundamental understanding is needed. Many authors have made advances in these directions. Our understanding becomes more concrete and precise after Wilson's renormalization group formulation was developed.

The first half of this paper is devoted to discussions of basic ideas and to set up the machinery which is to carry out these ideas. It is emphasized that the basic idea is very simple even though the machinery required is rather complicated. It is hoped that these discussions will give a clear idea on what is precisely defined and what is unproved plausible hypotheses.

The second half of this paper is more technical. It illustrates how the renormalization group machinery works in the limiting case of large n. In this case, exact analytic expressions can be worked out by summing a special, infinite set of graphs, the so-called "tree graphs," which turn out to dominate in the large n limit.

We now give a brief sketch of what is done here using the tree graph summation. This is not a summary but will give the reader some idea of the content of the latter half of this paper.

In the large n limit, possible probability distributions for the classical field (order parameter fluctuation) $\phi_i(x)$, $i=1,\ldots,n$, take the form $P \propto \exp(-\mathcal{H})$ with

$$\mathcal{H}=\int d^dx[(\nabla\phi)^2+U(\phi^2(x))], \tag{1.2}$$

where

$$(\nabla\phi)^2=\tfrac{1}{2}\sum_i(\nabla\phi_i)^2, \quad \phi^2=\tfrac{1}{2}\sum_i\phi_i^2,$$

and $\phi_i(x)$ contains Fourier components of wave vectors up to a cutoff Λ. $U(\phi^2)$ is any function of ϕ^2 which approaches infinite as $\phi^2 \to \infty$. A renormalization-group transformation R_s, $1 \leq s < \infty$, takes a probability distribution P to another P', and is expressed as a transformation in the space of U's:

$$R_s : U \to U'. \tag{1.3}$$

At the same time, $\phi(x)$ is replaced by $s^{-d/2+1-\eta/2}\phi(x/s)$ ($\eta=0$ in the large n limit) so that the average of any function of ϕ over P is the same as the transformed function of ϕ taken over P'. In other words, R, behaves like a scale transformation for the probability distribution. The work done in the second half of this paper includes (a) the determination of the fixed point U^* (satisfying $R_s : U^* \to U^*$, a plot of which is given in Fig. 1. (b) A subspace of the space of U's is called the critical surface defined by properly fixing one parameter in U. A system at its critical temperature is represented by a point on this critical surface. We work out the details of $R_s : U \to U'$ for any U on or close to the critical surface (it does not matter whether it is above or below). The transformation of ϕ^2 is also discussed as an illustration of how products of ϕ transform under R_s. Critical behavior of various correlation functions are then determined by the transformations at large s. The fact that results are independent of the details of U is an illustration of universality. Note that (1.2) is not the most general form of $\mathcal{H}$. However, it is sufficiently general for the discussion within the framework of free graph summation.

The tree graphs summed here correspond to those of the self-consistent Hartree approximation in many-body theory. As will be seen, the tree graphs for the renormalization group in the large-n limit displays a very rich and appealing structure in addition to being exact in this limit. This is in contrast to the Hartree approximation in other applications.

The outline of the paper is the following:

Sec. II: Basic concepts are explained in detail. The meaning of coupling parameters with respect to a cutoff Λ is clarified. The renormalization group is defined as transformations in a "parameter space." Each point in the parameter space represents a possible probability distribution describing the statistical mechanical system. Formalism is set up. The notion of the fixed point is introduced.

Sec. III: Critical behaviors are related to the characteristics of the renormalization group acting close to a subspace called the "critical surface." Critical exponents related to the correlation function and susceptibility are introduced. The discussion is qualitative.

Sec. IV: Graphs are introduced and the graph representation of the renormalization group is explained.

Sec. V: Tree graphs are introduced. They are shown to dominate in the large-n limit. The fixed point is generated from a simple interaction. General features of the fixed point are illustrated.

Sec. VI: The details of the renormalization group transformation in general in the large-n limit are worked out. The critical surface, critical exponents, approach to the fixed point, and other concepts and assertions discussed in Sec. III are demonstrated explicitly. All results are exact in the large n limit. Corrections will be of $O(1/n)$.

Sec. VII: The effect of a uniform external field and general features below T_c are discussed. Characteristics of longitudinal and transverse susceptibilities are examined in detail in the large n limit. The exponents δ and β are discussed.

Sec. VIII: This section is devoted to a careful study of the free energy under the transformations of the renormalization group. Weakness of the usual scaling argument given by (1.1) is illustrated.

Sec. IX: A detailed study of the simplest composite variable ϕ^2 is carried out. The transformation of ϕ^2 under the renormalization group is described and the critical behavior of related correlation functions is examined. The concept of dimensions of variables is discussed.

Sec. X: Basis for perturbation theory calculation of critical exponents is discussed.

Sec. XI: Concluding remarks are made.

Much of Secs. II–IV is devoted to explaining the basic definitions. Those readers who are already familiar with the basics may read through Sec. II quickly to get an idea of the notation and then proceed to Sec. V. The range of material covered in this paper is very small. The emphasis is on the details of the renormalization group, not on reviewing its accomplishments.

Table I lists frequently occurring symbols and their defining equations.

II. RENORMALIZATION GROUP DEFINED

A renormalization group can be defined for any large system such as a thermodynamical system or a quantum field. We shall define a renormalization group for a model thermodynamical system analyzed in the framework of classical statistical mechanics. But before we proceed with our definitions, we would like to remind the reader of some truly trivial facts concerning probability distributions.

A. Digression on trivial observations

Let $P(y_1, y_2, y_3)$ be the probability distribution function for the random variables $-\infty < y_1, y_2, y_3 < \infty$. To calculate the average value of any function $f(y_1, y_2, y_3)$ of these random variables, for example, $f=y_1y_2$, we simply do the integral

$$\langle f\rangle_P=\langle y_1y_2\rangle_P=\int_{-\infty}^{\infty} dy_1dy_2dy_3y_1y_2P(y_1, y_2, y_3). \qquad (2.1)$$

We notice that for those f which do not depend on y_3, we

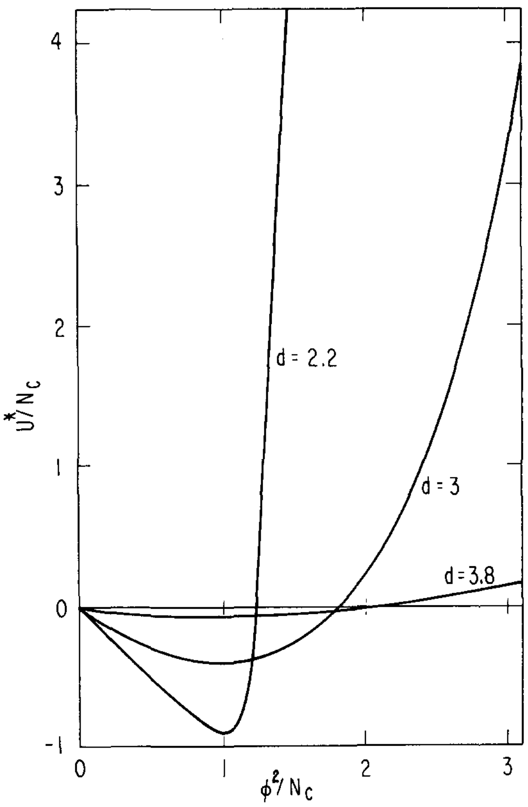


FIG. 1. The fixed point in the large n limit depicted as the function $U^*(\phi^2)$ for $d=2.2$, 3, and 3.8. See (5.45)–(5.54). The unit N_c is given by (5.15) and depends on d.

TABLE I. Symbols and where they are defined.

A	(8.4), (8.5)	$\mathcal{H}$	(2.12)	Π	(7.32)
A_l	(9.10)	$\mathcal{H}'$	(2.18)	R_s	(2.17)
α_s	(2.18), (2.27)	J	(7.33)	R_s^L	(3.2)
e_1, e_j	above (3.3)	K_d	(5.11)	s	(2.18)
η	(2.26)	L	above (2.6)	Σ	(4.11)
ξ	(3.17)	Λ	(2.11)	Σ_s	(4.20)
$F(T)$	(8.1)	λ	(5.38)	t_j	(3.4)
$\mathcal{J}$	(8.2)	μ	(2.13)	$t=t(\lambda)$	(6.15)
$\mathcal{J}'$	below (8.2), (8.4)	μ'	(2.17)	$t'=t'(\lambda)$	(5.38)
Φ	(5.23)	μ^*	(2.25)	t_0	(5.3)
$\overline{\phi}, \delta\overline{\phi}$	(4.12)	μ_1	(5.4)	t_1	(3.4), (6.18)
$\phi_i(x)$	(2.14)	$\mu(T)$	Sec. III. 2	t_0^*	(5.42)
ϕ_{ik}	(2.6)	$\delta\mu$	(3.1)	t^*	(5.41)
ϕ^2	(5.6)	N	(5.10)	u_{2m}	(2.12)
$(\phi^2)'$	(9.5), (9.6)	N_a, N_b	(5.30)—(5.32)	$\overline{u}_{2m}$	(4.20)
$(\nabla\phi)^2$	(5.46)	N_c	(5.15)	u'_{2m}	(4.23)
G	(2.8), (2.20), (2.29)	N'	(5.13)	$u(N)$	(7.26)
G_0	(4.2)	$N(\lambda)$	(5.38), (6.12)	U	(6.3), (6.16)
G_0'	(4.22)	N_s	(5.17)	U'	(6.17)
$G_\parallel, G_\perp$	(7.19), (7.20)	ν	(3.8)	U^*	(5.48)
H	(7.1)	P	(2.12)	y_1, y_2, y_j	(3.3)
$H(\Lambda)$	(2.11)	P'	(2.18)	ζ	(6.19)

can obtain an equivalent distribution function $P'(y_1, y_2)$ by integrating out the variable y_3 from $P(y_1, y_2, y_3)$, i.e.,

$$P'(y_1, y_2) \equiv \int_{-\infty}^{\infty} dy_3 P(y_1, y_2, y_3). \tag{2.2}$$

Therefore, let us remember

Fact 1: P', obtained from P by integrating out certain random variables, is equivalent to P provided we are not interested in these integrated variables. Next, we observe that if we obtain a new probability distribution $P'(y_2, y_4, y_6)$ from $P(y_1, y_2, y_3)$ by changing the name of random variables, we won't get anything new. For example,

$$P'(y_2, y_4, y_6) \equiv P(y_2, y_4, y_6), \tag{2.3}$$

i.e., just replacing 1, 2, 3 in $P(y_1, y_2, y_3)$ by 2, 4, 6. The only thing we must watch out for is that when we calculate averages, we must change labels accordingly. For example,

$$\begin{aligned}\langle y_2\rangle_P &= \int dy_1 dy_2 dy_3 P(y_1, y_2, y_3) y_2 \\ &= \int dy_2 dy_4 dy_6 P'(y_2, y_4, y_6) y_4 \\ &= \langle y_4\rangle_{P'},\end{aligned} \tag{2.4}$$

i.e., we must calculate the average of y_4 over P' if we want to get the average of y_2 over P. This sounds too trivial, but must be remembered:

Fact 2: P', obtained from P by relabeling the random variables, is equivalent to P provided that when average values are computed we relabel the random variables of interest accordingly. Finally, if α is a positive constant and

$$P'(y_1, y_2, y_3) \equiv \alpha^3 P(\alpha y_1, \alpha y_2, \alpha y_3) \tag{2.4'}$$

then P' clearly says nothing new. Any average calculated over P' is easily related to that over P. For example,

$$\langle y_1\rangle_P = \alpha\langle y_1\rangle_{P'}, \langle y_1^2\rangle_P = \alpha^2\langle y_1^2\rangle_{P'}. \tag{2.5}$$

Therefore, let us remember

Fact 3: P', obtained from P by changing random variables by a constant factor, is equivalent to P provided we multiply the random variables of interest by the same factor when average values are computed.

We list the above three trivial observations so that it will be easier for the reader to understand the more complicated, but basically the same procedures later. A transformation in the renormalization group essentially transforms a given probability distribution to an equivalent one in the above mentioned three steps: integration, relabeling, and multiplying random variables by a constant.

B. Model and notation

Imagine a d-dimensional crystal lattice of volume L^d, where L is measured in units of lattice spacing. At each lattice site x, there is an n-component vector "spin" $\phi(x) \equiv (\phi_1(x), \phi_2(x) \ldots \phi_n(x))$. Let ϕ_k denote Fourier components of $\phi(x)$:

$$\phi_i(x) = L^{-d/2} \sum_k \phi_{ik} \exp(ik\cdot x), \tag{2.6}$$

where the sum over wavevectors k is taken over the L^d discrete points in the first Brillouin zone. The density of points, $L^d(2\pi)^{-d}$, is very large since L is a very large number. Each ϕ_{ik} is regarded as a random variable. There are nL^d of them. The probability distribution for these random variables is given by

$$P_{\text{micro}} \propto \exp(-H_{\text{micro}}/T), \tag{2.7}$$

where T is the temperature and H_{micro} is the Hamiltonian which is assumed to be a given function of all the random variables. We assume that H_{micro} is invariant under rotations in the n-dimensional spin vector space.

The *correlation function* $G(k)$ is defined as

$$\begin{aligned}G(k)\delta_{ij} &= \int d^dx \langle\phi_i(x)\phi_j(0)\rangle \exp(-ik\cdot x) \\ &= \langle|\phi_{ik}|^2\rangle\delta_{ij},\end{aligned} \tag{2.8}$$

where the average $\langle\cdots\rangle$ is taken over P_{micro} given by (2.7). If a term

$$-\int d^dx \phi_1(x) H \tag{2.9}$$

is added to the Hamiltonian, i.e., when a "magnetic field" is turned on in the 1 direction, we can define the susceptibility as

$$\partial\langle\phi_i\rangle/\partial H. \tag{2.10}$$

It is very easy to show that the susceptibility per unit volume is just $G(0)$. Other quantities of interest will be defined later.

Since the probability distribution is assumed to be invariant under rotations in spin space, we expect $G(k)$ to be independent of i if there is no external field. However, a rotationally invariant probability distribution can still produce average values which are not rotationally invariant. This happens below T_c, where one of the components, say ϕ_1, has nonzero average even when $H=0$. In our discussions, we shall always assume that $H=0$ unless otherwise specified.

C. The idea of an effective Hamiltonian

What we are interested in is the behavior of long-wavelength fluctuations, i.e., that of ϕ_k with small k. The Hamiltonian is usually given by nearest-neighbor interactions. Since we expect that the characteristics of long wavelength fluctuations are independent of the microscopic details, we should be able to obtain an effective Hamiltonian with these irrelevant details removed. In other words, this effective Hamiltonian should not involve any ϕ_k with large k. Of course, the effective Hamiltonian must lead to the same results as the original Hamiltonian would when averages involving ϕ_k's with small k are calculated. How do we find this effective Hamiltonian? It is very easy in principle. Remember the trivial Fact 1 mentioned at the beginning of this section: We may simply integrate out the irrelevant random variables. Thus, P_{micro}, as given by (2.7), is equivalent to, apart from a normalization constant,

$$\prod_{i,k>\Lambda}\int d\phi_{ik}\exp(-H_{\text{micro}}/T)\equiv\exp[-H(\Lambda)/T], \tag{2.11}$$

where the multiple integral is taken over all ϕ_{ik}'s with all $i=1,\ldots,n$ and all k larger than Λ. The cutoff Λ is taken to be much less than the inverse lattice spacing but still much larger than the small range of k which is of interest ultimately. $H(\Lambda)$ defined by (2.11) is the desired effective Hamiltonian. Note that we set Λ this way to leave ϕ_k's in the intermediate k range unintegrated. This is because, besides the random variables in the small k range themselves, those in the intermediate k range also play an important part in determining the small k behavior. The effective Hamiltonian $H(\Lambda)$ tells us about the interactions down to a minimum distance Λ^{-1}. The finer details beyond this distance are averaged out. The multiple integrals in (2.11) will not be easy to carry out explicitly. However, we expect that $H(\Lambda)$ in general will look very different. For example, if the microscopic Hamiltonian has only quadratic and quartic terms in ϕ, the multiple integral of (2.11) will generate all powers of ϕ for $H(\Lambda)$. This will become more evident later. The important point to remember is that the cutoff Λ is an inseparable part of the definition of a Hamiltonian. The fluctuations over a distance less than Λ^{-1} does play a role in determining the structure of $H(\Lambda)$.

The ultimate task is to derive singular behavior of physical quantities such as the correlation function near the critical point from a generally nonsingular Hamiltonian. Constructing $H(\Lambda)$ does not seem to help in this task. No singularity is expected in $H(\Lambda)$ since we only smeared out fluctuation over very short distances. If we are now to study critical behaviors starting from $H(\Lambda)$, then the task would appear to be much worse than before because $H(\Lambda)$ would look far more complicated than the microscopic Hamiltonian. However, we will be able to see the major characteristics of the critical behavior, which are independent of the details of $H(\Lambda)$, by examining how $H(\Lambda)$ would behave under the renormalization group, which is a set of transformations and will be defined shortly.

D. The parameter space

We shall now be more general and consider a large class of probability distributions for ϕ_{ik}. We now forget about our spin model introduced above and regard ϕ_{ik}'s just as a set of random variables. But we still want the label k to range over discrete points in a sphere of radius Λ in k space. The density of points is $L^d(2\pi)^{-d}$. Of course, $1\leqslant i\leqslant n$, as before.

Any probability distribution for these random variables can be specified by a set of parameters. Let us imagine that each set of parameters is a point in a *parameter space*, so that any probability distribution P is represented by a point μ in this space. To make our discussion more concrete, let us illustrate how such a parameter space can be constructed. Write

$$P\propto\exp(-\mathcal{H}),$$

$$\mathcal{H}=\sum_{m=1}^{\infty}L^{-(m-1)d}\sum_{k_1,k_2\ldots k_{2m-1}}\;\sum_{i_1,i_2\ldots i_{2m}}\phi_{i_1k_1}\phi_{i_2k_2}\cdots\phi_{i_{2m}k_{2m}}u_{2m}$$

$$+\text{const}, \tag{2.12}$$

where $k_{2m}=-(k_1+k_2+\cdots+k_{2m-1})$ and u_{2m} is a function of $k_1,k_2\ldots k_{2m-1}$. We now define our parameter space as the space of all possible μ,

$$\mu\equiv(u_2,u_4,u_6,\cdots). \tag{2.13}$$

We shall refer to the u_{2m}'s as "coupling parameters." This space is of course enormous. The region of interest in this space will be very limited. Symmetry and other restrictions will be required for u_{2m}, for example. The additive constant in $\mathcal{H}$ is not included as a parameter. Odd powers of ϕ_k are not included but they will be needed when discussing external field. Anyway, further restrictions and adjustments can always be made when necessary. We shall stick to (2.12) for our general discussion.

We do want to emphasize that Λ, the cutoff in k space, is, unless otherwise specified, always fixed for all probability distributions. The coupling parameters are meaningless without fixing Λ. Another important point is that L, which tells us how many random variables there are, is not included as a parameter. This is because we are interested in the limit of infinite L. Averages of interest are always L-independent in this limit. In fact we shall write $\mu=\mu'$ as long as $u_{2m}=u'_{2m}$ for all m even if $L\neq L'$.

A slightly more appealing way of writing (2.12) is by introducing $\phi(x)$:

$$\phi(x) \equiv L^{-d/2} \sum_{k<\Lambda} \phi_k \exp(ik \cdot x); \tag{2.14}$$

then we have

$$\mathcal{H} = \sum_{m=1}^{\infty} \sum_{i_1, i_2 \ldots i_{2m}} \int d^d x_1 \ldots d^d x_{2m} \phi_{i_1}(x_1) \phi_{i_2}(x_2) \ldots \phi_{i_{2m}}(x_{2m}) \times v_{2m}(x_1 - x_{2m}, x_2 - x_{2m}, \ldots x_{2m-1} - x_{2m}), \tag{2.15}$$

where v_{2m} are related to u_{2m} via

$$u_{2m} = \int \prod_{l=1}^{2m-1} (d^d y_l \exp(-ik_l \cdot y_l)) v_{2m}(y_1, y_2 \ldots y_{2m-1}). \tag{2.16}$$

We shall assume that v_{2m} represents short range interactions (i.e., $v_{2m} \to 0$ if one or more of the y's becomes large) so that u_{2m} can be expanded in powers of k.

Finally, to those readers who are too used to statistical mechanical terminology, we want to emphasize that $\mathcal{H}$, defined by (2.12), is not to be thought of as "energy divided by temperature." It is just the logarithm of the probability distribution. As far as our parameter space is concerned, the concepts of energy and temperature are irrelevant. They enter only in (2.11) as inputs in determining a particular probability distribution corresponding to a particular point in the parameter space.

E. Renormalization group

Consider the following transformation which takes a probability distribution P to another probability distribution P'. We want to represent this transformation as

$$\mu' = R_s \mu \tag{2.17}$$

which transforms the point μ to μ' in the parameter space. Of course, μ and μ' represent P and P', respectively. This transformation R_s is defined implicitly by

$$P' \propto \exp(-\mathcal{H}') = \left[\prod_{i, \Lambda/s<k'<\Lambda} \int d\phi_{ik'} \exp(-\mathcal{H}) \right]_{\phi_k \to \alpha_s \phi_{sk}}. \tag{2.18}$$

Equation (2.12) defines μ, and μ' is to be extracted from $\mathcal{H}'$ by writing $\mathcal{H}'$ in the form of (2.12) and identifying the coefficients of products of random variables. Three steps are involved in (2.18). First, we integrate out those $\phi_{k'}$ with k' between Λ/s and Λ. Second, we relabel the random variables by enlarging the wavevectors by a factor s. Third, we multiply all random variables by a constant factor α_s. The three trivial facts listed at the beginning of this section imply that P' is equivalent to P as far as random variables ϕ_k with $k < \Lambda/s$ are concerned and provided that proper relabeling and multiplying by α_s are done when averages are computed. For example,

$$\langle |\phi_{ik}|^2 \rangle_P = \alpha_s^2 \langle |\phi_{isk}|^2 \rangle_{P'}. \tag{2.19}$$

If we define $G(k, \mu) \equiv \langle |\phi_{ik}|^2 \rangle_P$, (2.19) says

$$G(k, \mu) = \alpha_s^2 G(sk, R_s \mu). \tag{2.20}$$

Note that the number of random variables in P' is smaller by a factor s^{-d} than that in P owing to the multiple integral in (2.18). The change of scale $k \to sk$ makes the density of points in k space smaller by the same factor. These simply mean that the volume of the system described by P' is $L'^d \equiv s^{-d} L^d$, i.e., shrunk by a factor s^{-d}. To identify μ' from $\mathcal{H}'$ given by (2.18), we must write $\mathcal{H}'$ in the form of (2.12) with L' replacing L; and the density of points in k space is now $L'^d (2\pi)^{-d}$. As was mentioned earlier, L' or L plays no role in calculating quantities of interest and is not included as a parameter. The set of R_s, $1 \leq s < \infty$, will be called the "renormalization group." We did not define the inverse of R_s; so it is not quite a group.

So far nothing has been said about the α_s in (2.18). The only role of α_s is in the last substitution in (2.18). If we have two successive transformations R_s and $R_{s'}$, then it is clear from (2.18) that they have the same effect as a single transformation $R_{ss'}$ except that the substitution is $\phi_k \to \alpha_s \alpha_{s'} \phi_{ss'k}$ not $\phi_k \to \alpha_{ss'} \phi_{ss'k}$. Thus, in order to observe

$$R_s R_{s'} \mu = R_{ss'} \mu \tag{2.21}$$

for any μ, we must demand

$$\alpha_s \alpha_{s'} = \alpha_{ss'}. \tag{2.22}$$

We shall so restrict our choice of α_s. Equation (2.22) is a severe restriction. It requires that

$$\alpha_s = s^y, \tag{2.23}$$

where y is a constant. If we regard the substitution

$$\phi_k \to s^y \phi_{sk} \tag{2.24}$$

in (2.18) as a scale change, then y can be interpreted as the dimension of ϕ_k in units of length. The dimension of ϕ_k can be defined by the microscopic Hamiltonian. However the dimension so defined is not useful. Instead, we shall determine y with respect to a *fixed point*.

A fixed point μ^* in the parameter space is that satisfying

$$R_s \mu^* = \mu^*. \tag{2.25}$$

It will play a major role in later discussions. Equation (2.25) may be viewed as an equation to be solved for μ^*. It is not expected to have a solution unless the y in $\alpha_s = s^y$ is properly chosen. This seems reasonable if we consider the case $s \to \infty$. We expect that all factors of s (and hence y) must delicately balance to achieve (2.25). In some sense (2.25) is an "eigenvalue equation" for the eigenvalue y and eigenvector μ^*. Of course, (2.25) is not a linear equation. We have no theorem so far to tell us whether (2.25) has a discrete, or continuous set of solutions, or even any solution at all. For the moment, we simply assume that there is at least one solution. We shall concentrate on a particular one with a definite y. We define the quantity η for this y:

$$y = 1 - \eta/2, \tag{2.26}$$

then

$$\alpha_s = s^{1-\eta/2}. \tag{2.27}$$

We shall identify η as a critical exponent later. Equation (2.20) now takes the form

$$G(k, \mu) = s^{2-\eta} G(sk, R_s \mu). \tag{2.28}$$

This formula will be used very often later.

More general correlation functions can be defined. For example, let

$$\begin{aligned}G_{i_1 i_2 \ldots i_m}(k_2, k_3 \ldots k_m, \mu) &\equiv \int d^d x_2 d^d x_3 \ldots d^d x_m \exp(-ik_2\cdot x_2 - \cdots - ik_m \cdot x_m) \\ &\quad \times \langle \phi_{i_1}(0)\phi_{i_2}(x_1)\ldots\phi_{i_m}(x_m)\rangle_P \\ &= L^{(d/2)m-d}\langle \phi_{i_1 k_1}\phi_{i_2 k_2}\cdots\phi_{i_m k_m}\rangle_P, \qquad (2.29)\end{aligned}$$

where $k_1 = -k_2 - k_3 - \cdots - k_m$ and none of the subsums of the k's is zero. It is easy to generalize (2.28) to

$$\begin{aligned}G_{i_1\ldots i_m}(k_2 \ldots k_m, \mu) &= s^{(m/2)(d+2-\eta)-d} G_{i_1 \ldots i_m}(sk_2 \ldots sk_m, R_s\mu) \qquad (2.30)\end{aligned}$$

provided that $k_1, k_2, \ldots, k_m < \Lambda/s$.

F. R_s as a refined scale transformation

The transformation R_s is basically a scale transformation. It tells how coupling parameters change when the system is shrunk by a factor s. However, the multiple integral and the determination of α_s by a fixed point equation make R_s very different from a naive change of scale. The multiple integral in (2.18) is necessary to keep the cutoff Λ fixed under R_s, i.e., it changes Λ to Λ/s and then lets the scale change bring Λ/s back to Λ. This is an extremely important point. The coupling parameters are defined with respect to a definite Λ. To compare two sets of coupling parameters, we must make sure that they are defined with respect to the same cutoff. Therefore, to define a sensible scale transformation, it is necessary to keep Λ fixed. The multiple integral is an unambiguous way. Thus, R_s can be viewed as a refined scale transformation keeping the cutoff fixed.

As was mentioned below (2.24), the quantity y can be interpreted as the dimension of ϕ_k in units of length. In (2.26) we have chosen $y = 1 - \frac{1}{2}\eta$ to be an interaction-dependent quantity based on the fixed point equation (2.25). Thus, the concept of dimension of a random variable under our refined scale transformation becomes an interaction dependent concept. We shall return to this point later.

G. Smoothed cutoff

The multiple integral in (2.18) implies a sharp cutoff in k space. That is to say for k immediately below Λ/s, ϕ_k is not integrated but it would be integrated if k is immediately above Λ/s. This sharp cutoff leads one to expect oscillating tails in the new coupling parameters of $\mathcal{H}'$ in the coordinate representation. This is analogous to the Friedel oscillation, which comes from the sharp Fermi surface, in the theory of Fermi gases. However unlike the Friedel oscillation, the oscillating tails here are of a purely mathematical origin and will lead to no physical consequence. It simply introduced complications in intermediate steps of calculation. It is desirable to remove the sharp cutoff by making the transition from "integrated" to "unintegrated" smooth. This can be done (see Ref. 1, Sec. XI) but is too complicated to be worth the effort here. In the graph representation to be introduced later, this can be done easily. What we want to point out here is that the fixed point μ^* will depend on how the cutoff is effected. This will become clearer in later discussions.

H. An important remark

Note that in the definition of R_s no reference is made to the average values that a probability distribution generates. In particular, whether $\langle\phi_i(x)\rangle$ vanishes or not is irrelevant in (2.18). The definition of R_s is separated from the concepts of averages, above or below critical point, etc. So far the concept of temperature simply has not entered. R_s simply takes one point in the parameter space to another.

I. Recursion formula and wavepacket variables

As will be evident later that the transformation of interest is R_s with large s. The usefulness of the renormalization group is not affected if we restrict s to

$$s = 2^l, \quad l = 0, 1, 2, 3, \cdots \qquad (2.31)$$

so that R_s is just applying R_2 l times:

$$R_s = (R_2)^l. \qquad (2.32)$$

One then works out $R_2\mu$ for a general μ. The result is the recursion formula of Wilson.[2] The renormalization group is then obtained by repeated applications of the recursion formula.

Note that regarding R_s as R_2 repeated l times is not just a change of terminology. It exhibits the two distinctive features of R_s of large s, i.e., first the transformation R_2 and second, the *repetitions*. It is the large number of repetitions that will be directly related to the singularities in critical behavior. R_2 is a completely nonsingular object. It is the "generator" of the renormalization group.

Separating the task of obtaining R_2 and that of repeating R_2 also allows some flexibility in computing and making approximations. For example, Wilson's approximate recursion formula for R_2 was obtained by using "wavepacket variables" as integration variables in the multiple integral of (2.18). We shall briefly sketch the basic idea, which can be generalized for other applications. The reader should consult Ref. 2 for details.

The random variable ϕ_k denotes the fluctuating amplitude of a plane wave configuration $\exp(ik\cdot x)$, which is spread over the whole volume. We expect $\mathcal{H}$ to be simpler when it is written in terms of more "localized" fluctuations because the interactions are assumed to be short-ranged. Thus, it should be useful to introduce the new variables (wavepacket variables)

$$\overline{\phi}(x_m) \equiv L^{-d/2} \sum_{1/2\Lambda<k<\Lambda} \phi_k \exp(ik\cdot x_m), \qquad (2.33)$$

where the points x_m form a lattice. The spacing between lattice points is such that the total number of variables $\overline{\phi}(x_m)$ is the same as the number of $\overline{\phi}_k$'s with k in the shell $\frac{1}{2}\Lambda < k < \Lambda$. The new variable $\overline{\phi}(x_m)$ represents the fluctuating amplitude of the wavepacket configuration

$$L^{-d} \sum_{1/2\Lambda<k<\Lambda} \exp[ik\cdot(x - x_m)] \qquad (2.34)$$

centered around x_m. This is the "most localized" configuration one can construct by superimposing plane waves of wave vectors in the shell $\frac{1}{2}\Lambda < k < \Lambda$. By smoothing the wavepacket and using $\bar{\phi}(x_m)$ as integration variables in (2.18), Wilson worked out an approximate formula for $R_2\mu$, which is suitable for numerical work and also as a basis for further approximations.

In the following sections, we shall always use R_s with arbitrary s and will make no use of the wavepacket variables. The above brief discussion is to point out some important features of the renormalization group which are more explicit in the recursion formula approach. For numerical investigation, the recursion formula approach is a powerful tool.

III. RENORMALIZATION GROUP NEAR THE FIXED POINT AND CRITICAL EXPONENTS

We shall now study R_s operating near a fixed point μ^* defined by $R_s\mu^* = \mu^*$ [see (2.25)]. The characteristics of critical phenomena will be related to those of R_s operating near μ^*.

So far our definition of R_s has been purely formal since we have not indicated how the multiple integral in (2.18) can be carried out, nor have we found a way to solve (2.25) for μ^* and η. Explicit illustrations will be given after we discuss the graphic representation of the renormalization group. In this section, our discussion will still be purely formal, and far from being rigorous. The validity of many assumptions and conclusions will not be evident till later sections.

A. The linearized equation

If μ is near μ^*, we write formally

$$\mu = \mu^* + \delta\mu, \tag{3.1}$$

where $\delta\mu$ is small in some sense. The equation $\mu' = R_s\mu$ can be written as

$$\delta\mu' = R_s^L \delta\mu \tag{3.2}$$

since $R_s\mu^* = \mu^*$, $\mu' \equiv \mu^* + \delta\mu'$. R_s^L becomes a linear operator when $O((\delta\mu)^2)$ terms are dropped in calculating $\delta\mu'$ from (3.2). In principle, at least, we can construct a matrix to represent R_s^L in (3.2); and we can determine the eigenvalues and eigenvectors of this matrix. Suppose that the eigenvalues are found to be $\lambda_j(s)$ and corresponding eigenvectors to be e_j, $j = 1, 2, 3, \ldots, \infty$. We label the eigenvalues in the order $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots$. Note that since $R_s R_{s'} e_j = R_{ss'} e_j$, we have

$$\lambda_j(s)\lambda_j(s') = \lambda_j(ss'),$$

$$\therefore\ \lambda_j(s) = s^{y_j}, \tag{3.3}$$

where y_j are constants and $y_1 \geq y_2 \geq y_3 \cdots$ since $s \geq 1$. We write $\delta\mu$ as a linear combination of the eigenvectors e_j:

$$\delta\mu = \sum_j t_j e_j; \tag{3.4}$$

then from (3.2)

$$\delta\mu' = \sum_j t_j s^{y_j} e_j. \tag{3.5}$$

Apparently, we have made no progress since we do not know y_j or e_j. But simplicity appears if it turns out that only $y_1 > 0$, all other y_j's are negative. In this case

$$\delta\mu' = R_s^L \delta\mu = t_1 s^{y_1} e_1 + O(s^{y_2}) \tag{3.6}$$

if s is so large that the first term dominates but $t_1 s^{y_1}$ is still small enough so that the linear approximation for R_s is valid. If $t_1 = 0$ to start with, then $R_s^L\delta\mu \to 0$ as s increases, i.e., μ will be "pushed" toward the fixed point by R_s. Wilson calls t_1 a "relevant" variable and other t_j's "irrelevant."[2]

We can imagine that the eigenvectors e_j span the linear vector space which is the neighborhood of μ^*. The subspace defined by $t_1 = 0$ will be called the "critical surface." Points on the critical surface will be pushed to the fixed point by R_s, and points not on the critical surface will be pushed toward e_1 but away from the fixed point as (3.6) indicates. (See Fig. 2.)

The linear approximation for R_s is expected to break down when μ, μ' are not very close to μ^*. But we expect the general picture of a critical surface and the approach to the e_1 axis of $R_s\mu$ for large s to remain valid.

B. Critical exponents and the correlation length

So far no physical concept has appeared in our discussion of the renormalization group. R_s simply transforms one probability distribution to another in a peculiar way. Now we shall examine the effect of R_s on the probability distribution (2.12), which describes fluctuations in a physical system at a definite temperature. This particular probability distribution is represented by a certain point $\mu(T)$ in the parameter space. This point corresponds to a set of coupling parameters which depend on the temperature T. They must be smooth functions of T. Because we have integrated out $\phi_{k'}$ with $k' > \Lambda$ in the microscopic Hamiltonian [see (2.12)], $H(\Lambda)$ would depend on T also. It is important to note that the integrations are over $\phi_{k'}$ with large k' and we would not expect any singular temperature dependence of $H(\Lambda)$ due to such integrals. If we vary T continuously, we would trace out a trajectory in the parameter space. This trajectory should be very smooth, and hits the critical surface at a special temperature T_c as shown in Fig. 2. At a temperature T which is very close to T_c and assuming $\mu(T)$ is close to μ^*, the distance from $\mu(T)$ to the critical surface, which is t_1, is then proportional to $T - T_c$.[11] Let us assume that $\mu(T)$ is close to μ^* write $\mu(T) = \mu^* + \delta\mu(T)$. Then (3.6) reads

$$R_s^L \delta\mu(T) = A(T - T_c)s^{1/\nu} e_1 + O(s^{y_2}) \tag{3.7}$$

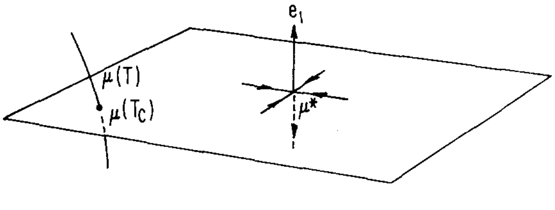


FIG. 2. Qualitative picture of a critical surface and a fixed point μ^* in the parameter space. The arrows point in directions of motion of $R_s\mu$ as s increases. The trajectory on the left is $\mu(T)$ for a continuous range of T, and $\mu(T_c)$ is the intersection of the trajectory and the critical surface.

where we have defined ν by

$$1/\nu = y_1, \tag{3.8}$$

and A is a constant. Applying (3.7) to (2.28), we obtain, for large s,

$$G(k, \mu(T)) = s^{2-\eta}[G(sk, \mu^* + A(T - T_c)s^{1/\nu}e_1 + O(s^{y_2}))]. \tag{3.9}$$

Consider first the case $T = T_c$. Since s is arbitrary, we choose s to be proportional to $1/k$, say $s = \Lambda/2k$. We then get from (3.9)

$$G(k, \mu(T_c)) = k^{-2+\eta}(\Lambda/2)^{2-\eta}[G(\Lambda/2, \mu^*) + O((\Lambda/2k)^{y_2})]. \tag{3.10}$$

In the limit of small k, this means

$$G(k, \mu(T_c)) \propto k^{-2+\eta}, \tag{3.11}$$

which is the equation defining the critical exponent η. Thus, the critical exponent η is related to the fixed point equation (2.23). The power law (3.11) for $G(k)$ at T_c is seen as a consequence of the fact that $R_s\mu(T_c)$ approaches μ^* for large s. How small must k be in order that (3.11) is a good approximation? Equation (3.10) says that $(2k/\Lambda)^{-y_2}$ must be small, much smaller than 1/2. say, i.e.,

$$2k/\Lambda \ll 2^{1/y_2}. \tag{3.12}$$

Equation (3.12) is an estimate of the size of the critical region *in k space*, namely the region in which (3.11) holds. This size therefore strongly depends on y_2. Recall that s^{y_2} is the second largest eigenvalue of R_s in the linear approximation, and y_2 is assumed to be negative.

Now we consider the case $T - T_c > 0$, $k = 0$. We choose $s = t_1^{-\nu}$. Here we write t_1 for $A(T - T_c)$. Equation (3.9) gives

$$G(0, \mu(T)) = t_1^{-(2-\eta)\nu}[G(0, \mu^* + e_1) + O(t_1^{-\nu y_2})]. \tag{3.13}$$

In the limit of small t_1, i.e., small $T - T_c$, we have

$$G(0, \mu(T)) \propto (T - T_c)^{-\gamma}, \tag{3.14}$$

$$\gamma = \nu(2 - \eta). \tag{3.15}$$

Equation (3.14) is the definition of the critical exponent γ. Equation (3.15) is a "scaling law" relating the exponents γ, η, and ν. Equation (3.14) holds when $t_1^{-\nu y_2}$ is much smaller than order unity, say $\frac{1}{2}$, as (3.13) indicates. This means

$$t_1 \ll 2^{1/\nu y_2}. \tag{3.16}$$

Similar to (3.12), (3.16) estimates the size of critical region in $T - T_c$. Equations (3.12) and (3.16) are oversimplified to exhibit the role of y_2. Many other parameters will in general enter in determining the size of the critical region. In other words, instead of $2^{1/y_2}$, we should have a complicated model dependent constant raised to the power $1/y_2$. The relevant question to answer for determining the size of the critical region is how large s must be so that $R_s\mu(T)$ is well approximated by $\mu^* + t_1 s^{1/\nu}e_1$. Intuitively, we expect that the farther away $\mu(T)$ is from μ^*, the larger an s is required, and hence the smaller the critical region becomes. This expectation is misleading sometimes, however. We shall have an opportunity to examine this point more explicitly later.

We now define the quantity ξ as

$$\xi = |t_1|^{-\nu} \tag{3.17}$$

which we shall call "correlation length." Then (3.7) reads

$$R_s^L \delta\mu = (s/\xi)^{1/\nu}e_1 + O(s^{y_2}). \tag{3.18}$$

The effect of R_s is thus to decrease the correlation length by a factor s. If we ignore the $O(s^{y_2})$ term, we would then arrive at the scaling hypothesis discussed in the Introduction. Thus the scaling hypothesis is valid if R_s, in its linear approximation near μ^*, is dominated by one eigenvalue for large s.

What about the case $T - T_c < 0$? In this case, $t_1 < 0$, we can simply set $s = (-t_1)^{-\nu}$ and replace (3.13) by

$$G(0, \mu(T)) = (-t_1)^{-\gamma}[G(0, \mu^* - e_1) + O((-t_1)^{-\nu y_2})]. \tag{3.19}$$

This is a correct statement but it turns out, for this case, to contain no information because $G(0, \mu) = \infty$ for $t_1 < 0$ as we shall discuss in Sec. VII. For other cases (see Sec. VIII on the free energy, for example), this kind of result may be useful.

The assumption that μ must be near μ^* can in fact be relaxed. The critical surface can be taken as a curved surface extending away from μ^*. Any μ on this surface has the property that

$$\lim_{s\to\infty} R_s\mu = \mu^*. \tag{3.20}$$

For s large enough, $R_s\mu$ will get into the neighborhood of μ^*, and the linear approximation will then apply. It is clear that if μ is not close to μ^* but is very close to the critical surface, then there is some range of s for which $R_s\mu$ is not far away from μ^*. There is no need to find all the eigenvalues and eigenvalues of R_s^L. All we need to know is $1/\nu$ and y_2, which should be regarded as specifying the leading s dependence of $R_s\mu$ for large s.

Therefore, the qualitative conclusion obtained in this section should hold for $\mu(T)$ close to the critical surface, i.e., for $T - T_c$ very small, but not necessarily close to the fixed point.

What we have described in this section is clearly a plausible conjecture not supported by any proof. In fact it is just one possible outcome. It is the simplest set of predictions that we expect from a renormalization group analysis. Many other possibilities exist. It may turn out that, besides fixing T at T_c, one has to fix something else to get on a critical surface. It might happen that there are more than one fixed point, or there are important complex eigenvalues for R_s in the linear approximation. Different possibilities in the behavior of R_s are expected to be consequences of different symmetry restrictions and other features of the parameter space. It is extremely desirable to have more rigorous work done to classify various possibilities. The difficulty is mathematical complication, not in principle. In principle, the renormalization group is well defined, and can always be carried out approximately by numerical means.

IV. REPRESENTATION BY GRAPHS

To demonstrate some basic features of the renormalization group defined above, the graph expansion is a very useful formal device. In particular, the multiple integral in (2.18) can be formally performed at the expense of introducing an infinite number of graphs. In the limiting cases where $\epsilon = 4 - d$ is small, or n is large, the graph expansion becomes useful for calculation as well.

A. Introducing graphs

The integrations over a virtually infinite number of random variables ϕ_k are very difficult except when most of these random variables are statistically independent, i.e., when P is a product of distributions each involving only one or two random variables. The graph expansion starts with separating $\mathcal{H}$ into two pieces

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1, \tag{4.1}$$

where

$$\mathcal{H}_0 = \tfrac{1}{2} \sum_{k,i} |\phi_{ki}|^2 G_0^{-1}(k) \tag{4.2}$$

is the $m=1$ term in (2.12). We shall use the symbol G_0^{-1} for u_2. The rest of $\mathcal{H}$ are included in $\mathcal{H}_I$. if $\mathcal{H}_I$ is ignored, $P \propto \exp(-\mathcal{H}_0)$ is a product of independent Gaussians since $\mathcal{H}_0$ is a sum of quadratic terms. Averages are easily computed. For example,

$$\begin{aligned}\langle\phi_{ik}\rangle_0 &= 0,\\ \langle\phi_{ik}\phi_{i-k}\rangle_0 &= \int d\phi_{ik}d\phi_{i-k}\{\exp[-|\phi_{ik}|^2/2G_0(k)]\}|\phi_{ik}|^2\\ &\quad\times\{\int d\phi_{ik}d\phi_{i-k}\exp[-|\phi_{ik}|^2/2G_0(k)]\}^{-1}\\ &= G_0(k),\end{aligned} \tag{4.3}$$

where $d\phi_{ik}d\phi_{i-k}$ means integrating over the complex ϕ_{ik} plane. Note that $\phi_{ik}^* = \phi_{i-k}$ [see (2.6)], so that ϕ_{ik} and ϕ_{i-k} are not independent complex variables. They are combinations of two real, independent random variables $\mathrm{Re}\phi_{ik}$ and $\mathrm{Im}\phi_{ik}$. We shall always denote such Gaussian averages by the subscript 0.

Now we write

$$\exp(-\mathcal{H}) = \exp(-\mathcal{H}_0)\exp(-\mathcal{H}_I), \tag{4.4}$$

and any average $\langle A\rangle$ over the full distribution becomes

$$\begin{aligned}\langle A\rangle &= \langle\exp(-\mathcal{H}_I)A\rangle_0/\langle\exp(-\mathcal{H}_I)\rangle_0\\ &= \sum_{n=0}^{\infty}\frac{(-)^n}{n!}\langle\mathcal{H}_I^n A\rangle_0\Big/\sum_{n=0}^{\infty}\frac{(-)^n}{n!}\langle\mathcal{H}_I^n\rangle_0.\end{aligned} \tag{4.5}$$

Let us assume that A, as well as $\mathcal{H}_I$, are sums of products of the ϕ_k's. Since the Gaussian average of a product of ϕ_k's is a product of pairwise averages [each ϕ_{ik} has to pair up with ϕ_{i-k} to give $\langle\phi_{ik}\phi_{i-k}\rangle_0 = G_0(k)$], the numerator and the denominator of (4.5) are complicated sums of products of $G_0(k)$'s. To introduce graphic representations, it is more convenient to use the random variables $\phi(x)$ [see (2.14)], instead of ϕ_k's, because the coordinate space is easier to visualize. The Gaussian average of a product of $\phi(x)$'s (with different x's in general) is a sum of products of pairwise averages since $\phi(x)$ is a linear combination of ϕ_k's. Each pair gives, writing $(2\pi)^{-d}\int d^dk$ for $L^{-d}\sum_{k<\Lambda}$,

$$\begin{aligned}\langle\phi_i(x)\phi_j(x')\rangle_0 &= (2\pi)^{-d}\int d^dkG_0(k)\exp[ik\cdot(x-x')]\delta_{ij}\\ &\equiv G_0(x-x')\delta_{ij},\end{aligned} \tag{4.6}$$

which can be represented by drawing a line between x and x'. Various averages can then be represented by graphs. As an illustration, suppose that

$$\begin{aligned}\langle A\rangle &= \langle\phi_i(y)\phi_i(0)\rangle = G(y),\\ \mathcal{H}_I &= (u_4/2)\int d^dx[\phi^2(x)]^2,\end{aligned} \tag{4.7}$$

where

$$\phi^2(x) \equiv \tfrac{1}{2}\sum_{i=1}^{n}\phi_i^2(x). \tag{4.8}$$

Then (4.5) is a power series in u_4. To zeroth order in u_4, we simply have $G(y) = G_0(y)$. To first order, we have an additional term

$$-u_4(\tfrac{1}{2}n+1)\int d^dx'G_0(y-x')G_0(x-x')G_0(x') \tag{4.9}$$

as represented by Fig. 3(a). We use a dashed line for u_4 only to separate the two $\phi^2(x)$ factors in (4.7). The second-order terms are given in Fig. 3(b). Those readers who are not familiar with graphs should write out the second-order terms explicitly. Note that disconnected graphs appear both in the numerator and in the denominator of (4.5). The net result is that only connected graphs contribute to $\langle A\rangle$. Note also that if A is of the form $A_1A_2\ldots A_m$ then there will be disconnected graphs of the form $\langle A_1A_2\ldots A_l\rangle\langle A_{l+1}\ldots A_m\rangle$ provided that neither of the two averages vanishes. The coordinate representation is useful only for visualization. In practice, the wavevector representation is more convenient. Any random variable A to be averaged over is regarded as a product of ϕ_k's. So are powers of $\mathcal{H}_I$ as given by (2.12). Every line in a graph will be labeled by a wavevector. The sum over wavevectors is now a well defined integral in k space. In each graph, those lines whose wavevectors are integrated over will be called *internal lines*. Those lines with wavevectors fixed by the ϕ_k's in A will be called *external lines*.

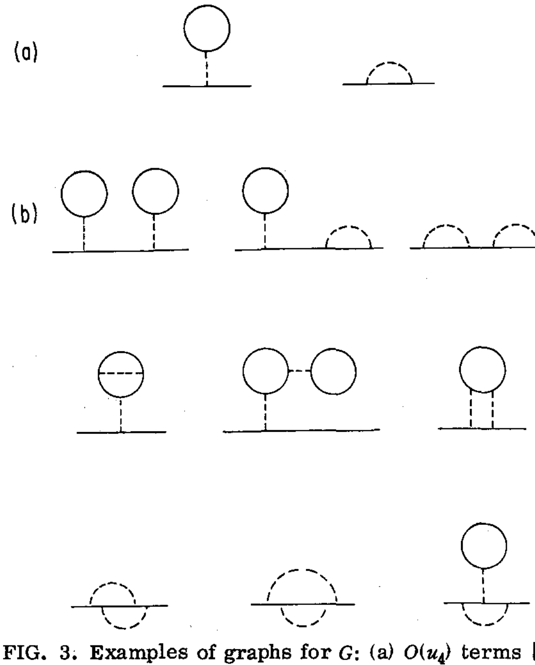


FIG. 3. Examples of graphs for G: (a) $O(u_4)$ terms [see (4.10)]; (b) $O(u_4^2)$ terms.

Of great importance is the "linked cluster theorem," which says that

$$\langle\exp(-\mathcal{H}_I)\rangle_0=\exp\langle[\exp(-\mathcal{H}_I)-1]\rangle_{0c} \tag{4.10}$$

where the subscript c denotes the sum of connected graphs only. The disconnected graphs are generated by exponentiation. The proof is left as an exercise in counting graphs.

A frequently occurring phrase is the "self-energy" Σ defined by

$$G^{-1}(k)=G_0^{-1}(k)+\Sigma(k). \tag{4.11}$$

The self-energy graphs are simply those graphs of $G(k)$ with G_0 lines of wavevector identical to k dropped.

What we have just gone through is the same as the Wick's theorem and Feynman graph expansion in field theory, if the time variable there is taken as imaginary and counted as a space dimension.

B. The multiple integral

The multiple integral in (2.18) is the first step in defining R_s. Let us denote those random variables to be integrated over by $\overline{\phi}$ and those not to be integrated by ϕ. To save writing, we shall introduce the notation

$$\int\delta\overline{\phi}\equiv\prod_{i,\Lambda/s<k'<\Lambda}\int d\phi_{ik'}. \tag{4.12}$$

We shall also write (2.12) as a sum:

$$\mathcal{H}\equiv\mathcal{H}(\phi)+\mathcal{H}(\phi,\overline{\phi}), \tag{4.13}$$

for the $\mathcal{H}$ in (2.18). Here $H(\phi)$ is the part depending only on ϕ, and $\mathcal{H}(\phi,\overline{\phi})$ depending on both ϕ and $\overline{\phi}$. More explicitly, $\mathcal{H}(\phi)$ is given by (2.12) with all wave vectors restricted to less than Λ/s, and $\mathcal{H}(\phi,\overline{\phi})$ is the rest. The graphic representation of $\int\delta\overline{\phi}\exp(-\mathcal{H})$ can be introduced as we did previously. Similar to (4.1) and (4.2), we write

$$\mathcal{H}(\phi,\overline{\phi})=\mathcal{H}_0(\overline{\phi})+\mathcal{H}_I(\phi,\overline{\phi}), \tag{4.14}$$

$$\mathcal{H}_0(\overline{\phi})\equiv\sum_{\Lambda/s<k<\Lambda}\sum_{i=1}^{n}|\phi_{ik}|^2G_0^{-1}(k). \tag{4.15}$$

We then define the Gaussian average by dropping $\mathcal{H}_I$ as before:

$$\langle A\rangle_{\bar{0}}\equiv\int\delta\overline{\phi}A\exp[-\mathcal{H}_0(\overline{\phi})]/\int\delta\overline{\phi}\exp[-\mathcal{H}_0(\overline{\phi})]. \tag{4.16}$$

The additional "bar" in the subscript of $\langle\cdots\rangle_{\bar{0}}$ denotes that the average is taken over the random variables $\overline{\phi}$. Then the multiple integral in (2.18) can be written as

$$\begin{aligned}\int\delta\overline{\phi}&\exp[-\mathcal{H}(\phi)-\mathcal{H}(\phi,\overline{\phi})]\\&=\exp[-\mathcal{H}(\phi)]\langle\exp[-\mathcal{H}_I(\phi,\overline{\phi})]\rangle_{\bar{0}}\\&\quad\times\int\delta\overline{\phi}\exp[-\mathcal{H}_0(\overline{\phi})].\end{aligned} \tag{4.17}$$

The last factor is a constant independent of ϕ. The average in the middle of (4.17) can be expanded and represented by graphs:

$$\begin{aligned}\langle\exp[-\mathcal{H}_I(\phi,\overline{\phi})]\rangle_{\bar{0}}&=\sum_{n=0}^{\infty}\frac{(-)^n}{n!}\langle\mathcal{H}_I^n(\phi,\overline{\phi})\rangle_{\bar{0}}\\&=\exp\langle(\exp[-\mathcal{H}_I(\phi,\overline{\phi})]-1)\rangle_{\bar{0}c},\end{aligned} \tag{4.18}$$

where the last line follows from the linked cluster theorem (4.11). Remember that $\overline{\phi}$ denotes the random variables $\phi_{k'}$, with $\Lambda/s<k'<\Lambda$. Thus, the internal lines in the graphs now have wave vectors ranging between Λ/s and Λ in magnitude, i.e., wave vectors in a "shell" in k space. Now we substitute (4.18) and (4.17) in (2.18), we obtain $\mathcal{H}'$ apart from an additive constant

$$\mathcal{H}'=[\mathcal{H}(\phi)-\langle(\exp[-\mathcal{H}_I(\phi,\overline{\phi})]-1)\rangle_{\bar{0}c}]_{\phi_k\to\alpha_s\phi_{sk}}. \tag{4.19}$$

This is then the graphic representation of (2.18).

C. The change of scale; R_s defined graphically

The new parameters $\mu'=(G_0'^{-1},u_4',u_6',\cdots)$ are now available in (4.19). For clarity, we shall extract μ' in two steps. First, let us write what is in the square bracket of (4.19) as

$$\begin{aligned}&\mathcal{H}(\phi)-\langle\exp[-\mathcal{H}(\phi,\overline{\phi})]-1\rangle_{\bar{0}c}\\&\quad=\sum_{i,k}|\phi_{ik}|^2(G_0^{-1}+\Sigma_s)+\sum_{m=2}^{\infty}L^{-(m-1)d}\sum_{k_1\cdots k_{2m-1}}\sum_{i_1\cdots i_{2m}}\\&\qquad\times\phi_{i,k_1}\cdots\phi_{i_{2m}k_{2m}}\overline{u}_{2m},\end{aligned} \tag{4.20}$$

i.e., we made an expansion in powers of unintegrated random variables. The wavevectors in (4.20) all have magnitude less than Λ/s. In terms of graphs, Σ_s is the self-energy, i.e., sum of all graphs (connected, of course) with two external lines, and $\overline{u}_{2m}$ is the sum of all graphs of $2m$ external lines. *All internal lines of these graphs have wave vectors in the shell $\Lambda/s<k'<\Lambda$, while all external lines have wave vectors restricted to $k<\Lambda/s$.*

The second step is to replace ϕ_k by $\phi_{sk}\alpha_s$ and write sL' for L, $s^{1-1/2\eta}$ for α_s [see (2.27)] in (4.20). We obtain

$$\begin{aligned}\mathcal{H}'=\sum_{i,k}|\phi_{ik}|^2G_0'^{-1}+\sum_{m=2}^{\infty}L'^{-(m-1)d}\sum_{k_1\cdots k_{2m-1}}\sum_{i_1\cdots i_{2m}}\\\times\phi_{i,k_1}\cdots\phi_{i_{2m}k_{2m}}u_{2m}',\end{aligned} \tag{4.21}$$

$$G_0'^{-1}=[G_0^{-1}(k/s)+\Sigma_s(k/s)]s^{2-\eta}, \tag{4.22}$$

$$u_{2m}'=\overline{u}_{2m}s^{-(m-1)d+m(2-\eta)}. \tag{4.23}$$

The quantity $\overline{u}_{2m}$ given by (4.20) of course depends on $k_1\cdots k_{2m-1}$. In (4.23), it is understood that they are replaced by $k_1/s\cdots k_{2m-1}/s$, like the k in (4.22). Now in (4.21) the wave vectors range from 0 to Λ in magnitude, but, as mentioned before, the density of points in k space is decreased from $L^d(2\pi)^{-d}$ to $L'^{-d}(2\pi)^{-d}$. We now have a system of a smaller volume.

From $\mu=(G_0^{-1},u_4,u_6,\cdots)$, which defines $\mathcal{H}$ via (2.12), we have arrived at (4.23) giving μ' by carrying out (2.18). We have thus established $\mu'=R_s\mu$ in terms of graphs.

D. The exponent η and self-energy

In Sec. II C we define η with respect to a fixed point μ^*. We shall now observe a simple relationship between η and the derivative of the self-energy at the fixed point. Since $R_s\mu^*=\mu^*$, it follows from (4.22) that

$$G_0^{*-1}(k)=[G_0^{*-1}(k/s)+\Sigma_s^*(k/s)]s^{2-\eta}. \tag{4.24}$$

We expand $G_0^{*-1}(k)$ in powers of k:

$$G_0^{*-1}(k)=t_0^*+r_1^*k^2+r_2^*k^4+\cdots. \tag{4.25}$$

We can always choose the unit of k such that $r_1^*=1$. Let us expand $\Sigma_s^*(k)$ also,

$$\Sigma_s^*(k)=\Sigma_s^*(0)+\left(\frac{\partial\Sigma_s^*}{\partial k^2}\right)_{k=0}k^2+\cdots \qquad (4.26)$$

and define

$$Z_s^{*-1}=1+\left(\frac{\partial\Sigma_s^*}{\partial k^2}\right)_{k=0}. \qquad (4.27)$$

Then (4.24) reads

$$t_0^*+k^2+\cdots=(t_0^*+\Sigma_s^*(0)+k^2s^{-2}Z_s^{*-1}+\cdots)s^{2-\eta}. \qquad (4.28)$$

Thus,

$$t_0^*=(t_0^*+\Sigma_s^*(0))s^{2-\eta}, \qquad (4.29)$$

$$Z_s^*=s^{-\eta}. \qquad (4.30)$$

Therefore, $\eta=0$ only if $Z_s^*=1$, i.e., if $\Sigma_s^*(k^2)$ is independent of k.

E. The case where $\langle\phi(x)\rangle\neq 0$

In this section we have assumed that $\langle\phi(x)\rangle=0$. If there is an external field or in the case μ gets below the critical surface ($t_1<0$), this would no longer be true, and the graphs will have some additional features which can be easily included.

V. THE FIXED POINT IN THE LARGE n LIMIT

So far our discussion has been abstract. Important conclusions in Sec. III are qualitative and not yet substantiated. In fact, no explicit example of R_s has been given. In the following sections, we shall illustrate all of what we have said about the renormalization group by explicit calculation for the case of large n. Our analysis will be exact in the large-n limit, i.e., terms neglected are of $O(1/n)$ compared to terms kept. Of course, our results will not constitute any general proof but will only serve as an example illustrating the ideas and qualitative conclusions explained before.

Our presentation might look somewhat unnatural to some reader, but it is designed to minimize mathematical complexity at the beginning. This section is devoted to the fixed point only. Everything else comes later.

We shall assume in this section and the next that $M\equiv\langle\phi(x)\rangle=0$ to avoid complication in discussion. As far as conclusions about the renormalization group is concerned, whether $M=0$ or not is irrelevant as we mentioned earlier. Our results on R_s in this section and the next will be valid both above and below the critical surface.

A. Generalization of the fixed point

Instead of solving the equation $R_s\mu^*=\mu^*$ for the fixed point μ^*, we shall show first that μ^* is easily generated by the limit

$$\lim_{s\to\infty}R_s\mu_1=\mu^* \qquad (5.1)$$

provided that μ_1 is on the critical surface. The reason that we use (5.1) to find μ^* is that a simple μ_1 can be found easily and the procedure itself serves to illustrate the simplifying features of the large n limit. Consider the probability distribution $P\propto\exp(-H)$,

$$H=\sum_{k,i}|\phi_{ik}|^2G_0^{-1}+L^{-d}\sum_{k_1k_2k}\sum_{ij}\phi_{ik_1}\phi_{ik_1-k}\phi_{jk_2}\phi_{jk_2+k^{u_4/8}}, \qquad (5.2)$$

where

$$G_0^{-1}=t_0+k^2 \qquad (5.3)$$

and t, u_4 are assumed to be constants. The point μ_1 in the parameter space defined by (2.12) is then

$$\mu_1=(G_0^{-1},u_4,0,0,0,\cdots). \qquad (5.4)$$

Note that we have changed the definition of u_4 in (2.12) slightly in including and 1/8 in (5.2). In terms of $\phi(x)$ defined by (2.14), the second term in (5.2) is simply

$$(u_4/2)\int d^dx[\phi^2(x)]^2, \qquad (5.5)$$

where

$$\phi^2(x)\equiv\tfrac{1}{2}\sum_i\phi_i^2(x). \qquad (5.6)$$

The reason for the factor $\frac{1}{2}$ in (5.6) is that in graphs a line can start from either factor of $\phi_i^2(x)$ thus requiring the multiplication by 2, and the factor $\frac{1}{2}$ then removes this 2. In the large n limit, graphs with the maximum number of closed loops dominate because every loop involves a summation over i and therefore a factor n. Figure 4 shows several graphs for the self-energy of order u_4^3. Graphs (a) and (b) are proportional to $u_4^3n^3$ while (c), (d), (e) are proportional to $u_4^3n^2$, u_4^3n, and u_4^3, respectively. Clearly, the dominating self-energy graphs of order u_4^l are proportional to $u_4^ln^l$ and are obtained from lower order ones by adding a loop whenever a dashed line is added. Graphs so generated are called "tree graphs." A tree graph can be separated into two disconnected pieces by removing one dashed line. We can choose units such that $u_4=O(1/n)$. This way, the tree graphs for the self energy are of $O(1)$. Other graphs are of $O(1/n)$ or smaller.

The most important simplifying feature of the tree graphs is that the self energy graphs are independent of the external wavevector. As a result, we have

$$\eta=0, \qquad (5.7)$$

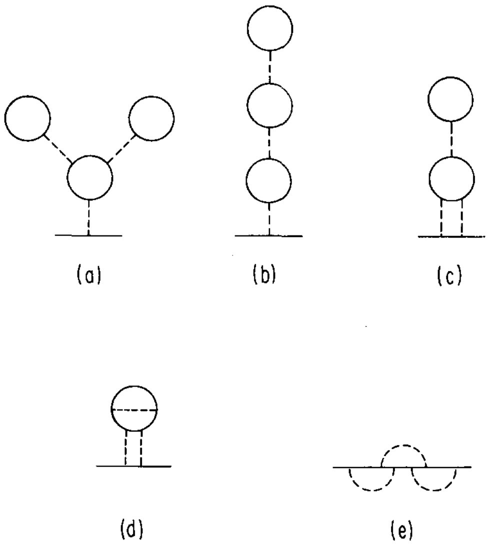


FIG. 4. Some self-energy graphs.

in the large n limit, according to (4.30) and (4.27). The tree graphs for the self-energy Σ are summed easily by solving the following equations:

$$G^{-1}(k,\mu_1)\equiv G^{-1}(k)$$
$$=t_0+\Sigma+k^2, \qquad (5.8)$$
$$\Sigma=u_4N, \qquad (5.9)$$
$$N=(2\pi)^{-d}\int d^dp\ G(p)$$
$$=(n/2)K_d\int_0^1 dp\, p^{d-1}(t_0+\Sigma+p^2)^{-1}, \qquad (5.10)$$

where

$$K_d\equiv 2^{1-d}\pi^{-d/2}/\Gamma(d/2), \qquad (5.11)$$

and we have set the cutoff $\Lambda=1$ to simplify writing.

Let $G'(k)$ denote $G(k,\mu_1')=G(k,R_s\mu_1)$. Since $\eta=0$, we have

$$G'^{-1}(k)=t_0'+\Sigma'(N')+k^2$$
$$=s^2G^{-1}(k/s), \qquad (5.12)$$

according to (2.28), and we defined

$$N'\equiv(n/2)K_d\int_0^1 dp\, p^{d-1}G'(p)$$
$$=(n/2)K_d\int_0^1 dp\, p^{d-1}s^{-2}G(p/s)$$
$$=(n/2)K_ds^{d-2}\int_0^{1/s} dp\, p^{d-1}G(p). \qquad (5.13)$$

It is important to note that $\Sigma'(N')$ is not u_4N'. It has the more complicated form[12]

$$\Sigma'(N')=\sum_{m=1}^{\infty}u'_{2(m+1)}N'^m, \qquad (5.14)$$

where $u'_{2(m+1)}$ will be determined in terms of u_4 later.

If μ_1 is on the critical surface, we have $G^{-1}(0)=0$, i.e., $G(p)=p^{-2}$, and

$$N=N_c=(n/2)K_d\int_0^1 dp\, p^{d-1}p^{-2}$$
$$=(n/2)K_d/(d-2), \qquad (5.15)$$
$$t_0+\Sigma=t_0+u_4N_c=0.$$

This gives the condition for μ_1 to be on the critical surface. The simplicity of (5.15) is peculiar to the large n limit, where $N=N_c$ for all μ on the critical surface. It will be clear later that $t_0+\Sigma=0$ implies (3.20). To determine t_0', we obtain from (4.22)

$$t_0'=(t_0+\Sigma_s)s^2,$$
$$\qquad (5.16)$$
$$\Sigma_s=u_4N_s,$$
$$N_s=\frac{n}{2}K_d\int_{1/s}^{1} dp\, p^{d-1}\frac{1}{t_0+\Sigma_s+p^2}\ . \qquad (5.17)$$

Equation (5.17) is the same as (5.10) except that p is restricted to $1/s<p<1$ in (5.17). Its solution gives the sum of tree graphs for the self-energy with internal line wavevectors so restricted. As $s\to\infty$, $G_0'^{-1}\to G_0^{*-1}=t_0^*+k^2$. Equation (5.17) is not convenient for taking the $s\to\infty$ limit. Let us subtract $1/p^2$ in the last integrand of (5.17) and add

$$(n/2)K_d\int_{1/s}^1 dp\, p^{d-1}p^{-2}=(n/2)K_d(1-s^{2-d})/(d-2) \qquad (5.18)$$

to balance the subtraction. We then obtain

$$s^{d-4}\left(\frac{N_s}{N_c}-1\right)s^2=-1+(d-2)\int_1^s dp\, p^{d-1}\left(\frac{1}{t_0'+p^2}-\frac{1}{p^2}\right)$$
$$\qquad (5.19)$$

after dividing (5.17) by $s^{2-d}N_c$. N_c is given by (5.15). From (5.15) and (5.16), we obtain

$$t_0'=(t_0+u_4N_s)s^2$$
$$=t_0(1-N_s/N_c)s^2. \qquad (5.20)$$

In the large s limit, $t_0'\to t_0^*$. Therefore, the left-hand side of (5.19) must approach zero as

$$s^{d-4}t_0'/t_0\to(t_0^*/t_0)s^{d-4}\to 0 \qquad (5.21)$$

as $s\to\infty$. Now the limit $s\to\infty$ for (5.19) is clear. We write

$$1+(d-2)(t_0^*/2)\Phi(-t_0^*,1,2-d/2)=0. \qquad (5.22)$$

We have expressed the p integral in (5.19) in terms of the transcendental function Φ [13]:

$$\int_1^\infty dp\, p^{d-1}\left(\frac{1}{t_0^*+p^2}-\frac{1}{p^2}\right)=-\frac{t_0^*}{2}\int_0^\infty dx\frac{\exp[-(1-d/2)x]}{e^x+t_0^*}\ ,$$
$$=-\frac{t_0^*}{2}\ \Phi\left(-t_0^*,1,2-\frac{d}{2}\right).$$
$$\qquad (5.23)$$

A useful series representation is[13]

$$\Phi(z,1,v)=\sum_{n=0}^{\infty}\frac{z^n}{n+v}\ . \qquad (5.24)$$

Equation (5.22) determines t_0^*. For $d=3$, it reduces to

$$1=(-t_0^*)^{1/2}\tanh(-t_0^*)^{1/2}, \qquad (5.25)$$

which implies $t_0^*=-0.69$. For large but finite s, (5.19) tells us how t_0' approaches t_0^*:

$$t_0'-t_0^*=s^{d-4}t_0^*(t_0^{-1}(d-2)^{-1}+(4-d)^{-1})$$
$$\times(\int_1^\infty dp\, p^{d-1}(t_0^*+p^2)^{-1})^{-1}+O(s^{d-6}). \qquad (5.26)$$

Note that we start from μ_1 to obtain (5.26). If we start from some other point on the critical surface, the s^{d-4} dependence will remain but the coefficient will be different as will be seen later.

We proceed to find the remaining of $\mu_1'=R_s\mu_1$, i.e., $u'_{2(m+1)}$, $m=1,2,3,\ldots$, and then take the limit $s\to\infty$ to obtain $u^*_{2(m+1)}$.

Figure 5(a) shows some graphs for u_4' to order u_4^3. The first two are proportional to $u_4^3n^2$ and are the dominating ones. Figure 4(b) shows some graphs for u_6' to order u_4^5. The dominating ones are proportional to $u_4^5n^3$. Generalization is clear: To order u_4^l, the dominating graphs for $u'_{2(m+1)}$ are proportional to $u_4^ln^{l-m}$. Since $u_4=O(1/n)$, we have,

$$u'_{2(m+1)}=O(n^{-m}). \qquad (5.27)$$

The dominating graphs are again tree graphs. They are just those for Σ_s with m pairs of external lines replacing m loops at ends of m branches. Equation (5.27) also implies that

$$u^*_{2(m+1)}=O(n^{-m}). \qquad (5.28)$$

Now the task is to sum tree graphs for $u'_{2(m+1)}$. Unlike the graphs for Σ_s, those for $u'_{2(m+1)}$ depend on wave vec-

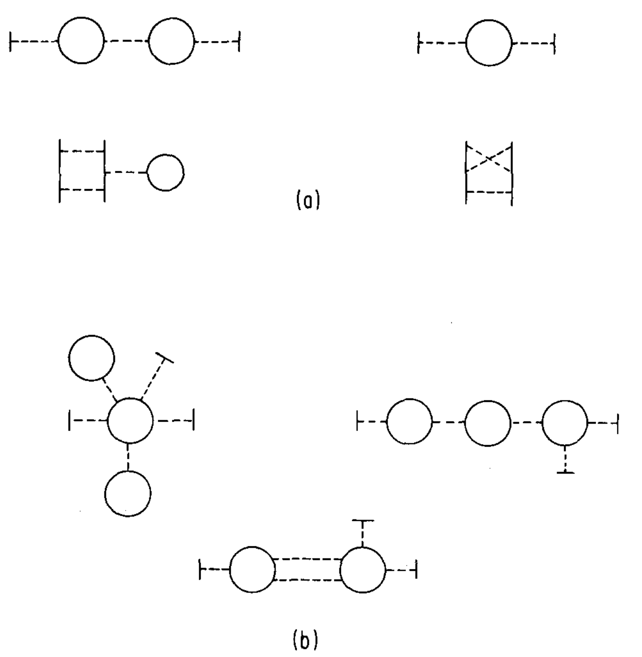


FIG. 5. (a) Some graphs for u_4' of $O(u_4^3)$; (b) some graphs for u_6' of $O(u_4^5)$.

tors of external lines, more precisely, on the wave vectors of pairs. This wave vector dependence is weak and is unimportant unless the wave vector of some pair gets close to the cutoff. In computing the self-energy $\Sigma'(N')$, each pair is closed to form a loop so that only the value of $u'_{2(m+1)}$ evaluated at zero wave vector is relevant. We shall ignore the wave vector dependence and regard $u'_{2(m+1)}$ as a constant. This means we regard $u'_{2(m+1)}$ as a local coupling, corresponding to a term

$$u'_{2(m+1)}\int d^dx(\phi^2(x))^{m+1}/(m+1) \tag{5.29}$$

in $\mathcal{H}'$. The numerical factor $(m+1)^{-1}$ is designed to make $\Sigma'(N')$ look simple [see (5.14)].

To determine $u'_{2(m+1)}$, we shall compute $\Sigma'(N')$ and identify the coefficient of N'^m, instead of counting graphs directly. Let us write the self-energy (5.9) as

$$\Sigma = u_4N = u_4(N_a+N_b), \tag{5.30}$$

where we define N_a, N_b by breaking up the integral in (5.10):

$$N_a=(n/2)K_d\int_{1/s}^{1}dp\,p^{d-1}G(p), \tag{5.31}$$

$$N_b=(n/2)K_d\int_0^{1/s}dp\,p^{d-1}G(p)=s^{2-d}N'. \tag{5.32}$$

The last equality in (5.32) follows from (5.13). Since $G(p)=(t_0+u_4(N_a+N_b)+p^2)^{-1}$, N_a clearly depends on N_b. So, let us write (5.31) as

$$\begin{aligned}N_a&=N_a(N_b)\\&=\frac{n}{2}K_d\int_{1/s}^{1}dp\,p^{d-1}\frac{1}{t_0+u_4(N_a(N_b)+N_b)+p^2}\,.\end{aligned} \tag{5.33}$$

Note that

$$N_a(0)=N_s \tag{5.34}$$

since we get (5.17) by setting $N_b=0$ in (5.33). From (5.14) and (5.12), we obtain

$$\begin{aligned}G'^{-1}(0)&=t_0'+\Sigma'(N')=t_0'+\sum_{m=1}^{\infty}u'_{2(m+1)}N'^m\\&=s^2G^{-1}(0)=s^2(t_0+u_4(N_a+N_b))\\&=s^2(t_0+u_4(N_a(N's^{2-d})+N's^{2-d})),\end{aligned} \tag{5.35}$$

where we have written $N's^{2-d}$ for N_b according to (5.32). We now expand the last line of (5.35) in powers of N', and identify the coefficient of N'^m as $u'_{2(m+1)}$:

$$u'_{2(m+1)}=\frac{1}{m!}\,s^{2+m(2-d)}\left[\frac{d^m}{dN_b^m}\,\Sigma\right]_{N_b=0}. \tag{5.36}$$

Here Σ means $u_4(N_a(N_b)+N_b)$ with $N_a(N_b)$ defined by (5.33). It is easy to visualize (5.36) in terms of graphs. Differentiation with respect to N_b means converting a close loop representing N_b to a pair of external lines. The self energy has one pair of external lines. In (5.36), we open up m loops to get a total of $2(m+1)$ external lines. The factor $s^{2+m(2-d)}$ is in accordance with (4.23). A neater formula can be obtained as follows. From (5.33) we obtain

$$\begin{aligned}N's^{2-d}&=N_b=N-N_a\\&=N-\frac{n}{2}K_d\int_{1/s}^{1}dp\,p^{d-1}\frac{1}{t_0+\Sigma+p^2}\,,\end{aligned} \tag{5.37}$$

which is equivalent to the statement

$$\begin{aligned}\lambda\equiv\frac{N'}{N_c}&=1+s^2\left(\frac{N(\lambda)}{N_c}-1\right)s^{d-4}\\&\quad-(d-2)\int_1^{s}dp\,p^{d-1}\left(\frac{1}{t'(\lambda)+p^2}-\frac{1}{p^2}\right),\end{aligned} \tag{5.38}$$

where t' is $s^2(t_0+u_4N)$. This equation is taken as defining N and t' as functions of λ. We shall always understand the symbol t' as $t'(\lambda)$. Equation (5.36) now takes the form

$$u'_{2(m+1)}=N_c^{-m}\frac{1}{m!}\left(\frac{d^mt'}{d\lambda^m}\right)_0. \tag{5.39}$$

The subscript 0 means setting $\lambda=0$. In particular, $t_0'=t'(0)$. By setting $\lambda=0$ in (5.38), we obtain (5.19). This is expected since setting $N'=0$ means $N_b=0$ and hence $N=N_a=N_s$ [see (5.34)]. Equation (5.38) provides a mathematical device for summing the tree graphs for $u'_{2(m+1)}$. It is important that N and t' in (5.38) must be taken as functions of λ, not as fixed quantities determined by (5.9). To sum up, we have carried out R_s on μ_1 and obtained μ_1'. Equations (5.38) and (5.39) give $u'_{2(m+1)}$. By setting $\lambda=0$ in (5.38), t_0' is determined.

Now $u^*_{2(m+1)}$ is readily obtained by taking the limit $s\to\infty$ of (5.38). The term $s^2(N(\lambda)/N_c-1)s^{d-4}$ vanishes in this limit since $u_4=-t_0/N_c$, $t'=s^2(t_0+u_4N(\lambda))$ so that

$$s^2(N(\lambda)/N_c-1)s^{d-4}=-(t'/t_0)s^{d-4}. \tag{5.40}$$

Therefore the limit $s\to\infty$ of (5.38) is

$$\begin{aligned}\lambda&=1+(d-2)t^*\int_1^{\infty}dp\,p^{d-3}(t^*+p^2)^{-1}\\&=1+(d/2-1)t^*\Phi(-t^*,1,2-d/2),\end{aligned} \tag{5.41}$$

where t^* means $t^*(\lambda)$. The fixed point μ^* is then given by

$$t_0^*=t^*(0)+O(n^{-1}), \tag{5.42}$$

$$u^*_{2(m+1)} = N_c^{-m}\frac{1}{m!}\left(\frac{d^m t^*}{d\lambda^m}\right)_0 + O(n^{-m-1}). \qquad (5.43)$$

For finite but very large s, (5.38) tells us how fast $R_s\mu_1$ approaches μ^*. We obtain from (5.38)

$$(t' - t^*)(d-2)\int_1^\infty dp\, p^{d-1}(t^* + p^2)^{-2}$$
$$= t^*(t_0^{-1} + (d-2)/(4-d))s^{d-4} + O(s^{d-6}). \qquad (5.44)$$

We thus observe the s^{d-4} behavior as we did before.

B. Some simple features of μ *

Let us examine the exact results (5.41)–(5.43) more closely. The probability distribution represented by μ^* is $P^* \propto \exp(-H^*)$ with

$$H^* = \int d^dx\left((\nabla\phi)^2 + t_0^*\phi^2(x) + \sum_{m=1}^{\infty} u^*_{2(m+1)}\frac{[\phi^2(x)]^{m+1}}{m+1}\right) \qquad (5.45)$$

in the coordinate representation. We repeat some definitions here:

$$\phi_i(x) \equiv L^{-d/2}\sum_{k<1}\phi_{ik}\exp(ik\cdot x),$$
$$\phi^2(x) \equiv \tfrac{1}{2}\sum_{i=1}^{\infty}[\phi_i(x)]^2, \qquad (5.46)$$
$$(\nabla\phi)^2 \equiv \tfrac{1}{2}\sum_{i=1}^{n}[\nabla\phi_i(x)]^2.$$

The gradient term and t_0^* term of course comes from $G_0^{*-1} = t_0^* + k^2$ and the form of the other terms has been discussed before [see between (5.28) and (5.29)]. In view of (5.42) and (5.43), we can write (5.45) in the simple form by summing over m:

$$H^* = \int d^dx[(\nabla\phi)^2 + U^*(\phi^2(x))], \qquad (5.47)$$

where the function U^* is defined as

$$U^*(\phi^2) = N_c\int_0^{\phi^2/N_c} d\lambda\, t^*(\lambda). \qquad (5.48)$$

In other words, instead of specifying μ^* by an infinite set of parameters, we can represent it here by a real function U^*. Some qualitative features of $t^*(\lambda)$ and $U^*(\phi^2)$ can be obtained easily. It is clear from (5.41) that λ ranges from $-\infty$ to ∞ as t^* ranges from -1 to ∞. The integral is well defined for $t^* > -1$. Also, we have

$$t^*(1) = 0 \qquad (5.49)$$

and for $\lambda \to \infty$, we have

$$\lambda = 1 + \Gamma(d/2)\Gamma(2-(d/2))t^{*(d/2)-1} + O(t^{*(d/2)-2}), \qquad (5.50)$$

which means

$$t^* = \left(\frac{\lambda}{\Gamma(d/2)\Gamma(2-d/2)}\right)^{2/(d-2)}[1 + O(\lambda^{-2/(d-2)})] \qquad (5.51)$$

for large λ. According to (5.48), we obtain

$$U^*(\phi^2) = N_c(d-2)/(d\Gamma(d/2)\Gamma(2-d/2))$$
$$\times(\phi^2/N_c)^{d/(d-2)}[1 + O((\phi^2/N_c)^{-2/(d-2)})], \qquad (5.52)$$

for large ϕ^2/N_c. With the information (5.49), (5.51), (5.52), and $t^*(-\infty) = -1$, the general shape of the curve t^* vs λ and that of $U^*(\phi^2)$ vs ϕ^2 can be sketched. (See Figs. 1 and 6.) For the case $d=3$, (5.41) is expressible in terms of elementary functions. We have

$$\lambda = 1 - (a/2)\ln[(1+a)/(1-a)], \quad \text{for } -1 < t^* < 0,$$
$$= 1 + a\tan^{-1}a, \quad \text{for } t^* > 0, \qquad (5.53)$$
$$a \equiv (|t^*|)^{1/2}.$$

Thus, for $d=3$,

$$U^*(\phi^2) \propto \phi^6, \qquad (5.54)$$

for large ϕ^2/N_c.

The curves for $U^*(\phi^2)$ vs ϕ^2 are flatter as $d\to 4$ and steeper as $d\to 2$. They all have the general shape of that obtained by Wilson[2] through numerical work for $d=3$. In particular a ϕ^6 behavior was observed in Ref. 2 also for large ϕ^2.

C. Limit of small ϵ

Since our results are valid for arbitrary d between 2 and 4, we can easily extract the small $\epsilon \equiv 4-d$ limit. It is convenient to use the series representation (5.24) for Φ in (5.41):

$$\lambda = 1 + \left(1 - \frac{\epsilon}{2}\right)t^*\sum_{n=0}^{\infty}\frac{(-t^*)^n}{n+\epsilon/2}. \qquad (5.55)$$

The $n=0$ term in the series dominates, when ϵ is small. If we keep only this term, we get

$$t^* = (\epsilon/2)(\lambda - 1) + O(\epsilon^2), \qquad (5.56)$$

which means

$$t_0^* = -(\epsilon/2) + O(\epsilon^2), \qquad (5.57)$$
$$u_4^* = N_c^{-1}\epsilon/2 + O(\epsilon^2)$$
$$= 16\pi^2\epsilon/n + O(\epsilon^2), \qquad (5.58)$$

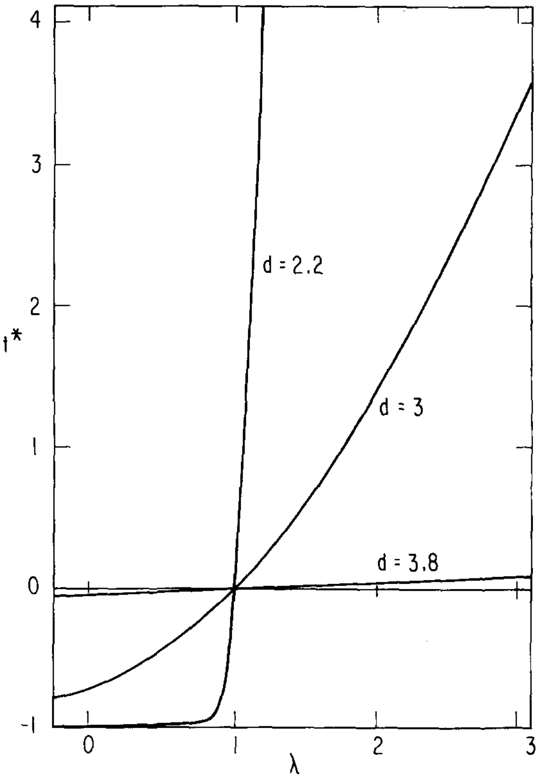


FIG. 6. Plot of t^* vs λ for $d=2.2$, 3, and 3.8. See (5.41). Note that U^* is obtained from t^* by integration. See (5.48).

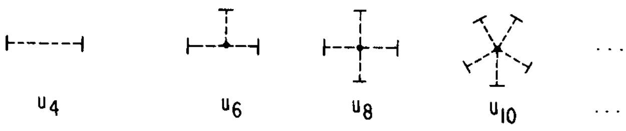


FIG. 7. Graph representation of $u_4, u_6, u_8, \cdots$. See (6.1)—(6.4).

where N_c is evaluated at $d=4$ via (5.15). These are the large n limits of well known results.[3] The important point is that $u^*_{2(m+1)}$ with $m>2$ do not appear to this order. It is easy to obtain from (5.55) that

$$u^*_{2(m+1)} = \tfrac{1}{2}\epsilon(-16\pi^2\epsilon/n)^m/(m-1) + O(\epsilon^{m+1}), \tag{5.59}$$

for $m>1$.

VI. AWAY FROM THE FIXED POINT

In the previous section we determined $R_s\mu_1$ for a special μ_1 [see (5.2)—(5.4)]. The reason for picking μ_1 instead of an arbitrary μ is that μ_1 has a very simple structure so that we could avoid mathematical complications which would have covered up the basic features of interest. However, these complications will no longer be any problem once we have become familiar with the basic features illustrated in the previous section. We shall now generalize the previous results and determine $\mu'=R_s\mu$ in the large n limit for any μ of the form

$$\mu = (t_0, u_4, u_6, \cdots). \tag{6.1}$$

Note that we write simply t_0 instead of $G_0^{-1} = t_0 + k^2$ in (6.1) since the k^2 term never changes under R_s for our cases. All entries in (6.1) are taken as constants. Thus, μ specifies a probability distribution $P \propto \exp(-\mathcal{H})$ with

$$\mathcal{H} = \int d^dx((\nabla\phi)^2 + U(\phi^2)), \tag{6.2}$$

$$U(\phi^2) = t_0\phi^2 + \sum_{m=1}^{\infty} u_{2(m+1)}(\phi^2)^{m+1}/(m+1). \tag{6.3}$$

It is understood that $t_0 = O(1)$, $u_{2(m+1)} = O(n^{-m})$. After $\mu'=R_s\mu$ is determined, various consequences will be discussed.

A. The transformation $\mu' = R_s\,\mu$

The graphs representing $u_4, u_6, u_8 \cdots$ are shown in Fig. 7. This multitude of coupling parameters is the major complication, which we avoided in the previous section by dropping all but u_4. However, the approach we took in the previous section was so designed that its generalization to include the new coupling parameters is straightforward. In fact, we can simply copy the formulas there and replace $\Sigma = u_4N$ there by the more complicated self-energy

$$\Sigma(N) = \sum_{m=1}^{\infty} u_{2(m+1)}N^m. \tag{6.4}$$

The tree graphs here are taken to be those that fall into two disconnected pieces when a dashed line is cut. The tree graphs for the self-energy is summed by solving the equations

$$N = (n/2)K_d \int_0^1 dp\, p^{d-1}G(p), \tag{6.5}$$

$$G(p) \equiv G(p,\mu) = [t_0 + \Sigma(N) + p^2]^{-1}, \tag{6.6}$$

together with (6.4), in the same way as solving (5.8)—(5.10). Let $G'(k)$ denote $G(k,\mu') = G(k, R_s\mu)$ and let N' and $\Sigma'(N')$ be defined by (5.13) and (5.14). We now write

$$N = N_a + N_b = N_a + N's^{2-d}, \tag{6.7}$$

with N_a, N_b defined by (5.31) and (5.32) with $G(p)$ given by (6.6). The equality $N_b = N's^{2-d}$ still holds. The generalization of (5.33) is now

$$N_a(N_b) = (n/2)K_d \int_{1/s}^1 dp\, p^{d-1}[t_0 + \Sigma(N_a(N_b) + N_b) + p^2]^{-1}. \tag{6.8}$$

Since $s^2G^{-1}(0) = G'^{-1}(0)$, we have, similar to (5.35),

$$\begin{aligned} t_0' + \Sigma'(N') &= t_0' + \sum_{m=1}^{\infty} u_{2(m+1)}N'^m \\ &= s^2[t_0 + \Sigma(N_a(N's^{2-d}) + N's^{2-d})]. \end{aligned} \tag{6.9}$$

Provided that we use the new $\Sigma(N)$, the equations (5.36)—(5.39) remain unchanged and μ' is thereby determined.

We summarize the four steps of obtaining $\mu' = R_s\mu$:

(a) Given $\mu = (t_0, u_4, u_6, \cdots)$ we construct

$$\Sigma(N) = \sum_{m=1}^{\infty} u_{2(m+1)}N^m \tag{6.10}$$

as a function of N.

(b) Define t' as

$$t' = s^2(t_0 + \Sigma(N)). \tag{6.11}$$

(c) Solve for N and hence t' via (6.11) as *functions of* λ, $t'(\lambda)$ and $N(\lambda)$ from the equation

$$\lambda = 1 + s^2(N(\lambda)/N_c - 1)s^{d-4} + (d-2)t' \int_1^s dp\, p^{d-3}(t' + p^2)^{-1}. \tag{6.12}$$

(d) Obtain $R_s\mu = \mu' = (t_0', u_4', u_6', \cdots)$ from

$$t_0' = t'(0), \tag{6.13}$$

$$u'_{2(m+1)} = N_c^{-m}\frac{1}{m!}\left(\frac{d^mt'}{d\lambda^m}\right)_0. \tag{6.14}$$

These four steps define $\mu' = R_s\mu$ regardless of whether μ, μ' are above, on, or below the critical surface.

We can express $R_s\mu = \mu'$ as the transformation from $U(\phi^2)$ to $U'(\phi^2)$. U is related to μ via (6.3). Write

$$t \equiv t(\lambda) \equiv t_0 + \Sigma(\lambda N_c); \tag{6.15}$$

then

$$U(\phi^2) = N_c\int_0^{\phi^2/N_c} t(\lambda)\, d\lambda, \tag{6.16}$$

and

$$U'(\phi^2) = N_c\int_0^{\phi^2/N_c} t'(\lambda)\, d\lambda, \tag{6.17}$$

with t' obtained from (6.12). Of course, $\mu' = R_s\mu$ can also be viewed as the transformation from t to t'.

B. The critical surface and the fixed point

We assert that the critical surface is given by the condition $G^{-1}(0,\mu) = 0$, i.e.,

$$t_1 \equiv t(1) = t_0 + \Sigma(N_c)$$

$$= t_0 + \sum_{m=1}^{\infty} u_{2(m+1)} N_c^m$$
$$= 0. \qquad (6.18)$$

This equation is a linear equation in $(t_0, u_4, u_6, \cdots)$ and defined a hyperplane in the parameter space. We need to prove that (6.18) implies (3.20), that is, if $t_1 = 0$ then $R_s\mu$ approaches μ^* as $s \to \infty$. Let us define

$$\zeta \equiv \zeta(\lambda) \equiv s^2(N(\lambda)/N_c - 1) \qquad (6.19)$$

which appears in the second term of (6.12). Following the notation (6.15), we write (6.11) as

$$t' = s^2[t_0 + \Sigma(N_c(1 + \zeta/s^2))]$$
$$= s^2 t(1 + \zeta/s^2). \qquad (6.20)$$

If μ satisfies (6.18), then (6.20) implies that for large s

$$t' = \zeta\left(\frac{dt}{d\lambda}\right)_1 + O(s^{-2}), \qquad (6.21)$$

where the subscript 1 means $\lambda = 1$. Thus, for $s \to \infty$, ζ is proportional to t', the second term of (6.12) thus vanishes as $s \to \infty$ and the equation (5.41) for the fixed point is obtained. Thus, our assertion has been proved.

C. Behavior of R_s μ for finite but large s

In view of our discussion in Sec. III, what is relevant to the theory of critical phenomena is the behavior of $R_s\mu$ for large s when μ is on the critical surface or very close to the critical-surface.

Let us begin by writing (6.12) as

$$\lambda = 1 + (d-2)t' \int_1^\infty dp\, p^{d-3}(t' + p^2)^{-1}$$
$$+ s^{d-4}[\zeta - (d-2)t' \int_1^\infty dp\, p^{d-3}(t'/s^2 + p^2)^{-1}], \qquad (6.22)$$

where ζ is related to t' via (6.20). Clearly, if $t' = t^*$, (6.22) is satisfied without the last square bracket $s^{d-4}[\cdots]$ in view of (5.41). This square bracket vanishes, too, if $t' = t^*$. This statement is just saying that $R_s\mu^* = \mu^*$ for all $s \geq 1$, and is easy to see as follows. Since (5.41) is true for any λ, we can set $\lambda = 1 + \zeta^*/s^2$, where ζ^* is obtained from (6.20) by setting $t' = t^*$, $t(1 + \zeta/s^2) = t^*(1 + \zeta^*/s^2)$. We then obtain

$$1 + \zeta^*/s^2 = 1 + (d-2)(t^*/s^2) \int_1^\infty dp\, p^{d-3}(t^*/s^2 + p^2)^{-1}, \qquad (6.23)$$

where we have written t^*/s^2 for $t^*(1 + \zeta^*/s^2)$. Clearly, (6.23) says that the square bracket of (6.22) must vanish if $t' = t^*$. In view of (5.41) and (6.23) we can write (6.22) as

$$-(d-2)\int_1^\infty \left(\frac{1}{t' + p^2} - \frac{1}{t^* + p^2}\right) p^{d-1}\, dp$$
$$= -s^{d-4}\left[\zeta - \zeta^* - s^2(d-2)\int_1^\infty dp\, p^{d-1} \frac{1}{t'/s^2 + p^2}\right.$$
$$\left. - \frac{1}{t^*/s^2 + p^2})\right]. \qquad (6.24)$$

Let us consider first the case where μ is on the critical surface, i.e., $t_1 = 0$ [see (6.18)]. We then have, from (6.20),

$$t' = \left(\frac{dt}{d\lambda}\right)_1 \zeta(1 + O(\zeta/s^2)), \qquad (6.25)$$

$$t^* = \left(\frac{dt^*}{d\lambda}\right)_1 \zeta^*(1 + O(\zeta^*/s^2)). \qquad (6.26)$$

Thus,

$$(\zeta - \zeta^*)\left(\frac{dt^*}{d\lambda}\right)_1 = t' - t^* - \zeta\left(\frac{d(t - t^*)}{d\lambda}\right)_1 + O\left(\frac{\zeta}{s^2}\right). \qquad (6.27)$$

Substituting (6.27) in (6.24), we obtain, for large s

$$(t' - t^*)[\int_1^\infty dp\, p^{d-1}(t^* + p^2)^{-2} + O(s^{d-4})]$$
$$= s^{d-4}\zeta\left(\frac{d(t - t^*)}{d\lambda}\right)_1 \Big/ \left(\frac{dt^*}{d\lambda}\right)_1 (d-2)(1 + O(s^{-2})). \qquad (6.28)$$

It is then obvious that $\mu' = R_s\mu$ approaches μ^* as fast as s^{d-4}.

In Sec. III, the probability distribution for our spin system at critical temperature is represented by a point $\mu(T_c)$ on the critical surface. What (6.28) shows is that $R_s\mu(T_c)$ will approach the fixed point. No interesting result appears for $G(k, \mu(T_c))$ since we already know that $G(k, \mu) = k^{-2}$ if μ is on the critical surface because only tree graphs have been included.

Next, we turn to the case where μ is not on the critical surface, i.e., $t_1 = t(1) \neq 0$. However, we shall assume that t_1 is very small. While s will be taken as a large number, we still want t' to be of $O(1)$. This is always possible by making t_1 small. Just how small will be clear later.

Similar to (6.25), we obtain from (6.20)

$$t' = s^2 t_1 + \left(\frac{dt}{d\lambda}\right)_1 \zeta\left(1 + O\left(\frac{\zeta}{s^2}\right)\right). \qquad (6.29)$$

We can still solve (6.20) for ζ by iteration even though $s^2 t_1$ and hence ζ can become very large, because ζ/s^2 is proportional to t_1 which is assumed to be small. The term $O(\zeta/s^2)$ in (6.29) will therefore be small, and we have

$$\zeta - \zeta^* = -s^2 t_1 \left(\frac{dt}{d\lambda}\right)_1 (1 + O(t_1)) + t' \Big/ \left(\frac{dt}{d\lambda}\right)_1 - t^* \Big/ \left(\frac{dt^*}{d\lambda}\right)_1 + O(s^{-2}). \qquad (6.30)$$

Substituting (6.30) in (6.24), we obtain

$$-(d-2)\int_1^\infty dp\, p^{d-1}\left(\frac{1}{t' + p^2} - \frac{1}{t^* + p^2}\right)$$
$$= s^{d-2} t_1 \Big/ \left(\frac{dt}{d\lambda}\right)_1 (1 + O(t_1)) + O(s^{d-4}). \qquad (6.31)$$

It is now clear that our assumption that $t' = O(1)$ is valid if we take $s^{d-2}t_1$ to be of $O(1)$. Now we want to solve (6.31) for t'. It is sufficient for our purpose to know that the solution is of the form

$$t' = F(z; \lambda), \qquad (6.32)$$

$$z \equiv s^{d-2} t_1(1 + O(t_1)) + O(s^{d-4}), \qquad (6.33)$$

where $F(z;\lambda)$ is some complicated function with no explicit dependence on s or t_1. Of course, if $t_1 = 0$, (6.28) should be recovered.

Let us write the equation $G(k, \mu) = G(sk, R_s\mu)s^{2-\eta}$ as

$$G(k, t) = s^2 G(sk, t'). \qquad (6.34)$$

Then we have

$$G(0,t)=s^2G(0,F(s^{d-2}t_1(1+O(t_1))+O(s^{d-4});\lambda)). \qquad (6.35)$$

Setting $s=|t_1|^{-1/(d-2)}$, we obtain

$$G(0,t)=|t_1|^{-2/(d-2)}G(0,F(1+O(t_1))+O(|t_1|^{(4-d)/(d-2)});\lambda). \qquad (6.36)$$

In view of our discussion in Sec. III, if μ (or t) represents a probability distribution at T very close to T_c, then $T-T_c$ is proportional to t_1.[11] Therefore, (6.36) says that, for $T\to T_c$,

$$G(0,\mu(T))\propto|T-T_c|^{-2/(d-2)} \qquad (6.37)$$

and hence

$$\gamma=2/(d-2). \qquad (6.38)$$

As we mentioned in Sec. III, and shall see later, (6.36) says only $0=0$ if $T<T_c$. Notice that, in constrast to the analysis in Sec. III, we have not introduced the linear approximation here for R_s near the fixed point. Nor do we need the concept of eigenvalues and eigenvectors. The factor s^{d-2} plays the role of $s^{1/\nu}$ in Sec. III, and s^{d-4} plays the role of s^{y_2}. The linear approximation tells us what to expect and can certainly be carried out here if desired. What we showed here by explicit construction of R_s is that, at least in the large n limits, conclusions of Sec. III are valid without linear approximation.

It is important to notice that the details of the coupling parameters do not appear in our discussion. The restriction that μ be close to the critical surface depends on just one parameter, namely, t_1 being small. The critical surface is thus a very enormous subspace of the parameter space. The explicit demonstration of the fact that $R_s\mu$ approaches μ^* for large s for any μ on the critical surface is an illustration of universality. How large s has to be so that $R_s\mu$ gets close to μ^* is the relevant information needed to determine the size of the critical region mentioned in Sec. III. Equation (6.28) gives some idea about the nature of this information. It depends on the details of the parameters and very little more can be said without explicit evaluation. One particular feature is that if $(d(t-t^*)/d\lambda)_1$ vanishes, then $t'-t^*$ becomes of $O(s^{d-6})$, and the critical region becomes much larger. This feature will be discussed in Sec. X in connection with the ϵ-expansion of critical exponents. The important point is that certain parameters are far more important than others in determining the size of the critical region, which cannot be estimated by simply looking at the magnitudes of coupling parameters and ask how far μ is from μ^*.

In the above, we did not do our analysis in the most general manner in the large n limit. For example, we could have included another term in the parameter G_0^{-1} so that

$$G_0^{-1}(k)=t_0+k^2+ak^4 \qquad (6.39)$$

instead of just t_0+k^2. It is left as an exercise for the reader to show that the fixed point remains exactly the same as before. The effect of ak^4 diminishes in $R_s\mu$ as s increases. Of course, the additional parameter a makes the critical surface and other details slightly more complicated.

On the other hand, if the form of the cutoff is changed, from our sharp cutoff $\int_{1/s}^1 dp$ to a smoothed cutoff

$$\int_0^\infty dp\exp(-p^6)\{1-\exp[-(sp)^6]\}, \qquad (6.40)$$

for example, there will be quantitative, not qualitative, modifications on the fixed point and details of $R_s\mu$. The critical exponents will stay the same.

It should be noted that (6.39) and (6.40) have very different meanings. A change of cutoff given by (6.40) is a quantitative change in the definition of the renormalization group. Equation (6.39) is just a modification of a coupling parameter, not a change in the renormalization group.

There is room for more work on the characteristics of $R_s\mu$ in the large n limit. For example, we have not inquired into the question of whether equation (6.12) can actually be solved for a general s. What happens if there are multiple solutions? Would our conclusions for large s, which were based on expansions in inverse powers of s, remain valid in that case? More work will be needed to answer these questions.

VII. BROKEN ROTATION SYMMETRY IN SPIN SPACE

The average value of $\phi_i(x)$ is zero as a result of the assumed rotation invariance in the n-dimensional spin vector space of the probability distribution and the assumption that μ is above the critical surface. This average value becomes nonzero when an external field H [see (2.9)] is turned on. It can be nonzero also when μ lies below the critical surface even when H is turned off. In the latter case, we have the rather striking phenomenon that a rotationally symmetric probability distribution produces apparently nonsymmetric average values. This is, of course, the most conspicuous feature of a phase transition. In this section we discuss cases with $\langle\phi\rangle\neq0$.

A. Transformation of H and M under R_s

In defining the parameter space [see (2.12) and (2.13)], odd powers of ϕ were excluded. Now we introduce one more parameter H by adding to $\mathcal{H}$ a term

$$H\int d^dx\ \phi_1(x)=HL^{d/2}\phi_{10}, \qquad (7.1)$$

where ϕ_{10} means $(\phi_{1k})_{k=0}$. The parameter H can be identified as proportional to a uniform external field in the 1 direction. (It should not be confused with a Hamiltonian.) It is easy to find out how H changes under R_s through (2.18). Since ϕ_{10} is never involved in the multiple integral, the only thing that happens is the replacement $\phi_{10}\to\alpha_s\phi_{10}$. Thus, in $\mathcal{H}'$, there appears a term

$$HL^{d/2}s^{1-\eta/2}\phi_{10}=H'L'^{d/2}\phi_{10}, \qquad (7.2)$$

$$H'=s^{(d-\eta)/2+1}H. \qquad (7.3)$$

Recall that $\alpha_s=s^{1-\eta/2}$, $sL'=L$. We can therefore write formally

$$(H',\mu')=R_s(H,\mu)=(H',R_s\mu), \qquad (7.4)$$

with H' given by (7.3) and $R_s\mu$ defined previously, as the renormalization group transformation in the extended parameter space. The average "magnetization" M is given by

$$M(H, \mu)=\langle\phi_1(x)\rangle_P=L^{-d/2}\langle\phi_{10}\rangle_P, \tag{7.5}$$

where P of course denotes the probability distribution represented by (H, μ). We know that

$$\langle\phi_{10}\rangle_P=s^{1-\eta/2}\langle\phi_{10}\rangle_{P'}, \tag{7.6}$$

where P' stands for the probability distribution represented by (H', μ). [If (7.6) is not obvious, please go back to the three trivial facts discussed at the beginning of Sec. II. See (2.5) in particular.] Substituting (7.6) in (7.5), we obtain an equation analogous to (2.28):

$$\begin{aligned} M(H, \mu) &= L'^{-d/2}s^{-d/2}s^{1-\eta/2}\langle\phi_{10}\rangle_{P'} \\ &= M(H', \mu')s^{-(d+\eta)/2+1} \\ &= M(Hs^{(d-\eta)/2+1}, \mu')s^{-(d+\eta)/2+1}. \end{aligned} \tag{7.7}$$

Before we proceed further, let us emphasize the fact that as long as M and H are uniform, the renormalization group transformation $\mu'=R_s\mu$ discussed previously is not affected, regardless whether μ is above, on, or below the critical surface.

B. The exponents δ and β

If $\mu=\mu(T_c)$ is a point on the critical surface, μ' will approach μ^* for large s. If H is small enough (i.e., weak external field), we can choose

$$s=H^{-[(d-\eta)/2+1]^{-1}} \tag{7.8}$$

so that (7.7) becomes

$$M(H, \mu(T_c))=H^{1/\delta}M(1, \mu^*+O(H^{-y_2[(d-\eta)/2+1]^{-1}})), \tag{7.9}$$

where

$$\delta=(d+2-\eta)/(d-2+\eta). \tag{7.10}$$

In the limit of small H,

$$M\propto H^{1/\delta} \tag{7.11}$$

which is the equation defining the exponent δ.

If $H=0$, and μ is below the critical surface, we choose $s=|t_1|^{-\nu}$ and obtain from (7.7)

$$M(\mu)=|t_1|^{\beta}M(\mu^*-e_1+O(|t_1|^{-\nu y_2})), \tag{7.12}$$

$$\beta=\tfrac{1}{2}\nu(d-2+\eta). \tag{7.13}$$

The exponent β is defined by $M\propto|T-T_c|^{\beta}$ in the limit of small $|T-T_c|$ below T_c.

C. Correlation functions and susceptibilities

Although (7.3) and (7.4) are all we need to know about the renormalization group in addition to what we should know when $H=0$, a finite M clearly makes the physical situation very different. An important difference is that the rotation symmetry in the spin space is broken when $M\neq 0$, and (2.8) is no longer a convenient definition.

We defined the new random variable $\psi_1(x)$ by

$$\phi_1(x)=M+\psi_1(x). \tag{7.14}$$

For the Fourier component with $k=0$, $i=1$,

$$L^{-d/2}\phi_{10}=M+L^{-d/2}\psi_{10}, \tag{7.15}$$

and $\phi_{ik}=\psi_{ik}$ otherwise. Clearly, by definition,

$$\langle\psi_{10}\rangle=0. \tag{7.16}$$

We need to define two (longitudinal and transverse) correlation functions now since the 1 direction is distinct:

$$G_{\parallel}(k, \mu, H)=\langle|\phi_{1k}|^2\rangle_P, \quad k\neq 0, \tag{7.17}$$

$$\delta_{ij}G_{\perp}(k, \mu, H)=\langle\phi_{ik}\phi_{j-k}\rangle_P, \quad i, j\neq 1. \tag{7.18}$$

When the symbol H is omitted, we shall mean $H=0$. By $G_{\parallel}(0, \mu, H)$ we shall mean the limit $k\to 0$, not at $k=0$.

If we add a small field δH in addition to H, there will be a δM in addition to M as a result. One easily obtains similar to (2.10)

$$G_{\parallel}(0, \mu, H)=\left(\frac{\delta M}{\delta H}\right)_{\parallel}. \tag{7.19}$$

If the additional field is perpendicular to the original field, say in the 2 direction, δM would be also in the 2 direction. We have

$$G_{\perp}(0, \mu, H)=\left(\frac{\delta M}{\delta H}\right)_{\perp}. \tag{7.20}$$

The quantities (7.19) and (7.20) are called the longitudinal and transverse susceptibilities, respectively. We note the important fact that applying a field δH perpendicular to the original field H is the same as rotating the original field by a small angle $\delta H/H$ in the spin space. Thus, the result must be rotating the original M by the same angle since μ is assumed to be invariant under this rotation. We therefore conclude that

$$\frac{\delta M}{M}=\frac{\delta H}{H}, \tag{7.21}$$

which implies that $(\delta H/\delta M)_{\perp}=H/M$, and

$$G_{\perp}(0, \mu, H)=M/H. \tag{7.22}$$

This is a very important observation, for it fixes the small k limit of the transverse correlation function. When μ is below the critical surface, $M\neq 0$ even when $H=0$. Equation (7.22) asserts that in this case

$$G_{\perp}^{-1}(0, \mu)=0, \tag{7.23}$$

i.e., the transverse correlation function must diverge as $k\to 0$. This statement is a form of the Goldstone theorem often encountered in many-body theory and in field theory.

D. The large n limit

For an explicit illustration of the above discussion, let us consider the large n limit with $t_1<0$. As we mentioned earlier, the structure of R_s remains the same apart from the additional equation (7.3). Therefore, what we want to illustrate is mainly the effect of finite M on various averages. We shall assume that t_1 is negative but small.

Note that there are $n-1$ transverse components ($i=2, 3, \ldots, n$), but only one longitudinal ($i=1$). For large n, close loops of transverse components dominate.

The transverse correlation function has the same structure as G before except that an additional term shown in Fig. 8(a) appears in the self-energy. Similar to (6.5) and (6.6), we write

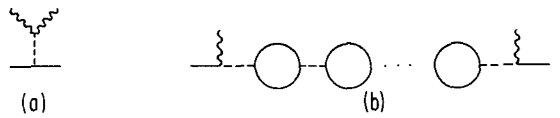


FIG. 8. Additional self-energy terms when $M\neq 0$. A wavy line here denotes a factor of M. A dashed line here denotes a factor of $u(N)=dt/dN$ [see (7.26)]. (a) The term for $G_\perp$ and (b) for $G_\parallel$.

$$G_\perp^{-1}(k,t,H)=t(N/N_c)+\tfrac{1}{2}M^2u+k^2, \tag{7.24}$$

$$N=\tfrac{1}{2}(n-1)K_d\int_0^1 dp\, p^{d-1}G_\perp(\mu,t,H), \tag{7.25}$$

$$u=u(N)\equiv\frac{dt}{dN}\,. \tag{7.26}$$

Since t_1 is assumed to be small, $M^2\ll N_c$. We are neglecting higher-order terms in M^2/N_c. Together with (7.22), i.e.,

$$t(N/N_c)+\tfrac{1}{2}M^2u(N)=H/M, \tag{7.27}$$

solutions for N, M in terms of t and H can be obtained. The factor $(n-1)$ in (7.25) will be replaced by n.

For $G_\parallel$, there is still another term in the self-energy as shown in Fig. 8(b). We have

$$G_\parallel^{-1}(k,t,H)=G_\perp^{-1}(k,t,H)+uM^2[1+(n/2)u\Pi(k)]^{-1}, \tag{7.28}$$

where

$$\Pi(k)=(2\pi)^{-d}\int d^dpG_\perp(p,t,H)G_\perp(p+k,t,H). \tag{7.29}$$

Consider first the case $H=0$. Then, by (7.27) and (7.24), we have

$$G_\perp(k,t)=k^{-2}. \tag{7.30}$$

Therefore, $N=N_c$, and

$$\Pi(k)=\int d^dp(2\pi)^{-4}p^{-2}(p+k)^{-2} \tag{7.31}$$

$$=\Pi(1)k^{d-4}(1+O(k^{4-d})), \tag{7.32}$$

where

$$\begin{aligned}\Pi(1)&=JB(d/2-1,d/2-1),\\ J&=\tfrac{1}{2}K_d\pi(d/2-1)/\sin\pi(d/2-1),\end{aligned} \tag{7.33}$$

and B is the beta function. Since $N=N_c$, (7.27) gives

$$t_1+\tfrac{1}{2}M^2u(N_c)=0, \tag{7.34}$$

which implies $M\propto|t_1|^{1/2}$ and $\beta=\frac{1}{2}$ as (7.13) implies for large n. Substituting (7.30)—(7.34) in (7.28), we obtain, for small k,

$$G_\parallel(k,t)=-\left[t_1^{-1}k^{d-4}\ \frac{n}{4N_c}\left(\frac{dt}{d\lambda}\right)_1\Pi(1)\right]+\text{const.} \tag{7.35}$$

By (7.19), (7.35) implies that

$$\left(\frac{\delta H}{\delta M}\right)_{\parallel,\,H=0}=0, \tag{7.36}$$

i.e., the H vs M curve at $H=0$ is flat.

We proceed to the case of nonzero but small H/M. Then

$$G_\perp^{-1}=H/M+k^2 \tag{7.37}$$

and therefore N is, by (7.25),

$$N=-(n/2)(H/M)^{d/2-1}J/(d/2-1)+N_c+O(H/M). \tag{7.38}$$

Since $d/2-1>0$, then $(1/N_c)(N-N_c)$ is small. We can expand t and u in (7.27) and obtain

$$t_1+u(N_c)(\tfrac{1}{2}M^2-(H/M)^{d/2-1}(n/2)J/(d/2-1))+O(H/M)=0. \tag{7.39}$$

This is the thermodynamic equation of state relating H, M, and $t_1\propto T-T_c$. $\Pi(k)$ becomes more complicated in this case.[14] Let us just note that

$$\Pi(0)=(H/M)^{d/2-2}J+O(1) \tag{7.40}$$

and (7.28) gives

$$\begin{aligned}G_\parallel(0,t,H)^{-1}&=\left(\frac{\delta H}{\delta M}\right)_\parallel\\ &=\frac{2M^2}{nJ}\left(\frac{H}{M}\right)^{2-d/2}\left[1+O\left(\frac{H}{M}\right)^{2-d/2}\right]\,.\end{aligned} \tag{7.41}$$

In conclusion, the presence of a finite uniform M does not modify the idea or formulation of the renormalization group in any essential way. On the other hand, some of the important consequences of a finite M such as (7.41) do not appear derivable from renormalization group arguments alone. Such consequences obtained here are general to the extent that no assumption is made about the function $t(N/N_c)$ which specifies the details of the interaction.

Discussions on critical exponents and the equation of state below T_c to $O(1/n)$ can be found in the work of Brezin and Wallace.[8]

VIII. THE FREE ENERGY

In this section we are interested in how the free energy transforms under the renormalization group. We shall derive the transformations from our basic definition explicitly and examine the validity of the usual scaling arguments. A study of free energy and corrections to scaling laws assuming general transformation properties has been carried out by Wigner.[15]

The free energy per unit volume $F(T)$ is defined by[16]

$$\exp[-L^dF(T)/T]=\prod_{0<k'<\Lambda}\int d\phi_{k'}\exp[-H(\Lambda)/T]. \tag{8.1}$$

Clearly an additive constant in $H(\Lambda)$ will make a difference in F. To apply the renormalization group to the study of the free energy, one must specify the additive constant in $\mathcal{H}$ so far ignored. We shall adopt the rule that the additive constant is always *written out explicitly* and symbols like $\mathcal{H}$, $\mathcal{H}'$ *will contain no additive constant*, i.e., $\mathcal{H}$, $\mathcal{H}'$ are zero if $\phi=0$. We now define $\mathcal{F}=\mathcal{F}(\mu)$ by

$$\exp(-L^d\mathcal{F})=\prod_{0<k<\Lambda}\int d\phi_k\exp(-\mathcal{H}). \tag{8.2}$$

Similarly we define $\mathcal{F}'=\mathcal{F}(\mu')$ by replacing in (8.2) $\mathcal{H}$ by $\mathcal{H}'$, $L^d\mathcal{F}$ by $L'^d\mathcal{F}'$ and keeping in mind that the density of points in k space over which the product Π runs through is changed to $L'^d(2\pi)^{-d}$. To relate $\mathcal{F}'$ to $\mathcal{F}$, we separate the multiple integral in (8.1) into two and write

$$\exp(-\mathcal{F}L^d)=\prod_{0<k<\Lambda/s}\int d\phi_k\prod_{\Lambda/s<k'<\Lambda}\int d\phi_{k'}\exp(-\mathcal{H}). \tag{8.3}$$

We then make the substitution $\phi_k\to\alpha_s\phi_{sk}$ for the second (the left) set of variables and obtain

$$\exp(-\mathcal{J}L^d)=\prod_{0<sk<\Lambda}\int d\phi_{sk}\alpha_s^{-1}\left[\prod_{\Lambda/s<k'<\Lambda}\int d\phi_{k'}\exp(-\mathcal{H})\right]_{\phi_k\to\alpha_s\phi_{sk}}$$
$$=\prod_{0<k<\Lambda}\int d\phi_k\exp(-\mathcal{H}')\exp[-L^d(A+A_0)]$$
$$=\exp[-L'^d\mathcal{J}'-(A+A_0)L^d]. \qquad (8.4)$$

We have applied the definition (2.18) for $\exp(-\mathcal{H}')$. The constants A and A_0 are defined by

$$\exp(-L^dA)=\left[\prod_{\Lambda/s<k'<\Lambda}\int d\phi_{k'}\exp(-\mathcal{H})\right]_{\phi_k=0}, \qquad (8.5)$$

where ϕ_k, $0<k<\Lambda/s$, the unintegrated variables, are set to zero, and

$$\exp(-L^dA_0)=\prod_{0<k<\Lambda/s}\alpha_s^{-1},$$

i.e.,

$$A_0=nK_d\int_0^{\Lambda/s}dp\,p^{d-1}[1-(\eta/2)]\ln s, \qquad (8.6)$$

since $\alpha_s=s^{1-\eta/2}$. The additive constant A would be just $\mathcal{J}$ if all ϕ_k with $0<k<\Lambda/s$ were set to zero. A_0 is to compensate the change of the size of the phase space produced by the substitution $\phi\to\alpha_s\phi$. We have therefore

$$\mathcal{J}(\mu)=s^{-d}\mathcal{J}(\mu')+A+A_0 \qquad (8.7)$$

from (8.4). Now let $H(\Lambda)/T$ be represented by $\mu(T)$, then the free energy is

$$F(T)=\mathcal{J}(\mu(T))T. \qquad (8.8)$$

We obtain from (8.7)

$$\mathcal{J}(\mu(T))-\mathcal{J}(\mu(T_c))$$
$$=s^{-d}[\mathcal{J}(\mu'(T))-\mathcal{J}(\mu'(T_c))]+A(T)-A(T_c). \qquad (8.9)$$

For large s, $\mu'(T_c)$ approaches the fixed point μ^*. If $T-T_c$ is very small, we choose

$$s=|t_1|^{-\nu}\propto|T-T_c|^{-\nu} \qquad (8.10)$$

as was done in (3.13) and obtain from (8.9)

$$\mathcal{J}(\mu(T))-\mathcal{J}(\mu(T_c))$$
$$=|t_1|^{\nu d}[\mathcal{J}(\mu^*\pm e_1)-\mathcal{J}(\mu^*)+O(|t_1|^{-\nu y_2})]$$
$$+(A(T)-A(T_c))_{s=|t_1|^{-\nu}}, \qquad (8.11)$$

where, in the argument of $\mathcal{J}(\mu^*\pm e_1)$, the + and − signs correspond to the cases $t_1>0$ and $t_1<0$, respectively. In the small $T-T_c$ limit, we have[17]

$$F(T)-F(T_c)\propto|T-T_c|^{\nu d}+\text{“less singular terms”} \qquad (8.12)$$

provided that the last two terms of (8.11) are truly less singular. Note that $\mathcal{J}(\mu^*+e_1)$ is expected to be different from $\mathcal{J}(\mu^*-e_1)$. Therefore the proportionality constant in front of $|T-T_c|^{\nu d}$ in (8.12) for $T>T_c$ is different from that for $T<T_c$.

The last two terms of (8.11), usually handwaved away as nonsingular terms, deserve more attention. They are the contribution from the ϕ_k's with $k>\Lambda/s$ as (8.5) indicates. Equation (8.12) assumes that such contribution is less singular compared to the contribution from ϕ_k's with $k<\Lambda/s$. The assumption is not obviously valid and, in some cases, invalid. To gain qualitative understanding of the effect of A, let us turn to the large n limit.

In the large n limit, A is the sum of tree graphs with no external line and with internal lines of wavevectors restricted between Λ/s and Λ. Instead of summing these graphs directly, we make use of the identity

$$\frac{\partial A}{\partial t_0}=N_a(0)=N(0)=\left(1+\frac{\zeta(0)}{s^2}\right)N_c, \qquad (8.13)$$

where $N_a(0)$ or $N(0)$ is given by (6.8) or (6.12) with $\lambda=0$ and ζ is defined by (6.19). The identity (8.13) is similar to the well-known thermodynamic relation $-\partial\Omega/\partial\mu=N$. The reader can convince himself by studying a couple of simple graphs.

Equation (6.29) tells us what ζ should be when t_1 is small and s is large so that $t'=O(1)$. We obtain

$$\zeta(0)=(-s^2t_1+t_0')\Big/\left(\frac{dt}{d\lambda}\right)_1+O(t_1,s^{-2}). \qquad (8.14)$$

Substituting (8.14) in (8.13), we obtain

$$\frac{\partial A}{\partial t_0}=N_c\left[1-t_1\Big/\left(\frac{dt}{d\lambda}\right)_1+s^{-2}t_0'\Big/\left(\frac{dt}{d\lambda}\right)_1+s^{-2}O(t_1,s^{-2})\right]. \qquad (8.15)$$

Since t_0 is linearly related to t_1 [$t_1=t_0+\Sigma(N_c)$, see (6.18)], and t_1 is proportional to $T-T_c$, we obtain from (8.15)

$$A(T)-A(T_c)$$
$$=N_c\int_0^{t_1}dt_1\left[1-t_1\Big/\left(\frac{dt}{d\lambda}\right)_1+s^{-2}t_0'\Big/\left(\frac{dt}{d\lambda}\right)_1\right]+O(t_1,s^{-2})t_1s^{-2}. \qquad (8.16)$$

While $(dt/d\lambda)_1$ is independent of t_1, t_0' does depend on t_1. For $t_0=0$, (6.28) shows that $t_0'-t_0^*\sim s^{d-4}$ [$\zeta=O(1)$ in this case], and (6.32) shows what happens if $t_1\neq0$. For our purpose, it suffices to take (6.32) as saying that

$$t_0'-t_0^*\sim t_1s^{d-2} \qquad (8.17)$$

so that (8.16) becomes

$$A(T)-A(T_c)$$
$$=N_c\left[t_1-\tfrac{1}{2}t_1^2\Big/\left(\frac{dt}{d\lambda}\right)_1+s^{-2}t_1t_0^*\Big/\left(\frac{dt}{d\lambda}\right)_1\right.$$
$$\left.+t_1^2O(s^{d-4})+O(t_1,s^{-2})t_1s^{-2}\right]. \qquad (8.18)$$

Now we set $s=|t_1|^{-\nu}$ as before [$\nu=1/(d-2)$ in the large n limit, see above (6.37)], and obtain

$$A(T)-A(T_c)$$
$$=N_c\left[t_1-\tfrac{1}{2}t_1^2\Big/\left(\frac{dt}{d\lambda}\right)_1+t_1|t_1|^{2\nu}t_0^*\Big/\left(\frac{dt}{d\lambda}\right)_1+O(|t_1|^{2+(4-d)\nu})\right.$$
$$\left.+t_1O(t_1,|t_1|^{2\nu})|t_1|^{2\nu}\right],\quad t_1\propto(T-T_c). \qquad (8.19)$$

Clearly, in order that the last two terms of (8.11) become less singular than the first term $|t_1|^{\nu d}$, we must have

$$\nu d<1+2\nu,\quad \nu d<2+(4-d)\nu \qquad (8.20)$$

and νd not being an integer in view of the terms in (8.19). But since $\nu=1/(d-2)$ for our case, we have

$$1+2\nu=2+(4-d)\nu=d/(d-2)=\nu d. \qquad (8.21)$$

Therefore, for the large n limit the $A(T)-A(T_c)$ term, which is usually thrown away by handwaving arguments, is just as singular as the $|t_1|^{\nu d}$ term. Equation (8.12) remains true, however, but not as a consequence of the scaling hypothesis discussed in the Introduction [see (1.1)]. We arrived at the inequality (8.20) as a criterion for the validity of the scaling hypothesis prediction about the free energy from our tree graph analysis. However, the appearance of the $t_1 s^{-2}$ term in (8.15), which leads to the $t_1|t_1|^{2\nu}$ term in (8.19), looks so general that we expect (8.20) to be meaningful even if n is not large.[18] Note that we are not particularly interested in the validity of (8.20) here. What we want to illustrate is the nature of approximation and ambiguities involved in some of the applications of the scaling hypothesis.

In order for νd to be a noninteger, we must exclude a special set of dimensions. Since $\nu = 1/(d-2)$ for large n, we must exclude

$$d = d_l = 2 + 2/l, \quad l = 2, 3, 4, \ldots, \tag{8.22}$$

where νd will take the value $l+1$. These special dimensions have received some attention lately. Further discussion about them is beyond our scope here.[19]

IX. TRANSFORMATION OF COMPOSITE VARIABLES

There are random variables of interest which involve ϕ_k's with large k in an essential way. They often appear as products of ϕ_k's. We shall call ϕ_k "elementary" and products of ϕ's "composite." Under the transformations of renormalization group the averages and the correlation functions of the composite variables usually do not obey simple formulas like (7.7) and (2.28). We shall not give a general discussion here,[20] but only give a qualitative illustration by examining the simplest composite variable

$$\begin{aligned}\phi^2(x) &\equiv \tfrac{1}{2}\sum_{i=1}^{n}(\phi_i(x))^2\\ &= \tfrac{1}{2}L^{-d}\sum_{i=1}^{n}\sum_{k,k'<\Lambda}\phi_{ik}\phi_{ik'}\exp[i(k+k')\cdot x].\end{aligned} \tag{9.1}$$

A. Transformation of random variables

First, we shall define what we mean by transforming random variables by R_s.

Recall that the probability distribution P represented by μ is equivalent to P' represented by $\mu' = R_s\mu$ in view of the three trivial facts given at the beginning of Sec. II. They are equivalent in the sense that the average value of any function of ϕ_k over P and that over P' are simply related. Equation (2.19) is an example. The transformation

$$\phi_k \rightarrow s^{1-\eta/2}\phi_{sk} \tag{9.2}$$

can be viewed as the transformation of ϕ_k under R_s in the sense that

$$\langle h(\phi_k)\rangle_P = \langle h(\alpha_s\phi_{sk})\rangle_{P'} \tag{9.3}$$

for any function h. One could write $R_s\phi_k = \alpha_s\phi_{sk}$ if desired. Note that there is the important condition

$$sk < \Lambda, \tag{9.4}$$

which must be satisfied in order that (9.3) makes sense.

Now we ask how a composite variable transforms under R_s, i.e., we ask what is $(\phi^2)'$ so that

$$(\phi^2) \rightarrow (\phi^2)' \tag{9.5}$$

and

$$\langle \phi^2 h(\phi_k)\rangle_P = \langle(\phi^2)' h(\alpha_s\phi_{sk})\rangle_{P'}, \tag{9.6}$$

where k satisfies (9.4). The answer is very complicated since ϕ^2 involves wavevectors violating (9.4). The multiple integral in (2.18) defining R_s plays no part in the transformation of the elementary variable (9.2), but plays an important role in (9.5). Physically, (9.2) tells how ϕ_k looks after a change of length scale. The quantity $(\phi^2)'$ is to tell what the square of $\phi(x)$ looks like after a change of scale. This change of scale also involves an averaging process so that the minimum resolvable distance after the scale change remains at Λ^{-1} as before.

Note that (9.6) would be the same as (9.3) if we make the replacement

$$\begin{aligned}&\langle(\phi^2)\cdots\rangle_P \rightarrow \langle\cdots\rangle_P,\\ &\langle(\phi^2)'\cdots\rangle_{P'} \rightarrow \langle\cdots\rangle_{P'}.\end{aligned} \tag{9.7}$$

Then it is clear that all we need to do to define $(\phi^2)'$ is to replace (2.18) by

$$(\phi^2)'\exp(-\mathcal{H}') = \left[\int \delta\bar{\phi}(\phi^2)\exp(-\mathcal{H})\right]_{\phi_k\rightarrow\alpha_s\phi_{sk}}, \tag{9.8}$$

where we have used the shorthand $\delta\bar{\phi}$ defined by (4.12) for the multiple integral of (2.18).

B. Transformation of ϕ^2 in the large n limit

The graph representation for (9.8) is obtained in the same fashion as that for (2.18) as discussed in Sec. IV. The only new ingredient here is the additional factor ϕ^2. In the large n limit, the sum over tree graphs gives the exact answer. We write

$$(\phi^2)' = \sum_{l=0}^{\infty} A_l(\phi^2)^l, \tag{9.9}$$

where

$$A_l = (\text{connected graphs with a } \phi^2 \text{ vertex})s^{l(2-d)} \tag{9.10}$$

similar to (4.23). We have set $\eta = 0$. A few tree graphs for A_l are shown in Fig. 9. Notice that we are again ignoring the dependence of the A_l's on the wavevectors of external lines, as we ignored that of graphs for u_{2m} in the discussions of Secs. V and VI. Under this approximation, $(\phi^2)'$ is a function of ϕ^2 only. [Recall that under the same approximation, $\mathcal{H}$, $\mathcal{H}'$ assume the simple form (6.2). See discussion below (5.28).] The coefficients are easily determined in the same way we determined u_{2m} in Sec. VI. We realize that if we close all l pairs of external lines of A_l to form l closed loops, we obtain a graph for $N = \langle\phi^2\rangle_P$. Each new loop is a factor $N' = \langle\phi^2\rangle_{P'}$. Thus,

$$N = \sum_{l=0}^{\infty} A_l N'^l. \tag{9.11}$$

To find A_l, all we have to do is to expand N as a power of N' via (6.7). Now let us define the function $N(\lambda)$ as before by (6.12) with t' given by (6.11) in terms of $N(\lambda)$. Since $\lambda = N'/N_c$, we have

$$N\left(\frac{N'}{N_c}\right) = \sum_{l=0}^{\infty} N_c^{-l}\frac{1}{l!}\left(\frac{d^l N(\lambda)}{d\lambda^l}\right)_{\lambda=0}. \tag{9.12}$$

Comparing (9.11) and (9.12) we obtain

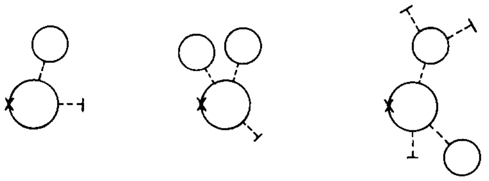


FIG. 9. Some graphs for A_l in the transformation $\phi^2 \rightarrow (\phi^2)'$. See (9.9), (9.10).

$$A_l = \frac{N_c^{-l}}{l!}\left(\frac{d^l N(\lambda)}{d\lambda^l}\right)_{\lambda=0} \tag{9.13}$$

and $(\phi^2)'$ is therefore obtained by substituting (9.13) in (9.9). We write simply

$$(\phi^2)' = N(\phi^2/N_c) \tag{9.14}$$

with the function $N(\lambda)$ defined by (6.12), (6.11). This is the desired transformation law of ϕ^2 under R_s in the large n limit.

C. Qualitative features and limiting values of $(\phi^2)'$

Now the task is to see explicitly how the function $N(\lambda)$ depends on λ and, more important, on s. It is convenient to introduce the function $\zeta(\lambda)$ defined by (6.19):

$$N(\lambda) = N_c(1 + \zeta/s^2), \tag{9.15}$$

because $\zeta = O(1)$ for large s if μ is on the critical surface.

First, let us see how $(\phi^2)'$ looks when $\mu = \mu^*$. Equation (6.23) gives $N(\lambda)/N_c$ in terms of t^*. We know how t^* depends on λ (see Fig. 6). We have, from (6.23),

$$\begin{aligned} N^*(\lambda)/N_c &= 1 + (d/2 - 1)t^*/s^2\Phi(-t^*/s^2, 1, 2 - d/2), \\ &\equiv f(t^*/s^2). \end{aligned} \tag{9.16}$$

The curve $\lambda = f(t^*)$ defines $t^*(\lambda)$ [see (5.41)]. We note that, since $t^*(1) = 0$, we have

$$N^*(1)/N_c = 1, \tag{9.17}$$

and

$$N^*(-\infty)/N_c \rightarrow f(-s^{-2}), \tag{9.18}$$

since $t^*(-\infty) \rightarrow -1$. For large and positive λ, (5.50) and (5.51) tells us that

$$N^*(\lambda)/N_c = \lambda s^{2-d} + O(1). \tag{9.19}$$

Equation (9.19) is useful if $t^*(\lambda)/s^2$ is large. If t^*/s^2 is small, we use (5.24) to expand (9.16) and obtain

$$\frac{N^*(\lambda)}{N_c} = 1 + \left(\frac{d}{2} - 1\right)\frac{f^*}{s^2}\left(\frac{t^*}{s^2}\ \frac{2}{4-d} - \frac{2t^*/s^2}{6-d} + O(t^{*2}s^{-4})\right). \tag{9.20}$$

With the information provided by (9.17)–(9.20), we can easily make a rough sketch of $N(\lambda)$ vs λ.

How would $N(\lambda)$ look if μ is away from μ^*? We shall discuss two cases, (a) μ is on the critical surface and (b) μ is not on the critical surface but very close to it.

Case (a). In this case $\mu' = R_s^l\mu$ will approach μ^* like s^{d-4} for large s, and ζ is of $O(1)$. From (6.25) and (9.15), we have

$$\begin{aligned} \frac{N(\lambda)}{N_c} &= 1 + s^{-2}t'\Big/\left(\frac{dt}{d\lambda}\right)_1 + O(s^{-4}) \\ &= 1 + s^{-2}t^*\Big/\left(\frac{dt}{d\lambda}\right)_1 [1 + O(s^{d-4})] + O(s^{-4}). \end{aligned} \tag{9.21}$$

For large λ, so large that t'/s^2 is also large, we obtain from (6.12)

$$N(\lambda)/N_c = \lambda s^{2-d} - (1 - s^{2-d})(1/t')[1 + O(s^2/t')]. \tag{9.22}$$

Case (b). In this case $t_1 \neq 0$, but is assumed to be small. From (9.15) and (6.30), we obtain

$$\frac{N(\lambda)}{N_c} = 1 - t_1\Big/\left(\frac{dt}{d\lambda}\right)_1 + s^{-2}t'\Big/\left(\frac{dt}{d\lambda}\right)_1 + s^{-2}O(s^{-2}, t_1). \tag{9.23}$$

For large λ and large t'/s^2, (9.22) still holds.

The above results can be summed up as, for large s,[21]

$$(\phi^2)' = \text{const} + s^{-2}t'(\phi^2)\Big/\left(\frac{dt}{d\lambda}\right)_1 + s^{-2}O(s^{-2}, t_1), \tag{9.24}$$

and, for large ϕ^2

$$(\phi^2)' = \phi^2 s^{2-d} + O(1), \tag{9.25}$$

where $t_1 = 0$ for μ on the critical surface and, $t' = t = t^*$ if $\mu = \mu^*$.

Note that our results on the transformation $\phi^2 \rightarrow (\phi^2)'$ are valid regardless of whether t_1 is positive or negative, the same situation as that for the transformation $\mu' = R_s\mu$ discussed in Sec. VI.

D. Correlation functions involving ϕ^2

We now apply the above results on $\phi^2 \rightarrow (\phi^2)'$ to examine certain correlation functions. Consider

$$\begin{aligned} K(k_1, k_2, \mu) &\equiv \int d^dx_1 d^dx_2 \exp(-ik_1\cdot x_1 - ik_2\cdot x_2) \\ &\quad \times \langle\phi^2(0)\phi(x_1)\phi(x_2)\rangle_P \\ &= L^d\langle\phi^2\phi_{k_1}\phi_{k_2}\rangle_P, \end{aligned} \tag{9.26}$$

where component indices are not written explicitly and ϕ^2 means $\phi^2(0)$.

For the rest of this section, we shall assume that μ is above the critical surface to avoid complications which are not essential in illustrating consequences of renormalization group arguments. The important new feature below the critical surface in addition to those discussed in Sec. VII is that in writing $\phi_1 = \psi_1 + M$ [see (7.14)], $K(k_1, k_2, \mu)$ will contain a term (only the component index 1 is written explicitly),

$$ML^{d/2}\langle\phi_{1-k_1-k_2}\phi_{k_1}\phi_{k_2}\rangle_P$$

because of an $M\psi_1$ term in ϕ^2. Note that averages of products of odd number of ϕ's are nonzero in general when $M \neq 0$.

We now proceed assuming $M = 0$. From (9.6) and (9.2), we obtain

$$\begin{aligned} K(k_1, k_2, \mu) &= L'^d s^{d+2-\eta}\langle(\phi^2)'\phi_{sk_1}\phi_{sk_2}\rangle_{P'} \\ &\equiv s^{-d_{\phi^2}}K''(sk_1, sk_2, \mu')s^{d+2-\eta}. \end{aligned} \tag{9.27}$$

We assume that k_1, k_2 and $k_1 + k_2$ are nonzero but so small that (9.4) is valid. Note that $K''(\cdot\cdot\,\mu')$ is not the same as $K(\cdot\cdot\,\mu')$ because $(\phi^2)' \neq \phi^2$. We use the double prime to emphasize the fact. The factor $s^{-d_{\phi^2}}$ is included as a part of the definition of K'' aimed at removing explicit s dependence from K'' and will be discussed later. Equation (9.27) is a generalization of (2.28). The critical behavior of K can be extracted as before by considering (9.27) at large s. We proceed to the large n limit.

For large s, $(\phi^2)'$ is given by (9.24)

$$K(k_1, k_2, t) = s^{-2}K''(sk_1, sk_2, t')s^{d+2}, \tag{9.28}$$

where

$$K''(sk_1, sk_2, t') = L'^d\langle t'(\phi^2)\phi_{sk_1}\phi_{sk_2}\rangle_{P'} \bigg/ \left(\frac{dt}{d\lambda}\right)_1, \tag{9.29}$$

$$d_{\phi^2} = 2.$$

The constant term in (9.24) does not contribute since $k_1 + k_2 \neq 0$. The higher-order terms of (9.24) are ignored.

If $t_1 = 0$, we choose $s = 1/k$ to obtain from (9.28)

$$K(k_1, k_2, \mu) = k^{-d}K''(k_1/k, k_2/k, t^* + O(k^{4-d})). \tag{9.30}$$

For $t_1 \neq 0$ but small, we choose $s = t_1^{-\nu} = \xi$ and obtain from (9.28)

$$K(k_1, k_2, t) = \xi^d K''(k_1\xi, k_2\xi, t'), \tag{9.31}$$

where t' would be independent of t_1 if $O(t_1^{\nu(4-d)})$ is ignored. In particular, for $k_1, k_2 \to 0$, [but not identically zero in order to keep the constant in (9.24) away], we have for very small t_1,

$$K(0, 0, t) \propto t_1^{-\nu d}, \tag{9.32}$$

with $\nu = 1/(d-2)$. A quantity $\Gamma(k_1 k_2)$ appears often in the literature. It is defined by

$$\Gamma(k_1 k_2 \mu)G(k_1, \mu)G(k_2, \mu) = K(k_1, k_2, \mu). \tag{9.33}$$

Corresponding to (9.30) and (9.32), we have

$$\Gamma(k_1, k_2, t) \propto k^{4-d}, \quad t_1 = 0, \tag{9.34}$$

$$\Gamma(0, 0, t) \propto t_1^{\nu(4-d)}, \quad t_1 \text{ small}. \tag{9.35}$$

The above analysis is easily generalized to apply to correlation functions of the form

$$\begin{aligned} K(k_1, k_2 \cdots k_m, \mu) \\ &= \int d^dx_1 \cdots d^dx_m \exp(-ik_1\cdot x_1 - ik_2\cdot x_2 - \cdots - ik_m\cdot x_m) \\ &\quad \times \langle\phi^2(0)\phi(x_1)\cdots\phi(x_m)\rangle_P \\ &= L^{md/2}\langle\phi^2(0)\phi_{k_1}\cdots\phi_{k_m}\rangle_P, \end{aligned} \tag{9.36}$$

with all subsums of the k's small but nonzero. Again there should be component indices for the ϕ's but these are not written for simplicity. Corresponding to (9.30) and (9.31), we obtain

$$\begin{aligned} K(k_1 \cdots k_m, \mu) \\ &= k^{-(m/2)(d+2-\eta)+d_{\phi^2}}K''(k_1/k, \ldots, k_m/k, \mu') \end{aligned} \tag{9.37}$$

for μ on the critical surface, which means $\mu' = \mu^* + O(k^{-y_2})$; and

$$\begin{aligned} K(k_1 \cdots k_m, \mu) \\ &= \xi^{(m/2)(d+2-\eta)-d_{\phi^2}}K''(\xi k_1, \ldots, \xi k_m, \mu') \end{aligned} \tag{9.38}$$

for μ slightly away from the critical surface and $\xi = |t_1|^{-\nu}$. Again, $\eta = 0$, $\nu = 1/(d-2)$, $d_{\phi^2} = 2$ in the large n limit.

E. Dimensions of random variables[20]

As we mentioned in Sec. II, the transformation (9.2), being thought of as a scale transformation, defines the dimension of ϕ_k to be $-1 + \eta/2$ in units of wavenumber, i.e., inverse length. In fact, much of our analysis can be viewed as simply changes of scales and dimensional analysis, provided that dimensions for various variables are properly defined. For example the dimension of

$$\phi(x) = L^{-d/2}\sum_{k<\Lambda}\phi_k \exp(ik\cdot x) \tag{9.39}$$

can be defined, via (9.2) which implies that

$$\begin{aligned} \phi(x) &\to L^{-d/2}\sum_{k<\Lambda} s^{1-\eta/2}\phi_{sk}\exp(ik\cdot x) \\ &= s^{1-(d+\eta)/2}L'^{-d/2}\sum_{k'<\Lambda}\phi_{k'}\exp(ik'\cdot x/s) \\ &= s^{1-(d+\eta)/2}\phi(x/s), \end{aligned} \tag{9.40}$$

provided that we drop the ϕ_k's with $\Lambda/s < k < \Lambda$ in the second step. Equation (9.40) says that the dimension of $\phi(x)$ is

$$d_\phi = \tfrac{1}{2}(d + \eta - 2). \tag{9.41}$$

More generally, we expect that the dimension d_A for a random variable $A(x)$ can be defined if A transform under R_s like

$$A(x) \to s^{-d_A}A(x/s). \tag{9.42}$$

However, a dimension can be defined to serve useful purposes even if (9.42) is not quite satisfied. The transformation $\phi^2 \to (\phi^2)'$ studied above furnishes an example. All we needed in obtaining the critical behavior of the correlation function K was the transformation at large s. For large s, (9.24) reads

$$\phi^2(x) \to s^{-2}t' \ \phi^2\left(\frac{x}{s}\right) \bigg/ \left(\frac{dt}{d\lambda}\right)_1 \tag{9.43}$$

if we drop the constant, which has no consequence in this case, and ignore the higher order terms in s^{-2}. Even though (9.42) does not look like (9.43), all that matters is the factor s^{-2} in front. Assigning $d_{\phi^2} = 2$ does make good sense as far as extracting critical behaviors of K is concerned.

If ϕ^2 is very large while s is not too large, (9.25) implies the transformation $\phi^2 \to s^{2-d}\phi^2$, which would naturally define $d_{\phi^2} = 2 - d$. In our application to the study of K, however, the relevant range of ϕ^2 is not large because the probability distribution P vanishes rapidly as ϕ^2 increases if $U(\phi^2)$ rises steeply as the sketches in Fig. 1 indicate. However, it is conceivable that one can construct correlation functions and probability distributions such that the range of large ϕ^2 becomes important and the dimension $2 - d$ becomes more applicable for ϕ^2 than the dimension 2.

X. BASIS FOR CALCULATION OF CRITICAL EXPONENTS BY PERTURBATION THEORY

The renormalization group analysis tells us how physical quantities such as the correlation function $G(k, \mu(T))$ should behave when T is very close to T_c, and how scaling laws appear. Our example of large n limit showed how things work out and also produced the critical exponents $\eta = O(1/n)$, $\gamma = 2/(d-2) + O(1/n)$. As was mentioned before, computing critical exponents in the large n limit, which is a trivial task, is not the purpose of our elaborate example. The purpose is to illustrate the qualitative aspects of renormalization group in a concrete fashion. Once the qualitative behavior of physical quantities is established, the critical exponents can be calculated as expansions in powers of $1/n$ by perturbation theory directly without studying the corrections to the renormalization group to higher orders in $1/n$. Consider the following example. Since we know from renormalization group arguments that at $T = T_c$

$$G^{-1}(k) \propto k^{2-\eta}(1 + O(k^{-y_2})) \tag{10.1}$$

and we found that, for large n, $\eta = O(1/n)$, $O(k^{-y_2}) = 0$ and $y_2 = d - 4 + O(1/n)$, then we can expand $k^{-\eta}$ in powers of $1/n$:

$$G^{-1}(k)k^{-2} \propto 1 - \eta \ln k + (\eta^2/2)\ln^2 k + \cdots + (1/n)O(k^{4-d}) + (1/n^2)O(k^{4-d}\ln k) + \cdots . \tag{10.2}$$

Since this is true for any μ on the critical surface, we can pick the simplest one, i.e., $\mu_1 = (t_0, u_4, 0, 0, \ldots)$, $u_4 = O(1/n)$, in Sec. V, to calculate $G(k)$ for the purpose of determining η. We can choose t_0 to make sure that μ is on the critical surface to every order we calculate. The calculation of $G^{-1}(k)k^{-2}$ will result in a power series of $(1/n)\ln k$ and η is then identified by comparison with (10.2). The coefficients of the powers of $(1/n)\ln k$ will not depend on u_4, which appears only in the $O(k^{-y_2})$ term as we argued in Sec. III. Recall that the $O(k^{-y_2})$ term reflects the approach to μ^* of $R_s\mu_1$ at a rate $s^{y_2} = s^{d-4+O(1/n)}$. This term is negligible in the critical region

$$k \ll 2^{-1/(4-d)} \tag{10.3}$$

given by (3.12). As long as d is not close to 4, the size of the critical region is of $O(1)$. The above discussion is the basis for the $1/n$ expansion of critical exponents by perturbation theory, which has been studied extensively.

It is instructive to point out a feature of the ϵ-expansion of critical exponents by perturbation theory, where one also starts with μ_1 with $u_4 = O(\epsilon)$. Complications appear because the rate of approach of $R_s\mu_1$ to the fixed point μ^* is still $s^{d-4} = s^{-\epsilon}$ for small ϵ, i.e., $y_2 = -\epsilon + O(\epsilon^2)$. The $O(k^{-y_2})$ term in (10.1) which has nothing to do with η, will also contribute a series in $\epsilon \ln k$, and one can no longer extract η by examining the coefficient of $\ln k$. In other words, the critical region vanishes in the small ϵ limit [see (10.3)]. One can get around this difficulty by choosing a special u_4 so that the s^{y_2} term in $R_s\mu_1$ vanishes. Then $R_s\mu_1$ will approach μ^* at a rate s^{y_3} which is s^{d-6} in the large n limit and the same for small ϵ and any n. With this choice, the $O(k^{-y_2})$ term vanishes and the $O(k^{y_3}) = O(k^{2+O(\epsilon)})$ term will not give rise to any $\ln k$. Effectively, the critical region is extended to $O(1)$. In the large n limit, this special u_4 is easily found from (6.28) or (5.44) by setting the coefficient of s^{d-4} to zero. We obtain

$$\begin{aligned} u_4 &= \left(\frac{4-d}{d-2}\right)\Big/ N_c + O(n^{-2}) \\ &= (2\epsilon/nK_{4-\epsilon}) + O(n^{-2}). \end{aligned} \tag{10.4}$$

It is important to remember that R_s does depend on how the cutoff is defined (see Sec. III). The special u_4 given by (10.4) has the same meaning as Wilson's $u_0(\epsilon)$ but appears different because of the different way of doing the cut off as well as the difference by a proportionality constant in definition.[4]

XI. CONCLUDING REMARKS

We have defined and illustrated the basic concepts and the working of the renormalization group in the context of classical statistical mechanics. In conclusion, we re-emphasize the following outstanding features.

(1) The renormalization group is a set of transformations R_s, $s \geq 1$, of *coupling parameters*, which are nonsingular. The singularities in observed quantities, which are average values calculated over probability distributions specified by the coupling parameters, appear as a result of large s behavior of R_s. The concept of scaling appears in the scale change which is a part of R_s. Universality is the statement that critical exponents are properties of R_s near a fixed point μ^* and is to a great extent independent of the details of the microscopic Hamiltonian.

(2) The renormalization group is precisely defined and its related concepts are concrete, but the mathematical machinery is complicated. Very little rigorous results exist at present to help handle this machinery.

(3) In spite of its usefulness, perturbation theory plays no part in the basic concept of the renormalization group and is not essential to the machinery. A valuable feature of the large n limit is that an infinite set of graphs can be summed exactly to illustrate many nonperturbative features of the renormalization group.

(4) In the large n limit, we are able to visualize an explicit and exact fixed point μ^* and the critical surface extending from it. We demonstrated how $R_s\mu$ approaches μ^* for μ being anywhere on the critical surface (not necessarily in the immediate vicinity of μ^*). This feature, which is fundamental to the idea of universality, makes the large n limit a more illustrative example than the small $\epsilon = 4 - d$ limit sample, where such a visualization is less transparent.

(5) Most important of all, the renormalization group shows besides the origin of scaling and universality hypotheses, precisely where the limitation and weakness of these hypotheses are. Our discussion on the free energy serves as an illustration of this aspect.

The renormalization group is not a hypothesis. It provides a basis for quantitative calculations, which are unfortunately complicated. What we have touched upon in this paper are most elementary aspects illustrated by a somewhat idealized model. Hopefully this paper has helped visualize these elementary aspects of the renormalization group.

ACKNOWLEDGMENTS

I am grateful to my colleagues at UCSD for helpful comments. I thank Ms. Annetta Whiteman for her masterful work in preparing this manuscript. This research is supported in part by the National Science Foundation.

*Alfred P. Sloan Foundation Fellow.

[1] Extensive references on the subject of renormalization group, both old and new, can be found in K. G. Wilson and J. Kogut, "The Renormalization Group and the ϵ Expansion," Phys. Rep. (to be published).

[2] The ground work was laid out by Wilson in Phys. Rev. B 4, 3174 (1971). Further numerical investigation has been carried out by many authors. See, for example, M.K. Grover, L.P. Kadanoff, and F. J. Wegner Phys. Rev. B 6, 311 (1972); G. Golner, Phys. Rev. B 8, 339 (1973).

[3] K. G. Wilson and M.E. Fisher, Phys. Rev. Lett. 28, 240 (1972).

[4] K. G. Wilson, Phys. Rev. Lett. 28, 548 (1972); E. Brezin, D. Wallace, and K. G. Wilson, Phys. Rev. Lett. 29, 591 (1972); B. G. Nickel, Phys. Rev. (to be published); E. Brezin, J.C. LeGuillou, and J. Zinn-Justin, Phys. Rev. B 8, 5530 (1973).

[5] H. E. Stanley, Phys. Rev. 176, 718 (1968).

[6] See, for example, R. Abe, Prog. Theor. Phys. 48, 1414 (1972); R. Abe and S. Hikami, Phys. Lett. A 42, 419 (1973); and Prog. Theor. Phys. 49, 442 (1973); K.G. Wilson, Phys. Rev. D 7, 2911 (1973); R.A. Ferrell and D.J. Scallapino, Phys. Rev. Lett. 29, 413 (1972); S. Ma, Phys. Rev. Lett. 29, 1311 (1972); M. Suzuki, Phys. Lett. A 42, 5 (1972), Prog. Theor. Phys. 49, 424, 1017 (1973).

[7] S. Ma, Phys. Rev. A, Phys. Rev. A 7, 2172 (1973).

[8] E. Brezin and D. Wallace, Phys. Rev. B 7, 1967 (1973).

[9] Much of the elementary discussions here was later used (by consent of the editor) as a part of a review, S. Ma, Rev. Mod. Phys. 45, 589 (1973). To preserve continuity and completeness of the present paper, we have made no attempt to eliminate duplication of material.

[10] For a review of earlier work on critical phenomena, see M.E. Fisher, Rep. Prog. Phys. 30, 615 (1967); L.P. Kadanoff *et al.*, Rev. Mod. Phys. 39, 395 (1967).

[11] In other words, we assume $t_1(T)$ can be expanded in a Taylor series $t_1 = A(T-T_c) + B(T-T_c)^2 + \cdots$ and also assume that $A \neq 0$.

[12] The $u'_{2(m+1)}$ here differs from that defined by (4.21) by a numerical factor so that Σ' can have a simpler appearance. [See (5.29).]

[13] A. Erdelyi *et al.*, *Higher Transcendental Functions*, (McGraw-Hill, New York, 1953) Vol. I, pp. 27, 30.

[14] Details of calculating $\Pi(k)$ can be found in Ref. 7. Note the difference in cutoff.

[15] F.J. Wegner, Phys. Rev. B 5, 4529 (1972).

[16] For simplicity of notation, we shall not write out the component labels i explicitly.

[17] The scaling law $2-\alpha = \nu d$ follows from (8.12). The exponent α is defined by the behavior of the specific heat near T_c taken as $-\partial^2 F/\partial T^2 \propto |T-T_c|^{-\alpha}$.

[18] For d very close to 4, (8.20) holds. The second inequality should be generalized to $\nu d < 2 - y_2 < \nu$.

[19] See the work of Abe and Hikami (Ref. 6) for details in this direction. See also Ref. 7, Appendix C.

[20] A more general discussion can be found in S. Ma, "Scaling Variables and Dimensions" (to be published).

[21] In subsequent formulas, $t(\phi^2)$ shall mean $t(\phi^2/N_c)$, i.e., $t(\lambda)$ with $\lambda = \phi^2/N_c$.

Quantization on hyperboloids and full space-time field expansion

A. diSessa

Laboratory for Nuclear Science and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 4 April 1974)

Beginning with free field quantization on hyperboloids in the forward lightcone, we extend field expansions to the full space-time. The role of boundary conditions on the propagator and of the particle-antiparticle distinction in establishing a unique field expansion is discussed. Some difficulties with the possibility of a surface to surface development of the S matrix are encountered in the massless case. Even in the massive case hyperboloids are unsuitable quantization surfaces when $x^2<0$, and we develop in some detail an alternative set of surfaces for that region in two dimensions.

I. INTRODUCTION

Recently there has been a revival of interest in the years-old suggestion by Dirac that quantization may be carried out on hyperboloids $x^2=\text{const}$. Fubini, Hanson, and Jackiw[1] first explicitly quantized fields on $x^2=\text{const}$ in Euclidean space where the surfaces of quantization are spheres. More recently Sommerfield[2] and Gromes, Rothe and Stech[3] (GRS) have independently quantized fields on hyperboloids in Lorentz space with rather different viewpoints. Both of their quantizations, however, give rather light treatment to the extension of the field outside the forward light cone. It will be the prime goal of this paper to obtain field expansions valid in all regions of space—time. The considerations which arise in this context shed light on the nature of the assumptions made by GRS (which we are led to make also) and their relation to the rather different assumptions made by Sommerfield.

We will quantize a real scalar field in two and four dimensions and a spin 1/2 field in two dimensions. The initial quantization will be in the forward light cone with canonical commutation relations given on surfaces of constant $x^\mu x_\mu$. The commutators are insufficient to determine a unique quantum expansion, but a polarization of the solution space of the field equation into positive and negative frequency solutions associated with destruction and creation operators, respectively, suffices to determine the expansion uniquely. This particular polarization is of course Poincaré invariant and thus our particle-antiparticle distinction will also be invariant. The polarization also insures that the boundary conditions of this quantization as expressed in the propagator function are identical to the usual equal time quantization. Our propagator, the vacuum expectation of $x^\mu x_\mu$ ordered (in the forward light cone) fields, will therefore be the same distribution as the usual one.

For $m\neq 0$ knowledge of the field in the forward lightcone is sufficient to determine the field everywhere providing one removes unphysical solutions which do not vanish at spacelike infinity. We can therefore uniquely extend the field to all of space—time. With this full space—time field we will show that translation invariance of the propagator requires that the field ordering in the region $x^\mu x_\mu<0$ be equivalent to time ordering (as is $x_\mu x^\mu$ ordering in the future, $x_\mu x^\mu>0$, $t>0$). Thus in $x^2<0$ as well as $x^2>0$ the development of the field from surface to surface must take place on spacelike surfaces, in particular not on hyperboloids $x^\mu x_\mu=-\text{const}$.

In the case $m=0$ one must add initial data on $x^\mu x_\mu$ in the backward light cone to data in the forward cone in order to specify a unique field everywhere. This makes it clear that the $m=0$ field cannot be achieved as the massless limit of the $m\neq 0$ field in contrast to Euclidean $x^\mu x_\mu$ quantization.[4] Nevertheless, one finds again that a translation invariant propagator requires a time ordered product in all space—time.

Sections II through V will concern only the two-dimensional case. Section VI will briefly consider the four-dimensional case for the scalar field.

II. MASSIVE SCALAR FIELD IN THE FORWARD LIGHT CONE

The field equation for a massive spin zero particle in two dimensions is, of course

$$(\Box+m^2)\Phi=0,\quad \Box=\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial x^2}=\frac{\partial}{\partial x^\mu}\frac{\partial}{\partial x_\mu}\,.$$

In the forward light cone, it will be extremely convenient to use coordinates α,β as follows:

$$e^{\alpha+\beta}=t+x=\sqrt{2}x_+,$$

$$e^{\alpha-\beta}=t-x=\sqrt{2}x_-,$$

where α is the group parameter of the dilation group which will parameterize development from surface to surface,[5] and β is the Lorentz group parameter. A Lorentz transformation is specified by $\beta\to\beta+\beta_0$ which of course leaves the quantization surfaces, $x^2=\exp(2\alpha)=\text{const}$, invariant. To simplify writing we can set $m=1$ which is equivalent to rescaling the coordinates $x_\mu\to x_\mu/m$. The field equation becomes

$$\left(\frac{\partial^2}{\partial\alpha^2}-\frac{\partial^2}{\partial\beta^2}+\exp(2\alpha)\right)\Phi=0.\qquad\text{(II.1)}$$

With the equation in this form it is obvious that we may introduce a canonical Hamiltonian structure which is form identical to equal time quantization except that mass is "time," i.e. α, dependent.

We posit then the equal α quantum relations:

$$[\Phi(\alpha,\beta),\Phi(\alpha,\beta')]=0,$$

$$[\pi(\alpha,\beta),\Phi(\alpha,\beta')]=-i\delta(\beta-\beta'),\qquad\text{(II.2)}$$

$$[\pi(\alpha,\beta),\pi(\alpha,\beta')]=0,$$

$$\pi=\frac{\partial\Phi}{\partial\alpha}.$$

The "Hamiltonian" or developmental generator which we denote by D is actually the generator of the dilation group

$$D(\alpha)=\frac{1}{2}\left[\pi^2+\left(\frac{\partial\Phi}{\partial\beta}\right)^2+\exp(2\alpha)\Phi^2\right]. \tag{II.3}$$

It is easy to check from (II.3) that $\partial D/\partial\alpha\neq 0$ in general. This occurs because D generates a broken symmetry. The fact that it is not a conserved charge, however, does not alter D's position as developmental generator; it generates the equations of motion.

$$i[D(\alpha),\Phi(\alpha,\beta)]=\frac{\partial\Phi}{\partial\alpha},$$

$$i[D(\alpha),\pi(\alpha,\beta)]=\frac{\partial^2\Phi}{\partial\alpha^2}=\frac{\partial^2\Phi}{\partial\beta^2}-\exp(2\alpha)\Phi.$$

To obtain a quantum expansion we first perform a Lorentz harmonic analysis of the classical field:

$$\Phi=\int_{-\infty}^{\infty}d\gamma\exp(i\beta\gamma)A_\gamma(\alpha).$$

This results in an equation for A_γ which is easily reduced to a Bessel equation:

$$\left(\frac{d^2}{d\alpha^2}+\gamma^2+\exp(2\alpha)\right)A_\gamma=0,$$

$$\left(\frac{d^2}{d\tau^2}+\frac{1}{\tau}\frac{\partial}{\partial\tau}+\frac{\gamma^2}{\tau^2}+1\right)A_\gamma=0,\quad \tau=e^\alpha.$$

We take independent solutions A_γ proportional to $H_{i\gamma}^{(1)}(\tau)$ or $H_{i\gamma}^{(2)}(\tau)$, Hankel functions of first and second kinds. We now have a complete set of functions for the classical field $\exp(i\beta\gamma)H_{i\gamma}^{(1)}(e^\alpha)$, $\exp(i\beta\gamma)H_{i\gamma}^{(2)}(e^\alpha)$.

The solution space of (II.1) is endowed with a natural Hermitian structure

$$(\Phi,\chi)=i\int d\beta\left(\Phi^*\frac{\partial\chi}{\partial\alpha}-\chi\frac{\partial\Phi^*}{\partial\alpha}\right). \tag{II.4}$$

With respect to this inner product it is trivial to verify that because of the Wronskian relation

$$H_\nu^{(1)}(\tau)\frac{d}{d\tau}H_\nu^{(2)}(\tau)-H_\nu^{(2)}(\tau)\frac{d}{d\tau}H_\nu^{(1)}(\tau)=\frac{-4i}{\pi\tau}$$

the following complete set of solution functions is orthonormal:

$$\begin{aligned}f_\gamma&=\frac{-i}{\sqrt{8}}H_{i\gamma}^{(2)}(\tau)\exp(i\beta\gamma)\exp(\pi/2)\gamma,\\ f_\gamma^*&=\frac{i}{\sqrt{8}}(H_{i\gamma}^{(2)}(\tau))^*\exp(-i\beta\gamma)\exp(\pi/2)\gamma\\ &=\frac{i}{\sqrt{8}}H_{-i\gamma}^{(1)}(\tau)\exp(-i\beta\gamma)\exp(\pi/2)\gamma,\end{aligned} \tag{II.5a}$$

$$\begin{aligned}&(f_\gamma,f_{\gamma'})=\delta(\gamma-\gamma'),\\ &(f_\gamma^*,f_{\gamma'}^*)=-\delta(\gamma-\gamma'),\\ &(f_\gamma,f_{\gamma'}^*)=0=(f_\gamma^*,f_{\gamma'}).\end{aligned} \tag{II.5b}$$

We expect a quantum expansion of the form

$$\Phi(\alpha,\beta)=\int d\gamma\, f_\gamma^*(\alpha,\beta)a_\gamma^\dagger+f_r(\alpha,\beta)\,a_\gamma,$$

$$[a_\gamma,a_\gamma^\dagger]=\delta(\gamma-\gamma'). \tag{II.6}$$

In fact, if f_γ is given by (II.5a), the canonical commutation relations will be satisfied. The nonuniqueness of this expansion, however, is clear. The total commutator is given by

$$[\Phi(x),\Phi(x')]=\int d\gamma(f_\gamma(x)f_\gamma^*(x')-f_\gamma(x')f_\gamma^*(x))$$

and is invariant (hence, so are the canonical commutators) under the following transformation:

$$\begin{aligned}f_\gamma&\to c(\gamma)f_\gamma+d(\gamma)f_\gamma^*,\\ f_\gamma^*&\to c^*(\gamma)f_\gamma^*+d^*(\gamma)f_\gamma,\end{aligned} \tag{II.7}$$

c and d are arbitrary functions of γ which satisfy

$$c(\gamma)c^*(\gamma)-d(\gamma)d^*(\gamma)=1.$$

This last is precisely the condition which also insures the preservation of the orthonormality relations (II.5b).

In order to specify a unique expansion we look to the propagator. Our propagator will be given by the developmentally ordered vacuum expectation of the field product

$$\begin{aligned}&\langle 0|D\Phi(x)\Phi(x')|0\rangle\\ &=\langle 0|\Phi(x)\Phi(x')|0\rangle\,\theta(\alpha-\alpha')+\langle 0|\Phi(x')\Phi(x)|0\rangle\,\theta(\alpha'-\alpha).\end{aligned} \tag{II.8}$$

Correspondingly, we need to define a Fock space generated from the vacuum $|0\rangle$ by creation operators $a_\gamma^\dagger$. We require the usual relations

$$a_\gamma|0\rangle=0=\langle 0|a_\gamma^\dagger.$$

Now one can check that the transformation (II.7) will change the propagator but only by adding a real homogeneous solution of the field equation, namely, $\int d\gamma[f_\gamma^*(x)f_\gamma(x')+f_\gamma(x)f_\gamma^*(x')]d(\gamma)d^*(\gamma)$.[6] Thus it is the boundary conditions we impose on the propagator which determine (up to inconsequential phases) exactly which of the infinitely many possible f's satisfying (II.2) will appear in the field expansion. In this paper we will make the natural assumption that the boundary conditions are the same as in equal time quantization. This is equivalent to insisting the f_γ are positive frequency solutions as later calculations will verify.

As an alternative procedure for unique specification of f_γ, f_γ^* one can simply require that the particle—antiparticle distinction be Poincaré invariant. This means that the subspace spanned by f_γ must be Poincaré invariant and orthogonal to that spanned by f_γ^*. Such a decomposition of the solution space is unique and is the usual positive—negative frequency distinction, the same polarization which we were led to from boundary condition considerations.

GRS use the positive—negative frequency distinction to determine a unique field expansion. Sommerfield on the other hand uses not time, but dilation frequency. This leads him to a theory where "particle" and "antiparticle" are names which, though Lorentz invariantly defined, do not coincide with the usual *Poincaré* invariant choice. It naturally follows that his propagator, which he does not discuss, will be different from the usual one.

We can identify positive frequency solutions from the property that $f(x)\to 0$ as $t\to -i\infty$. In terms of α and β this limit is $\alpha\to -i\pi/2+\infty$ or $\tau=e^{\alpha}\to -i\infty$. We check that our original choice for f provides the proper positive–negative frequency polarization:

$$f_{\gamma}(\alpha,\beta)=\frac{-i}{\sqrt{8}}\exp(i\beta\gamma)\exp(\pi\gamma/2)H_{i\gamma}^{(2)}(\tau)$$

$$\to\frac{i}{\sqrt{8}}\exp(i\beta\gamma)\exp(\pi\gamma/2)H_{i\gamma}^{(2)}[\exp(-i\pi/2)\infty]$$

$$=\frac{1}{\sqrt{2\pi}}\exp(i\beta\gamma)K_{i\gamma}(\infty)=0.$$

The independent solution $f_{\gamma}^{*}(\alpha,\beta)$ diverges in this limit but goes to zero as $t\to i\infty$, hence f_{γ} and f_{γ}^{*} are positive and negative frequency, respectively.

Finally, we can compute explicitly the propagator and commutator with the following integral[7]:

$$\int_{-\infty}^{\infty}d\gamma f_{\gamma}(x)f_{\gamma}(x')=\tfrac{1}{8}\int_{-\infty}^{\infty}d\gamma\exp[i\gamma(\beta-\beta')]$$

$$\times\exp(\pi\gamma)H_{i\gamma}^{(2)}(e^{\alpha})H_{-i\gamma}^{(1)}(e^{\alpha'})$$

$$=\frac{-i}{4}H_0^{(2)}([\exp(2\alpha)+\exp(2\alpha')-2\exp(\alpha+\alpha')$$

$$\times\cosh(\beta-\beta')-i(\alpha-\alpha')\epsilon]^{1/2})$$

$$=\frac{-i}{4}H_0^{(2)}([(x-x')^2-i(\alpha-\alpha')\epsilon]^{1/2}). \qquad \text{(II.9)}$$

Evaluating (II.8) we find

$$\langle 0|D\Phi(x)\Phi(x')|0\rangle=\frac{-i}{4}H_0^{(2)}([x-x')^2-i\epsilon]^{1/2}). \qquad \text{(II.10)}$$

This agrees with the usual propagator

$$\frac{i}{(2\pi)^2}\int d^2k\,\frac{\exp(-ik\cdot x)}{(k^2-1+i\epsilon)}=\frac{-i}{4}H_0^{(2)}([x^2-i\epsilon]^{1/2}).$$

The full commutator also agrees with the usual one:

$$[\Phi(x),\Phi(x')]=\frac{-i}{2}\theta((x-x')^2)\,\epsilon(\alpha-\alpha')\,J_0([(x-x')^2]^{1/2})$$

$$=\frac{-i}{2}\theta((x-x')^2)\,\epsilon(t-t')\,J_0([x-x')^2]^{1/2}). \qquad \text{(II.11)}$$

III. EXTENSION OUTSIDE THE FORWARD LIGHT CONE

In order to extend the field outside the forward lightcone we can express each of the set of functions f_{γ}, known inside the cone, in terms of solutions valid in all regions, $\exp(ik\cdot x)$.[8] We have no reason to presume initially that such a Fourier representation will be unique. In the massless case it will not be! We insist that the solutions f_{γ} be positive frequency, hence their Fourier representations are analytic as functions of x and t as long as $\operatorname{Im}t<0$ and $|\operatorname{Im}x|<|\operatorname{Im}t|$. Thus, though we initially know f_{γ} only in the region x_+, $x_->0$, we can find the values of the function in another region of space–time provided there exists a path from x_+, $x_->0$ (which we call region I) to the other region which lies entirely within the domain of analyticity of the Fourier representation. Representative paths for this analytic continuation are given below in terms of a real parameter, ϕ, which varies from 0 to 1. The salient point of the continuations is the avoidance of the branch point of the Bessel functions at $\tau=e^{\alpha}=0$ with a clockwise route for positive frequency solutions.[9]

The extensions of f_{γ}^{*} are just the complex conjugate of the extensions of f_{γ} and (II.6) is now a field expansion in all of space–time.

If we expect the ordered product of fields to play the invariant role of propagator in the quantization we must require it to have translation invariance. For example, in region II as in region I we want to have

$$\Delta(x,x')=\frac{-i}{4}H_0^{(2)}([(x-x')^2-i\epsilon]^{1/2})$$

$$=\frac{1}{2\pi}K_0([-(x-x')^2+i\epsilon]^{1/2})$$

$$=\langle 0|D\Phi(x)\Phi(x')|0\rangle$$

From (II.6) and Table I we can compute,[7] in region II,

$$\langle 0|\Phi(x)\Phi(x')|0\rangle=\frac{1}{2\pi^2}\int_{-\infty}^{\infty}d\gamma f_{\gamma}(\alpha,\beta)f_{\gamma}^{*}(\alpha',\beta')$$

$$=\frac{1}{2\pi^2}\int_{-\infty}^{\infty}d\gamma\exp[i\gamma(\beta-\beta')]\exp(-\pi\gamma)K_{i\gamma}(e^{\alpha})K_{i\gamma}(e^{\alpha'})$$

$$=\frac{1}{2\pi}K_0([\exp(2\alpha)+\exp(2\alpha')-2\exp(\alpha+\alpha')$$

$$\times\cosh(\beta-\beta')+i(\beta-\beta')\epsilon]^{1/2}). \qquad \text{(III.1)}$$

In this region $(x-x')^2=-\exp(2\alpha)-\exp(2\alpha')+2\exp(\alpha+\alpha')\cosh(\beta-\beta')$. Focusing attention on the branch specifying term $i(\beta-\beta')\epsilon$, in (III.1) one sees that we can identify the propagator with a developmentally ordered product only if developmental ordering agrees with β ordering. As surfaces of constant β are spacelike here, β ordering is equivalent to time ordering (as is a ordering in region I). If we want a developmental picture of the field evolving from surface to proximate surface, the surfaces we take in region II cannot be hyperboloids since they do not provide the proper ordering.

Similar reasoning requires ordering in region III to be along $-\beta$ and in region IV along $-\alpha$. Since fields commute at spacelike separations, all of these orderings are equivalent to time ordering. Even though our development is not time development, the ordering must be equivalent to time ordering. Having drawn that conclusion, one can verify that in *and between* all regions

TABLE I. Continuation from forward lightcone.

Region	$t+x$	$t-x$	Path from Region I	f_{γ}
I	$e^{\alpha+\beta}$	$e^{\alpha-\beta}$		$\frac{-i}{\sqrt{8}}e^{i\beta\gamma}e^{\pi\gamma/2}H_{i\gamma}^{(2)}(e^{\alpha})$
II	$e^{\alpha+\beta}$	$-e^{\alpha-\beta}$	$\alpha(\phi)=\alpha-\frac{i\pi}{2}\phi$ $\beta(\phi)=\beta+\frac{i\pi}{2}\phi$	$\frac{1}{\sqrt{2}\pi}e^{i\beta\gamma}e^{-\pi\gamma/2}K_{i\gamma}(e^{\alpha})$
III	$-e^{\alpha+\beta}$	$e^{\alpha-\beta}$	$\alpha(\phi)=\alpha-\frac{i\pi}{2}\phi$ $\beta(\varphi)=\beta-\frac{i\pi}{2}\phi$	$\frac{1}{\sqrt{2}\pi}e^{i\beta\gamma}e^{\pi\gamma/2}K_{i\gamma}(e^{\alpha})$
IV	$-e^{\alpha+\beta}$	$-e^{\alpha-\beta}$	$\alpha(\phi)=\alpha-i\pi\phi$ $\beta(\phi)=\beta$	$\frac{i}{\sqrt{8}}e^{i\beta\gamma}e^{-\pi\gamma/2}H_{i\gamma}^{(1)}(e^{\alpha})$

$$\langle 0|D\Phi(x)\,\Phi(x')|0\rangle=\Delta(x,x')=\frac{-i}{4}H_0^{(2)}([(x-x')^2-i\epsilon]^{1/2}).$$

We note that the set of integrals which verify the above statement may be derived by analytic continuation from the single integral below[10]:

$$\int_{-\infty}^{\infty} d\gamma\exp[i\gamma(\beta-\beta')]K_{i\gamma}(e^{\alpha})K_{-i\gamma}(e^{\alpha'})$$
$$=\pi K_0([\exp(2\alpha)+\exp(2\alpha')+2\exp(\alpha+\alpha')\cosh(\beta-\beta')]^{1/2})$$
$$|\mathrm{Im}(\beta-\beta')|+|\mathrm{Im}\,\alpha|+|\mathrm{Im}\,\alpha'|\leqslant\pi. \qquad \text{(III. 2)}$$

For example, the integral (III. 1) is obtained with $\beta\rightarrow\beta+i(\pi/2-\epsilon)$, $\beta'\rightarrow\beta'-i(\pi/2-\epsilon)$. It should be no surprise that the full commutator (II. 11) is also valid in all space–time, thanks to the same integral.

In a picture of field development, regions II and III should be taken together since knowledge of the field in either region alone is insufficient to determine the field everywhere. In region II we can write the field expansion

$$\Phi=\frac{1}{\sqrt{2}\pi}\int_{-\infty}^{\infty}d\gamma\exp(i\beta\gamma)K_{i\gamma}(e^{\alpha})\,[\exp(-\pi\gamma/2)\,a_{\gamma}$$
$$+\exp(\pi\gamma/2)\,a_{-\gamma}^{\dagger}].$$

Evidently only the linear combination $\exp(-\pi\gamma/2)a_{\gamma}+\exp(\pi\gamma/2)a_{-\gamma}^{\dagger}$ can be recovered from the field, not a_{γ} and $a_{\gamma}^{\dagger}$ separately. Region III provides knowledge of the linearly independent combination $\exp(-\pi\gamma/2)a_{\gamma}+\exp(\pi\gamma/2)a_{-\gamma}^{\dagger}$.

All the considerations of this section can be joined into a coherent picture of field evolution as follows: We shall introduce a single parameter of development q and a "spacelike" or kinematic coordinate p defined differently in each region.

In the past, evolution from surface to surface is on hyperboloids with $q=-\alpha$, $p=\beta$. Reaching the past part of the light cone at $q=\infty$, we begin a new development in the elsewhere simultaneously in II and III with $q=\beta$, $p=\alpha$ in II and $q=-\beta$, $p=\alpha$ in III. The future development is in terms of $q=\alpha$, $p=\beta$.

All the surfaces $q=\text{const}$ are spacelike and ordering within past, elsewhere, and future is given by q ordering, equivalent to time ordering. Ordering among the three reqions is obviously past, to elsewhere, to future. Finally, surfaces of neighboring q are topologically neighboring in Lorentz space so that when interactions are introduced, one still has a well-defined local development from surface to proximate surface.

The Hermitian structure defined in (II. 4) can be extended to all of space–time consistently with q development:

$$(\Phi,\chi)=i\int dp\left(\Phi^*\frac{\partial\chi}{\partial q}-\chi\frac{\partial\Phi^*}{\partial q}\right). \qquad \text{(III. 3)}$$

The integral is over the entire $q=\text{const}$ surface so in the elsewhere it is the sum of $\int d\alpha$ in region II and $\int d\alpha$ in region III. The orthonormality relations (II-5b) are now valid over the entire development from $q=-\infty$ past to $q=\infty$ future. The Hankel function Wronskian establishes (II. 5b) in the future and past. In the elsewhere one needs the integral[11]

$$\frac{2}{\pi^2}\int_{-\infty}^{\infty}d\alpha\,K_{i\gamma}(e^{\alpha})K_{i\gamma'}(e^{\alpha})=\frac{1}{\gamma\sinh(\pi\gamma)}\,\delta(\gamma-\gamma').$$

To complete the picture, canonical commutation relations can be verified with $\pi=\partial\Phi/\partial q$:

$$[\pi(p,q),\Phi(p',q)]=-i\delta(p-p'),$$
$$[\pi(p,q),\pi(p',q)]=0=[\Phi(p,q),\Phi(p',q)]=0.$$

The developmental operator is

$$D(\alpha)=\frac{1}{2}\int dp\left[\left(\frac{\partial\Phi}{\partial p}\right)^2+\left(\frac{\partial\Phi}{\partial q}\right)^2\pm\exp(2\alpha)\,\Phi^2\right]\cdot$$

The + sign is for the future and past and the – sign is needed for the proper development in the elsewhere.

The extension of forward light cone quantization that we developed in this section uses $\beta=\text{const}$ surfaces in the elsewhere. Though this was convenient because of the coordinate system we used, it is by no means necessary. Similar results should be obtainable from any appropriate spacelike set of surfaces.

IV. MASSLESS SCALAR FIELD

The massless scalar field cannot be achieved as a limit $m\rightarrow 0$ of the massive field. After reintroducing the mass parameter in the massive field expansion functions with the substitution $x_{\mu}\rightarrow mx_{\mu}$ $(e^{\alpha}\rightarrow me^{\alpha})$, one can easily check that there is no asymptotic function associated with the limit $m\rightarrow 0$. Fortunately, the massless field in two dimensions is trivial to solve, and our discussion will thereby be much simplified:

$$\Box\Phi=2\frac{\partial}{\partial x_+}\frac{\partial}{\partial x_-}\Phi=0,\quad \Phi=\Phi_+(x_+)+\Phi_-(x_-). \qquad \text{(IV. 1)}$$

Φ_+ and Φ_- are arbitrary functions of their arguments. From this fact it is easy to guess (and check) that in a harmonic decomposition of the field $\Phi=\int d\gamma\exp(i\beta\gamma)A_{\gamma}(\alpha)$ the functions $A_{\gamma}(\alpha)\exp(i\beta\gamma)$ must be proportional to $(x_+)^{i\gamma}$ or $(x_-)^{-i\gamma}$.

It is clear from the onset that the functions $(x_{\pm})^{i\gamma}$ (or linear combinations) are not intrinsically positive or negative frequency in the sense that the limits $t\rightarrow\pm i\infty$ are oscillatory. This prohibits any unique extension of the field expansion outside the forward light cone. Of course the impossibility also follows from the fact that $\Phi_{\pm}$ in the future where x_+, $x_->0$ do not determine $\Phi_{\pm}$ when $x_{\pm}<0$. Though the Cauchy problem in the future alone is well-defined, we want a field expansion in all space–time. Accordingly, we shall take the task of first expressing the propagator in all regions in harmonically decomposed form and from it deduce the field expansion.

We must now choose the boundary conditions on the propagator. This causes some difficulty as the momentum space representation of the two-dimensional propagator diverges. We can take the standard approach, however, with a cutoff near zero momentum and ignoring the remaining (infinite) constant one has

$$\Delta(x)=-\frac{1}{4\pi}\log(x^2-i\epsilon). \qquad \text{(IV. 2)}$$

Alternatively, this is the $m\rightarrow 0$ limit of the massive propagator $(-i/4)H_0^{(2)}([(mx)^2-i\epsilon]^{1/2})$, neglecting the infinite constant $\lim_{m\to 0} -(1/4\pi)\log m^2$.

Though $(x_\pm)^{i\gamma}$ are complete set for square integrable functions (i.e., square integrable Cauchy data) on hyperboloids, the logarithmic propagator is not in that space of functions. We will need to add a discrete point to the spectrum of the harmonic decomposition of the field with functions 1 and $\log x_\pm$. Unfortunately, these solutions will not allow the field to be Hermitian in this $\gamma=0$ subspace, as we shall see.

We now write the unique harmonic decomposition of the propagator $-(1/4\pi)\log[(x-x')^2 - i\epsilon]$:

$$\Delta(x;x') = \begin{cases} \dfrac{1}{8\pi}\sum_{j=\pm}\left[\int_{-\infty}^{\infty} d\gamma\,\dfrac{\exp(\pi\gamma)}{\gamma\sinh\pi\gamma}\left(\dfrac{x'\cdot n_j + i\epsilon}{x\cdot n_j - i\epsilon}\right)^{i\gamma} - \log(x\cdot n_j - i\epsilon) - \log(x'\cdot n_j + i\epsilon)\right], & t > t' \\ \dfrac{1}{8\pi}\sum_{j=\pm}\left[\int_{-\infty}^{\infty} d\gamma\,\dfrac{\exp(\pi\gamma)}{\gamma\sinh\pi\gamma}\left(\dfrac{x\cdot n_j + i\epsilon}{x'\cdot n_j - i\epsilon}\right)^{i\gamma} - \log(x'\cdot n_j - i\epsilon) - \log(x\cdot n_j + i\epsilon)\right], & t' > t. \end{cases} \tag{IV.3}$$

For the purposes of comparison with the four-dimensional case, we use a slightly more general notation in the following formulas. We use $x\cdot n_+ = t + x$ $[n_+ = (1,-1)]$ and $x\cdot n_- = t - x$ $[n_- = (1,1)]$ instead of x_+ and x_-.

The pole at $\gamma=0$ is regulated by a standard principle value regulation. We have chosen to use the positive and negative frequency extensions of $(x\cdot n_j)^{i\gamma}$, namely $(x\cdot n_j \mp i\epsilon)^{i\gamma}$, though at this point we could just as well have used, for example, even and odd extensions $|x\cdot n_j|^{i\gamma}$ and $|x\cdot n_j|^{i\gamma}\epsilon(x\cdot n_j)$.

With this propagator expansion it is not difficult to deduce the field expansion by demanding it satisfy

$$\langle 0|T\Phi(x)\Phi(x')|0\rangle = \Delta(x,x'). \tag{IV.4}$$

The use of T, time ordering, follows the considerations of the previous section and will be more thoroughly discussed shortly.

One finds

$$\Phi = \sum_{j=\pm}\left(\int_{-\infty}^{\infty} d\gamma (f_\gamma(x\cdot n_j)\,a_j(\gamma) + f_\gamma^*(x\cdot n_j)\,a_j^\dagger(\gamma)) + \frac{1}{\sqrt{8\pi}}(\log(x\cdot n_j + i\epsilon)\,a_j^\dagger - a_j + \log(x\cdot n_j - i\epsilon)\,b_j - b_j^\dagger)\right),$$

$$f_\gamma(x\cdot n_j) = \frac{1}{\sqrt{8\pi}}\left(\frac{\exp(-\pi\gamma)}{\gamma\sinh\pi\gamma}\right)^{1/2}(x\cdot n_j - i\epsilon)^{i\gamma}. \tag{IV.5}$$

The a's and b's as usual satisfy

$$[a_j(\gamma), a_{j'}^\dagger(\gamma')] = \delta(\gamma-\gamma')\,\delta_{j,j'},$$

$$[a_j, a_{j'}^\dagger] = \delta_{j,j'}, \quad [b_j, b_{j'}^\dagger] = \delta_{j,j'},$$

$$a_j(\gamma)|0\rangle = a_j|0\rangle = b_j|0\rangle = 0.$$

Our laxness in the consideration of cummutators in this section can be justified by the observation that $\Box_x\Delta(x,x') = -i\delta^2(x-x')$ implies, for example, in the forward light cone

$$\exp(2q)\Box_x\langle 0|T\Phi(x)\Phi(x')|0\rangle = \left(\frac{\partial^2}{\partial q^2} - \frac{\partial^2}{\partial\beta^2}\right)\langle 0|T\Phi(x)\Phi(x')|0\rangle$$
$$= \left\langle 0\left|\left[\frac{\partial\Phi(x)}{\partial q}, \Phi(x')\right]\right|0\right\rangle\delta(q-q') = -i\delta(\beta-\beta')\,\delta(q-q'). \tag{IV.6}$$

Since the vacuum expectation of the commutator is the commutator at least for the free field, the T product implies canonical hyperboloidal commutators. In fact, direct computation shows the field (IV-5) gives the usual full commutator

$$[\Phi(x), \Phi(x')] = \frac{-i}{2}\,\theta((x-x')^2)\,\epsilon(t-t'). \tag{IV.7}$$

The $\gamma\neq 0$ part of the field expansion was chosen to be Hermitian while there seems no way to make this so for the discrete part. In the following we ignore the discrete part as it seems characteristic only of the two-dimensional massless scalar field and is therefore not of importance for general considerations.

The field again admits a Hermitian form in terms of Cauchy data on hyperboloids:

$$(\Phi,\chi) = i\sum_{\mu=\pm}\int_{x^2=\mathrm{const}} d\beta\left(\Phi^*\epsilon(x_\mu)x_\mu\frac{\partial\chi}{\partial x_\mu} - \chi\,\epsilon(x_\mu)\,x_\mu\frac{\partial\Phi^*}{\partial x_\mu}\right) \tag{IV.8}$$

The integral here is over both branches of the hyperboloid in contrast to the massive case. β is as in the first sections of this paper. The Lorentz invariant measure $d\beta$ can be written alternatively $dx_+/|x_+|$ or $dx_-/|x_-|$. $\sum_{\mu=\pm}\epsilon(x_\mu)\,x_\mu(\partial/\partial x_\mu)$ is a more convenient form, for the purposes of this section, of $\partial/\partial q$.

With respect to (IV.8) f_γ and f_γ^* are orthonormal:

$$(f_\gamma(x\cdot n_j),\ f_{\gamma'}(x\cdot n_{j'})) = \delta(\gamma-\gamma')\,\delta_{j,j'},$$

$$(f_\gamma^*(x\cdot n_j),\ f_{\gamma'}^*(x\cdot n_{j'})) = -\delta(\gamma-\gamma')\,\delta_{j,j'}, \tag{IV.9}$$

$$(f_\gamma^*(x\cdot n_j),\ f_{\gamma'}(x\cdot n_{j'})) = (f_\gamma(x\cdot n_j),\ f_{\gamma'}^*(x\cdot n_{j'})) = 0.$$

These relationships are equally valid if the integration (IV.8) is performed on surfaces $x^2<0$. This means that the operators $a_j(\gamma)$ and $a_j^\dagger(\gamma)$ may be recovered from Cauchy data given on *any* hyperboloid $x^2=$ const, $x^2>0$ or $x^2<0$.

Let us review the results of this section in a slightly different order to parallel the logic of Sec. III. We have seen that Cauchy data on both sheets of any hyperboloid are necessary to determine the massless field. Consequently, our initial quantization involves both regions I and IV where $x^2>0$. We can then use knowledge of the quantum commutator and boundary conditions on the propagator [alternatively, just knowledge of the propagator per (IV.6)] to construct an appropriate field expansion in $x^2>0$. We took the usual propagator to guide us, and the result is the field expansion (IV.5). The expansion is evidently a solution of the field equation everywhere in space—time, and so we can proceed immediately to check that the vacuum expectation of the field product, when properly ordered, gives the chosen propagator in $x^2<0$ also. But regardless of region, $\langle 0|\Phi(x)\Phi(x')|0\rangle$ is given by the top right side of (IV.3) and $\langle 0|\Phi(x')\Phi(x)|0\rangle$ is the bottom right side. Also the harmonic decomposition (IV.3) is valid in all regions, so whenever ordering makes a difference, we must use time ordering.

The development picture we gave in the massive case cannot go through with a massless field. The reason is obvious; Φ and $\partial\Phi/\partial q$ given on $q=-\infty$ of the past is not sufficient Cauchy data. We do not know how to modify the picture to include massless fields. It may well be that this type of quantization for massless fields can only be useful in the future (or past) alone when interactions are introduced and an exact solution in terms of Cauchy data is therefore impossible.

V. THE DIRAC FIELD

The spin 1/2 field goes through in the same way that the scalar field did. In particular, continuation outside the forward light cone follows the same considerations and prescriptions.

The Dirac equation reads

$$\begin{gathered}(i\not\nabla-m)\psi=0, \quad \bar{\psi}(-i\overleftarrow{\not\nabla}-m)=0,\\ \gamma^0=\begin{pmatrix}0&1\\1&0\end{pmatrix}, \quad \gamma'=\begin{pmatrix}0&1\\-1&0\end{pmatrix},\\ \bar{\psi}=\psi^\dagger\gamma^0.\end{gathered} \tag{V.1}$$

It will be convenient inside the forward light cone to use again coordinates $\tau=e^\alpha$, β. We introduce also

$$\gamma^\tau=\begin{pmatrix}0&e^{-\beta}\\e^\beta&0\end{pmatrix}, \quad \gamma^\beta=\begin{pmatrix}0&e^{-\beta}\\-e^\beta&0\end{pmatrix}$$

in order to rewrite the Dirac equation (again scaling coordinates by m):

$$\left(i\gamma^\tau\frac{\partial}{\partial\tau}+\frac{i\gamma^\beta}{\tau}\frac{\partial}{\partial\beta}-1\right)\psi=0. \tag{V.2}$$

We can note the natural bilinear form associated with the Dirac equation in this form:

$$(\bar{\psi},\psi)=-\tau\int d\beta\,\bar{\psi}\gamma^\tau\psi. \tag{V.3}$$

As with the scalar case the integral is over a single sheet of any hyperboloid $x^2>0$.

Performing a harmonic decomposition of the field, one has

$$\begin{gathered}\psi_1=\int d\rho\exp[i\beta(\rho+i/2)]P_\rho(\tau),\\ \psi_2=\int d\rho\exp[i\beta(\rho-i/2)]Q_\rho(\tau),\\ i\left(\frac{d}{d\tau}-\frac{i}{\tau}(\rho+i/2)\right)P_\rho=Q_\rho,\\ i\left(\frac{d}{d\tau}+\frac{i}{\tau}(\rho-i/2)\right)Q_\rho=P_\rho.\end{gathered} \tag{V.4}$$

Solutions of these coupled equations are

$$P_\rho=\exp[\pi/2(\rho+i/2)]H_{i(\rho+i/2)}(\tau),$$
$$Q_\rho=\exp[\pi/2(\rho-i/2)]H_{i(\rho-i/2)}(\tau),$$

or

$$P_\rho=\exp[-\pi/2(\rho-i/2)]H_{i(\rho+i/2)}(\tau),$$
$$Q_\rho=\exp[-\pi/2(\rho+i/2)]H_{i(\rho-i/2)}(\tau).$$

The H's are Hankel functions of first or second kinds.

As in the scalar case we demand that the propagator, developmentally ordered vacuum expectation of the field product, must match the usual boundary conditions. Equivalently, we match positive frequency solutions with destruction and negative frequency solutions with creation operators:

$$\begin{gathered}\langle 0|D\psi(x)\bar{\psi}(x')|0\rangle=\Delta_s(x,x'),\\ \Delta_s(x,x')=(i\not\nabla_x+1)\,\Delta(x,x').\end{gathered} \tag{V.5}$$

The expansion which satisfies these requirements is given below:

$$\psi_j=\int_{-\infty}^{\infty}d\rho\,X_{j,\rho}\,a(\rho)+Y_{j,\rho}\,b^\dagger(\rho).$$

$$\begin{gathered}X_{1,\rho}=\frac{1}{\sqrt{8}}\exp[i\beta(\rho+i/2)]\exp[\pi/2(\rho+i/2)]H^{(2)}_{i(\rho+i/2)}(\tau),\\ X_{2,\rho}=\frac{1}{\sqrt{8}}\exp[i\beta(\rho-i/2)]\exp[\pi/2(\rho-i/2)]H^{(2)}_{i(\rho-j/2)}(\tau),\\ Y_{1,\rho}=\frac{1}{\sqrt{8}}\exp[i\beta(\rho+i/2)]\exp[-\pi/2(\rho-i/2)]H^{(1)}_{i(\rho+i/2)}(\tau),\\ Y_{2,\rho}=\frac{1}{\sqrt{8}}\exp[i\beta(\rho-i/2)]\exp[-\pi/2(\rho+i/2)]H^{(1)}_{i(\rho-i/2)}(\tau).\end{gathered} \tag{V.6}$$

The a's and b's are quantized with the usual anticommutation relations. The x's and y's satisfy orthonormality conditions with respect to the bilinear form (V.3):

$$\begin{gathered}(\bar{X}_\rho,X_{\rho'})=\delta(\rho-\rho')=(\bar{Y}_\rho,Y_{\rho'}),\\ (\bar{Y}_\rho,X_{\rho'})=0=(\bar{X}_\rho,Y_{\rho'}).\end{gathered} \tag{V.7}$$

Since the functions X and Y all satisfy the Klein–Gordon equation, we are in a position to continue the entire field into all regions of space–time by the prescription given in Table I, Sec. III.

Though in the interest of brevity we will not explicitly carry out this continuation, one finds as in the scalar case that in order to insure a translation invariant propagator, the ordering must be time ordering or an equivalent one. Not unexpectedly, therefore, the development picture of Sec. III is also appropriate for the massive Dirac field.

For completeness we give the full space–time field expansion for the massless Dirac field. The massless equation can be written simply

$$\not\nabla\psi=\sqrt{2}\begin{pmatrix}0&\dfrac{\partial}{\partial x_+}\\ \dfrac{\partial}{\partial x_-}&0\end{pmatrix}\psi=0. \tag{V.8}$$

The solution is

$$\psi_1=\psi_+(x_+), \qquad \psi_2=\psi_-(x_-).$$

The propagator is again derived from that of the scalar field:

$$\Delta_s(x,x')=-i\not\nabla_x\Delta(x,x'). \tag{V.9}$$

There is no logarithmic difficulty in the spinor case, and no discrete point in the Lorentz decomposition spectrum is needed. The expansion of the field which gives the above propagator as a time ordered vacuum expectation follows. The same considerations apply here as in Sec. IV where time ordering is also used:

$$\psi_j = \int_{-\infty}^{\infty} d\rho\, X_\rho(x\cdot n_j)\, a_j(\rho) + Y_\rho(x\cdot n_j)\, b_j^\dagger(\rho), \quad j=\pm,$$

$$X_\rho(x\cdot n_j) = \frac{1}{\sqrt{2\pi}}\,\frac{1}{[1+\exp(2\pi\rho)]^{1/2}}\,(x\cdot n_j + i\epsilon)^{-i\rho-1/2}, \qquad \text{(V.10)}$$

$$Y_\rho(x\cdot n_j) = \frac{1}{\sqrt{2\pi}}\,\frac{1}{[1+\exp(-2\pi\rho)]^{1/2}}\,(x\cdot n_j - i\epsilon)^{-i\rho-1/2}.$$

VI. FOUR-DIMENSIONAL FIELD

We shall limit our discussion of the four-dimensional case to a presentation of the field expansions for the massive and massless scalar Hermitian fields which satisfy the criteria developed in preceeding sections. To wit, the expansions will be in terms of a complete set of orthonormalized solutions with positive and negative frequency associated with destruction and creation operators, respectively. Explicit verification of the commutator and propagator relations is quite difficult, but we can rest on two less direct arguments. First, given a choice of Lorentz harmonics (ours are essentially those of Sommerfield, unlike those used by GRS), the expansions following the above criteria are unique up to phases. Second, one can check routinely by manipulating the resulting integral expressions that the propagator and commutator satisfy the appropriate differential equations and that the commutator satisfies the Cauchy data, i.e., canonical commutators, of the usual field commutator.

First of all we wish to perform a Lorentz harmonic analysis of the field. Recently the problem of determining Lorentz harmonics has been solved in the general contest of $O(n, N)$ harmonic analysis on hyperboloids by Strichhartz,[12] though the special case of interest, four-dimensional Lorentz space, was well known in the literature. The results we describe below follow straightforwardly from Strichhartz' work.

We begin in the forward sheet of the hyperboloid $x^2 = t^2 - x_1^2 - x_2^2 - x_3^2 = 1$. The space of square integrable functions on that base space with Lorentz invariant measure,[13] $d\underline{h} = d^3\underline{x}/(1-\mathbf{x}^2)^{1/2}$ can be decomposed into a direct integral of irreducible representations of the Lorentz group. The representations can be labeled by a real parameter $\rho > 0$. Function elements within a given representation can be parameterized by a point $\mathbf{n}$ in the two-dimensional unit sphere $n_1^2 + n_2^2 + n_3^2 = 1$. The decomposition can be carried out with the following functions:

$$S_{\rho,n}(\underline{x}) = \frac{1}{\sqrt{\pi^3}}\,\rho\,(n\cdot\underline{x})^{i\rho-1}. \qquad \text{(VI.1)}$$

As a 4-vector n denotes the null vector $(1, \mathbf{n})$. $S_{\rho,n}$ satisfy the Casimir eigenvalue equation

$$-\Box_{\underline{h}} S_{\rho,n} = (1+\rho^2)\, S_{\rho,n}. \qquad \text{(VI.2)}$$

$\Box_{\underline{h}}$ is the negative of the wave operator,

$$\Box = \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2} - \frac{\partial^2}{\partial x_3^2}$$

restricted to the unit hyperboloid. $S_{\rho,n}$ also satisfy the following completeness and orthogonality relations:

$$\int d\rho\, d^2\Omega_n\, S_{\rho,n}(\underline{x})\, S^*_{\rho,n}(x') = \delta^3(\underline{x} - \underline{x}'),$$

$$\int d\underline{h}\, S_{\rho,n}(\underline{x})\, S^*_{\rho',n'}(\underline{x}) = \delta(\rho-\rho')\,\delta^2(\mathbf{n}-n'). \qquad \text{(VI.3)}$$

The measure $d\Omega n$ is the Lebegue measure $[d(\cos\theta)\,d\phi]$ on the two-sphere, and $\delta^2(\mathbf{n}-n')$ is the corresponding Dirac distribution $\delta(\phi-\phi')\,\delta(\cos\theta - \cos\theta')$. $\delta^3(\underline{x}-\underline{x}')$ is the Dirac distribution on the hyperboloid, $\int d\underline{h}\,\delta^3(\underline{x}-\underline{x}')\, f(\underline{x}) = f(\underline{x}')$.

One additional fact about $S_{\rho,n}$ is needed in this context. $S^*_{\rho,n}$ and $S_{\rho,n'}$, ρ fixed, span the same space of functions. Thus $\int d\underline{h}\, S_{\rho,n}\, S_{\rho',n'} = 0$ unless $\rho = \rho'$. This will be useful in the proof of orthonormality relations for the field expansion functions.

Again, we will quantize the massive field first in the forward light cone. Performing a Lorentz harmonic analysis

$$\Phi(x) = \int d\rho\, d^2\Omega_n\, A_{\rho,n}(\tau)\, S_{\rho,n}(\underline{x}).$$

We are using scaled coordinates with homogeneous coordinates $\underline{x}_\mu = x_\mu/\tau$ for the "kinematic" dependence. The field equation $(\Box + 1)\Phi = 0$ is now

$$\left(\frac{\partial^2}{\partial\tau^2} + \frac{3}{\tau}\frac{\partial}{\partial\tau} - \frac{1}{\tau^2}\Box_{\underline{h}} + 1\right)\Phi = 0. \qquad \text{(VI.4)}$$

This results in an equation for $A_{\rho,n}(\tau)$:

$$\left(\frac{d^2}{d\tau^2} + \frac{3}{\tau}\frac{d}{d\tau} + \frac{1+\rho^2}{\tau^2} + 1\right) A_{\rho,n}(\tau) = 0.$$

We take independent solutions proportional to $\tau^{-1} H^{(2)}_{i\rho}(\tau)$ and $\tau^{-1} H^{(1)}_{i\rho}(\tau)$.

Again we introduce a natural Hermitian structure into the solution space of $(\Box + 1)\Phi = 0$:

$$(\chi, \Phi) = i\int dh\left(\chi^*\frac{\partial}{\partial\tau}\Phi - \Phi\frac{\partial}{\partial\tau}\chi^*\right). \qquad \text{(VI.5)}$$

The measure dh is the invariant $\tau^3 d\underline{h}$ and the integral may be performed on the forward sheet of any hyperboloid.

With respect to (VI.5) the following is a complete orthonormal set of solutions:

$$V_{\rho,n} = \frac{\rho}{2\pi}\,\frac{\exp(\pi\rho/2)}{\tau}\, H^{(2)}_{i\rho}(\tau)\,(n\cdot\underline{x})^{i\rho-1},$$

$$V^*_{\rho,n} = \frac{\rho}{2\pi}\,\frac{\exp(-\pi\rho/2)}{\tau}\, H^{(1)}_{i\rho}(\tau)\,(n\cdot\underline{x})^{-i\rho-1}; \qquad \text{(VI.6a)}$$

$$(V_{\rho,n}, V_{\rho',n'}) = \delta(\rho-\rho')\,\delta^2(\mathbf{n}-\mathbf{n}'),$$

$$(V^*_{\rho,n}, V^*_{\rho',n'}) = -\delta(\rho-\rho')\,\delta^2(\mathbf{n}-\mathbf{n}'), \qquad \text{(VI.6b)}$$

$$(V^*_{\rho,n}, V_{\rho',n'}) = 0 = (V_{\rho,n}, V^*_{\rho',n'}).$$

In the same way that we did in two dimensions it is simple to show that $V_{\rho,n}$ and $V^*_{\rho,n}$ are positive and negative frequency solutions, respectively. Thus we are led to the four-dimensional field expansion[14]

$$\Phi = \int_0^\infty d\rho \int d^2\Omega_n (V^*_{\rho,n} a^\dagger_{\rho,n} + V_{\rho,n} a_{\rho,n}). \qquad \text{(VI.7)}$$

Just as in two dimensions analytic continuation of V and V^* makes this a full space field expansion.

The zero mass field is slightly more complicated. To begin we enquire how the harmonics $S_{\rho,n}(\underline{x})$ can be extended to solutions of the wave equation

$$\Box A_{\rho,n}(\tau)\, S_{\rho,n}(x) = \left(\frac{\partial^2}{\partial \tau^2} + \frac{3}{\tau}\frac{\partial}{\partial \tau} - \frac{1}{\tau^2}\Box_{\underline{h}}\right) A_{\rho,n}\, S_{\rho,n}$$
$$= \left(\frac{\partial^2}{\partial \tau^2} + \frac{3}{\tau}\frac{\partial}{\partial \tau} + \frac{(1+\rho^2)}{\tau^2}\right) A_{\rho,n}\, S_{\rho,n} = 0.$$

Independent solutions for $A_{\rho,n}$ are $\tau^{\pm i\rho-1}$. We can note that $\tau^{i\rho-1}(n\cdot\underline{x})^{i\rho-1} = (n\cdot x)^{i\rho-1}$. Also since $S_{\rho,n}(\underline{x})$ and $S^*_{\rho,n'}(\underline{x})$ span the same space and since $\tau^{-i\rho-1}S^*_{\rho,n'}(\underline{x}) \propto \tau^{-i\rho-1}(n'\cdot\underline{x})^{-i\rho-1} = (n'\cdot x)^{-i\rho-1}$, we will find it notationally convenient to use $\tau^{-i\rho-1}S^*_{\rho,n}$ instead of $\tau^{-i\rho-1}S_{\rho,n}$ as a mate to $\tau^{i\rho-1}S_{\rho,n}$. We can then use $(x\cdot n)^{i\rho-1}$, $-\infty<\rho<\infty$, as a complete set of solutions.

Just as in two dimensions, the massless field requires Cauchy data on both sheets of the hyperboloid $x^2=\text{const}$. To get to the lower sheet we again use positive and negative frequency extensions of $(x\cdot n)^{i\rho-1}$, namely $(x\cdot n - i\epsilon)^{i\rho-1}$ and $(x\cdot n + i\epsilon)^{i\rho-1}$. The appropriate Hermitian form is also an integral over *both* sheets of the hyperboloid:

$$(\chi,\phi) = i\int dh\left(\chi^*\frac{\partial}{\partial\tau}\Phi - \Phi\frac{\partial}{\partial\tau}\chi^*\right)\epsilon(t). \qquad \text{(VI. 8)}$$

To sum up we list a complete orthonormal set of solutions to $\Box\Phi=0$ in all space—time:

$$V_{\rho,n}(x) = \left(\frac{\rho\exp(-\pi\rho)}{2\pi^3\sinh\pi\rho}\right)^{1/2}(x\cdot n - i\epsilon)^{+i\rho-1};$$
$$V^*_{\rho,n}(x) = \left(\frac{\rho\exp(-\pi\rho)}{2\pi^3\sinh\pi\rho}\right)^{1/2}(x\cdot n + i\epsilon)^{-i\rho-1}, \qquad \text{(VI. 9a)}$$

$$(V_{\rho,n}, V_{\rho',n'}) = \delta(\rho-\rho')\,\delta^2(\mathbf{n}-\mathbf{n}'),$$
$$(V^*_{\rho,n}, V^*_{\rho',n'}) = -\delta(\rho-\rho')\,\delta^2(\mathbf{n}-\mathbf{n}'), \qquad \text{(VI. 9b)}$$
$$(V^*_{\rho,n}, V_{\rho',n'}) = 0 = (V_{\rho,n}, V^*_{\rho',n'}).$$

We remind the reader that here $-\infty<\rho<\infty$. Verification of the orthonormality properties of the V's is a tedious but straightforward task using the orthonormality properties of the Lorentz harmonics $S_{\rho,n}(\underline{x})$.

The field expansion in four dimensions below and the expansion functions (VI. 9) bear a very close resemblance to their two-dimensional counterparts (IV. 5):

$$\Phi = \int_{-\infty}^{\infty} d\rho \int d^2\Omega_n (V^*_{\rho,n}\, a^\dagger_{\rho,n} + V_{\rho,n}\, a_{\rho,n}) \qquad \text{(VI. 10)}$$

Indeed our notation conventions were chosen to emphasize the similarity. In particular, the parameterization of Lorentz harmonics of a given irreducible representation by points on an $n-2$ sphere ($n=$ dimension space—time) accounts for the $\sum_{j=\pm}$ which represents the two points, ±1, of the zero-dimensional sphere.

The reader may wonder about the $\epsilon(x_\mu)$ or $\epsilon(t)$ factors which occur in the definitions of Hermitian forms for the massless case. They are necessary to match the $\partial/\partial q$ direction of field development we discussed in the massive case, and in the final analysis are the result of the choice of a positive direction for time. Positive and negative frequency solutions are naturally orthogonal with the chosen Hermitian form. On the other hand, if in the field expansion we used odd and even solutions $|n\cdot x|^{i\rho}$, $|n\cdot x|^{i\rho}\epsilon(n\cdot x)$ which are naturally orthogonal with the Hermitian form *excluding* $\epsilon(x_\mu)$ or $\epsilon(t)$, then the propagator would respect the time-reversing symmetry

$\Delta(x,x') = \Delta(-x,-x')$. This last is, of course, not true for the usual positive time-directed propagator.

VII. CONCLUSION

In investigating the problem of quantization on hyperboloids we have been led to full space field expansions which have the same propagator and commutator relations as the usual approach to quantization. In regions $x^2<0$ we found, to no great surprise, that the timelike hyperboloids are unsuitable for a surface-to-surface development picture. Nonetheless, one can salvage a Hamiltonian deployment of the S matrix from far past to far future by interpolating between backward and forward light cone with a set of cones having apex at the origin. We carried out the procedure explicitly in the two-dimensional massive case.

Unfortunately, massless fields are not amenable to this "patching up." That does not preclude application of the quantization on surfaces $x^\mu x_\mu=\text{const}$, however. First of all one has at the least achieved a reduction of the field by the homogeneous Lorentz group instead of the usual reduction by the translation group (Fourier analysis). Explicit presentation of invariance properties with respect to the homogeneous Lorentz group is just one of the advantages of this decomposition.

Secondly, if necessary, one can retreat to the forward light cone where problems with the massless case do not arise. The retreat can be made in two ways. One can content oneself with whatever information is contained in the forward light cone or, alternatively, simply place the zero of the coordinate system in the past of any relevant events, which is the suggestion of Sommerfield.

Finally, in Euclidean space surfaces $x_\mu x^\mu=\text{const}$ are connected spheres and the normal (radial) development spans all space. Massless fields there can cause no difficulties and perturbation theory is in principle straightforward.[15]

ACKNOWLEDGMENTS

I wish to express my thanks to Roman Jackiw who provided me with the framework for this study and also gave much assistance during its progress. I wish also to acknowledge conversations with C. Sommerfield.

*This work is supported in part through funds provided by the Atomic Energy Commission under Contract AT(11-1)-3069.

[1] S. Fubini, A.J. Hanson, and R. Jackiw, Phys. Rev. D 7, 1732 (1973).

[2] C.M. Sommerfield, Ann. Phys. (N.Y.) 84, 285 (1974).

[3] D. Gromes, H. Rothe, and B. Stech, Nucl. Phys. B 75, 313 (1974).

[4] A. diSessa, Phys. Rev. D 9 (1974).

[5] A thorough discussion of this will be found in Ref. 1.

[6] This can easily be shown to be the most general Lorentz invariant real solution of $(\Box_x + m^2)D(x,x') = 0$ and $(\Box_{x'} + m^2)\,D(x,x') = 0$.

[7] For evaluation of this integral see Eq. (III.2) and surrounding discussion.

[8] This is essentially the technique used by Sommerfield in Ref. 2 to examine the field outside the forward light cone.

[9] Properties of Bessel functions used in this paper, in particular, analytic continuations, will be found in G.N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge U. P., Cambridge, 1966), 2nd ed.

[10] A. Erdelyi *et al.*, *Tables of Integral Transforms*, Bateman Manuscript Project (McGraw-Hill, New York, 1954), Vol. II, p. 382.

[11] Reference 10, p. 173.

[12] R.S. Strichartz, J. Funct. Anal., **12** (4), 341 (1973).

[13] Notational note: We will often emphasize the restriction of a point to the unit hyperboloid with an underbar, e.g. $\underline{X}$, $\underline{dh}$.

[14] As always the creation and destruction operators are understood to satisfy commutation relations normalized to a product of δ functions of the labels;

$$[a_{\rho,n}, a^{\dagger}_{\rho',n'}] = \delta(\rho-\rho')\,\delta^2(\mathbf{n}-\mathbf{n}')$$

[15] A discussion of this point occurs in Ref. 1.

Propagation through an anisotropic random medium

M. J. Beran*

University of Pennsylvania, Philadelphia, Pennsylvania 19104

J. J. McCoy*

The Catholic University of America, Washington, D.C. 20017

(Received 16 January 1974)

An expression is derived, Eq. (59), for the mutual coherence of an initial plane wave signal that has propagated a distance, z, into an anisotropic random medium. This expression is valid for cases in which the characteristic radiation wavelength divided by 2π. $\bar{\lambda}/2\pi$ is roughly speaking of the same order as, or greater than, all characteristic correlation lengths in the direction perpendicular to the mean propagation direction, which is taken to lie in a horizontal plane. The exact condition is given in Eq. (48). We require $\bar{\lambda}/2\pi$ to be much smaller than all characteristic correlation lengths in the horizontal plane. Two derivation procedures are used. One follows that introduced by Beran (J. Opt. Soc. Am. **56**, 1475 (1966)] for cases in which $\bar{\lambda}/2\pi$ is much smaller than all characteristic correlation lengths and one is based on introducing simplifications in a Bethe-Salpeter equation. We discuss the problem of the loss of spatial coherence of an acoustic signal due to scattering by the ocean temperature microstructure in the light of the theory presented.

1. INTRODUCTION

Most work on the propagation of a radiation field through a random medium has been carried out under the assumption that $\bar{k}l_m \gg 1$, where $\bar{k}$ is the radiation wavenumber and l_m is the minimum correlation length associated with the random medium. In this case solutions are available for the mutual coherence function (which is directly related to the resolution limitation resulting from the presence of a random medium) for a wide variety of situations. In the case of propagation of a plane wave incident on the random medium, for example, the solution for the mutual coherence function, $\{\hat{\Gamma}\}$, defined by Eq. (29), is[1]

$$\{\hat{\Gamma}(x_{12}, y_{12}, z, \nu)\} = \hat{I}(\nu) \exp(\bar{k}^2 z[\bar{\sigma}(x_{12}, y_{12}) - \bar{\sigma}(0,0)]). \quad (1)$$

The coherence function is written here for two points in a plane a distance z into the medium. The coordinates x_{12} and y_{12} are difference coordinates defined by the points in the z plane. Further,

$$\bar{\sigma}(x_{12}, y_{12}) = \tfrac{1}{4}\int_{-\infty}^{\infty} \sigma(x_{12}, y_{12}, s_z)\, ds_z, \quad (2)$$

where $\sigma(x_{12}, y_{12}, s_z)$ is the correlation function associated with the inhomogeneities in index of refraction. It is not required that $\sigma(x_{12}, y_{12}, s_z)$ be statistically isotropic but the medium is taken to be statistically homogeneous. The precise conditions for the validity of the solution are

$$\bar{k}l_M\theta^2 \ll 1, \quad (3a)$$

$$\bar{k}^2\bar{\sigma}(0)\Delta z \ll 1, \quad (3b)$$

$$\bar{k}\theta^4\Delta z \ll 1, \quad (3c)$$

where Δz is some distance for which $\Delta z \gg l_M$, l_M being the largest important characteristic correlation length associated with σ. The angle θ represents the largest characteristic angular spread of the radiation. Condition (3a) is a somewhat stronger form of the condition $\bar{k}l_m \gg 1$. If $\theta = O(1/\bar{k}l_m)$ and $l_M = O(l_m)$, the conditions are the same. Condition (3b) essentially requires that the fluctuations in index of refraction be very weak.

One source of scattering of acoustic signals in the ocean is the temperature microstructure and it is this source that motivates the present study. Since the fluctuations in the index of refraction that are associated with the temperature microstructure are very weak, the above described theory is applicable for high frequency acoustic signals, where $\bar{k}l_m \gg 1$. The temperature microstructure in the ocean is very anisotropic, with correlation lengths defined by measurements taken in a horizontal direction being orders of magnitude greater than corresponding lengths defined by measurements taken in the depth direction. The theory presented in this paper is applicable to a highly anisotropic fluctuation field, where in place of the requirement that $\bar{k}l_m \gg 1$ we require that $\bar{k}l_{Hm} \gg 1$ and $\bar{k}l_{VM} \ll (\bar{k}l_{Hm})^{1/2}(l_{Hm}/l_{HM})^{1/2}$. Here, l_{Hm} is the minimum correlation length for measurements in a horizontal direction and l_{HM} and l_{VM} are maximum correlation lengths for measurements in a horizontal direction and the vertical direction, respectively. (We note that this second inequality is, in general, a good deal weaker than the inequality that $\bar{k}l_{VM} \ll 1$). We consider the mean propagation to be in the horizontal direction and, for convenience, we assume that the statistics of the refractive index field are isotropic in the horizontal plane.

For an isotropic medium, the single scattering solution shows that the characteristic angular spread of the scattered radiation is $O(1/\bar{k}l_m)$ when $\bar{k}l_m \gg 1$. This result is crucial in making suitable approximations necessary to obtain the solution given in Eq. (1). On the other hand, if $\bar{k}l_M \ll 1$ and the random field is isotropic then the angular scattering is isotropic. We might expect on this basis that if $kl_{VM} \ll 1$ and $kl_{Hm} \gg 1$, then scattering in the horizontal plane would have an angular spread of order $1/\bar{k}l_{Hm}$, whereas in the vertical plane it would be isotropic. This is not the case, nowever, and direct calculation shows that although $\theta_H = O(1/\bar{k}l_{Hm})$ we obtain $\theta_V = O(1/(\bar{k}l_{Hm})^{1/2})$.

In the next section we present a single scattering solution that yields the above results. These results are then used to motivate the approximations required to derive a multiple scatter solution analogous to that given in Eq. (1). The procedure is identical to that used by Beran.[1] The resulting expression is given by Eq. (59) and the conditions to be satisfied for this expression to be valid are summarized in Eqs. (63) and (64).

In an appendix an alternate derivation of Eq. (59) is presented, which is based on introducing simplifications to a Bethe—Salpeter type equation that may be written for the coherence function. In Sec. 4, we consider index of refraction fluctuations with a spectrum that may be described by a simple power law. Albegraic expressions are obtained for the horizontal coherence length; i.e., horizontal separation distance for the coherence function to decay to $1/e$ of its zero separation value. The coherence length is related to the maximum useful array length for the coherent addition of received signals.

2. SINGLE SCATTER SOLUTION

A simple way to determine the angular spectrum resulting from the scattering of a plane wave by a random medium is to consider the geometry given in Fig. 1.

A time harmonic plane wave impinges on a finite scattering volume and we calculate the intensity of the scattered radiation at a very distant point (R,θ), where $R \gg \bar{k}D^2$. For an isotropic random medium we find the well-known result that $\theta = O(1/\bar{k}l_m)$ if $\bar{k}l_m \gg 1$, i.e., $\{\hat{I}(R,\theta)\} \approx 0$ for $\theta \gg 1/\bar{k}l_m$. On the other hand $\{\hat{I}(R,\theta)\}$ is independent of θ if $\bar{k}l_m \ll 1$. In the present study we are interested in an anisotropic medium with different characteristic correlation lengths associated with differing directions. We thus have three pairs of lengths (l_{xm}, l_{xM}), (l_{ym}, l_{yM}) and (l_{zm}, l_{zM}). We shall study combinations of conditions to show the nature of the problem, but in this paper we shall be particularly interested in the case $\bar{k}l_{yM} \ll 1$, $\bar{k}l_{xm} \gg 1$, $\bar{k}l_{zm} \gg 1$. Here we shall find that $\theta_x = O(1/\bar{k}l_{zm})$ and $\theta_y = O(1/\bar{k}l_{zm})^{1/2})$.

The equation governing the propagation of a pressure signal in water with a variable index of refraction is taken here to be

$$\nabla^2 p = \frac{1}{C^2(\mathbf{x})} \frac{\partial^2 p}{\partial t^2}, \tag{4}$$

where $p(\mathbf{x},t)$ is the pressure field and $C^2(\mathbf{x})$ is the variable speed of sound. There are a number of approximations necessary to obtain Eq. (4) (see Chernov[2]). In particular we choose here $C^2(\mathbf{x})$ rather than $C^2(\mathbf{x},t)$ since we shall use ensemble averaging and can choose the properties of the water to be fixed in each realization of the ensemble. It is convenient to rewrite Eq. (4) as

$$\nabla^2 p = \frac{\{n^2\}}{\{C\}^2} [1 + \epsilon\mu(\mathbf{x})] \frac{\partial^2 p}{\partial t^2}, \tag{5}$$

where the braces $\{\ \}$ indicate an ensemble average. The term $\{C\}$ is the mean sound speed and $\mu(\mathbf{x})$ denotes a centered stochastic process of unit variance defined by the randomly varying index of refraction field. The term $\{n^2\}$ is defined by the equality

$$\{n^2\} = \{C\}^2/\{C^2\} \tag{6}$$

and ϵ is a measure of the strength of the index of refraction fluctuation field. In all of our studies we shall assume that $\epsilon \ll 1$. To first order in ϵ, therefore, $\{n^2\} = 1$.

For narrow band signals, with central frequency $\bar{\nu}$, it is convenient to introduce the approximation

$$p(\mathbf{x},t) = \mathrm{Re}[\hat{p}(\mathbf{x},\nu)\exp(2\pi i\bar{\nu}t)], \tag{7}$$

where $\hat{p}(\mathbf{x},\nu)$ is the complex pressure field. Substitution into Eq. (5) yields

$$\nabla^2\hat{p} + \bar{k}^2[1 + \epsilon\mu(\mathbf{x})]\hat{p} = 0, \tag{8}$$

where

$$\bar{k} = 2\pi\bar{\nu}/\{C\}. \tag{9}$$

The single scatter solutions are obtained by writting the solution in the form

$$\hat{p}(\mathbf{x}) = \hat{p}_0(\mathbf{x}) + \epsilon\hat{p}_1(\mathbf{x}), \tag{10}$$

where $\hat{p}_0(\mathbf{x})$ and $\hat{p}_1(\mathbf{x})$ are independent of ϵ, and by dropping all terms of order ϵ^2 and higher. For an initial plane wave incident on the scattering volume we have

$$\hat{p}_0(\mathbf{x}) = (\hat{I})^{1/2}\exp(i\bar{k}z), \tag{11}$$

where $\hat{p}_1(\mathbf{x})$ satisfies the equation

$$\begin{aligned}\nabla^2\hat{p}_1 + \bar{k}^2\hat{p}_1 &= -\bar{k}^2\mu(\mathbf{x})\hat{p}_0(\mathbf{x}) \\ &= -\bar{k}^2\mu(\mathbf{x})(\hat{I})^{1/2}\exp(i\bar{k}z).\end{aligned} \tag{12}$$

In Eq. (11), $\hat{I}$ might be termed the intensity of the initial plane wave although this definition differs slightly from the usual definition used in acoustics studies. The solution of Eq. (12) is to satisfy the radiation condition far from scattering volume and we write the result as

$$\hat{p}_1(\mathbf{x}) = -\frac{\bar{k}^2(\hat{I})^{1/2}}{4\pi} \int \frac{\exp\{i\bar{k}[r(\mathbf{x},\mathbf{x}') + z']\}}{r(\mathbf{x},\mathbf{x}')} \mu(\mathbf{x}')\,d\mathbf{x}', \tag{13}$$

where $r(\mathbf{x},\mathbf{x}')$ is the distance between $\mathbf{x}$ and $\mathbf{x}'$. The integration is over the scattering volume.

Defining the intensity of the scattered radiation field by

$$\{\hat{I}_S(\mathbf{x})\} = \epsilon^2\{\hat{p}_1(\mathbf{x})\hat{p}_1^*(\mathbf{x})\}, \tag{14}$$

yields the following expression for $\{\hat{I}_S\}$:

$$\begin{aligned}\{\hat{I}_S(\mathbf{x})\} = \frac{\bar{k}^4\hat{I}}{(4\pi)^2} \int\!\!\int \frac{\exp\{i\bar{k}[r(\mathbf{x},\mathbf{x}') - r(\mathbf{x},\mathbf{x}'') + (z' - z'')]\}}{r(\mathbf{x},\mathbf{x}')r(\mathbf{x},\mathbf{x}'')} \\ \times\sigma(\mathbf{x}',\mathbf{x}'')\,d\mathbf{x}'\,d\mathbf{x}'',\end{aligned} \tag{15}$$

where

$$\sigma(\mathbf{x}',\mathbf{x}'') = \epsilon^2\{\mu(\mathbf{x}')\mu(\mathbf{x}'')\} \tag{16}$$

is the spatial correlation function defined on the index of refraction fluctuations.

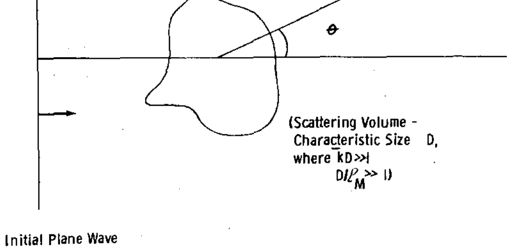


FIG. 1.

Introducing homogeneous statistics and the far field approximation enables us to accomplish a partial integration of the right-hand side of Eq. (15)—see, for example, Beran and Parrent.[3] We write the result

$$\{\hat{I}_S(\mathbf{x})\} = \frac{\bar{k}^4 \hat{I}}{(4\pi)^2 R^2} \int \sigma(\mathbf{u}) \exp\left[-i\bar{k}\left(\frac{\mathbf{u}\cdot\mathbf{x}}{R} - u_z\right)\right] d\mathbf{u}, \quad (17)$$

where $\mathbf{u} = \mathbf{x}' - \mathbf{x}''$ is the difference coordinate. We shall make use of Eq. (17) to demonstrate the dependence of the angular variation of the scattered intensity on correlation lengths defined by the index of refraction fluctuation field.

To this end we introduce a correlation function of the following form:

$$\sigma(\mathbf{u}) = \sigma_0 \exp\left[-\left(\frac{u_x^2}{l_x^2} + \frac{u_y^2}{l_y^2} + \frac{u_z^2}{l_z^2}\right)\right]. \quad (18)$$

[We emphasize that Eq. (18) is not meant to represent a realistic description of the index of refraction fluctuation field that is caused by the ocean temperature microstructure but is only introduced to provide the insight we require.]

Substituting Eq. (18) into Eq. (17) and integrating yields

$$\{\hat{I}_S(\mathbf{x})\} = \frac{\sigma_0 \bar{k}^4 \hat{I} V}{(4\pi)^2 R^2} \pi^{3/2} l_x l_y l_z \times \exp\left\{-\frac{\bar{k}^4}{4}\left[\frac{l_x^2 x^2}{R^2} + \frac{l_y^2 y^2}{R^2} + l_z^2\left(1 - \frac{z}{R}\right)^2\right]\right\}. \quad (19)$$

In spherical coordinates, i.e.,

$$x = R\sin\theta\cos\varphi,$$
$$y = R\sin\theta\sin\varphi, \quad (20)$$
$$z = R\cos\theta,$$

$$\{\hat{I}_S(\mathbf{x})\} = \frac{\sigma_0 \bar{k}^4 \hat{I} V}{(4\pi)^2 R^2} \pi^{3/2} l_x l_y l_z \times \exp\{-\tfrac{1}{4}\bar{k}^2[(l_x^2 \cos^2\varphi + l_y^2 \sin^2\varphi)\sin^2\theta + l_z^2(1-\cos\theta)^2]\}.$$

Consider now a few possibilities:

(A) $l_x = l_y = l_z = l$, where $\bar{k}l \gg 1$.

This is the usual case discussed in atmospheric propagation studies. Here

$$\{\hat{I}_S^A(\mathbf{x})\} = \frac{\sigma_0 \bar{k}^4 \hat{I} V}{(4\pi)^2 R^2} \pi^{3/2} l^3 \exp[-\tfrac{1}{2}\bar{k}^2 l^2 (1-\cos\theta)] \quad (21)$$

Thus, $\{\hat{I}_S^A(\mathbf{x})\}$ is only appreciable if

$$1 - \cos\theta \ll 1.$$

Therefore

$$1 - \cos\theta \approx \theta^2/2,$$

and

$$\theta = O(1/\bar{k}l). \quad (22)$$

(B) $l_x = l_y = l_z = l$, where $kl \ll 1$.

In this case all correlation lengths are small compared to the radiation wavelength. Here,

$$\{\hat{I}_S^B(\mathbf{x})\} = \frac{\sigma_0 \bar{k}^4 \hat{I} V}{(4\pi)^2 R^2} \pi^{3/2} l^3, \quad (23)$$

and the scattering is isotropic.

(C) $l_x = l_z$, $\bar{k}l_x \gg 1$, $\bar{k}l_y \ll 1$.

This is the case of interest in this paper. The vertical direction is the y direction. Here,

$$\{\hat{I}_S^C(\mathbf{x})\} = \frac{\sigma_0 \bar{k}^4 \hat{I} V}{(4\pi)^2 R^2} \pi^{3/2} l_x^2 l_y \exp\{-\tfrac{1}{4}\bar{k}^2 l_x^2 [\cos^2\varphi \sin^2\theta + (1-\cos\theta)^2]\}. \quad (24)$$

Scattering in the vertical direction is investigated by choosing $x = R\sin\theta\cos\varphi = 0$. We find

$$\{\hat{I}_S^C(\mathbf{x}_V)\} = \frac{\sigma_0 \bar{k}^4 \hat{I} V}{(4\pi)^2 R^2} \pi^{3/2} l_x^2 l_y \exp\{-\tfrac{1}{4}\bar{k}^2 l_x^2 (1-\cos\theta)^2]. \quad (25)$$

For the exponential term to be appreciable we must have

$$(1-\cos\theta)^2 \ll 1.$$

Therefore

$$\frac{\bar{k}^2 l_x^2}{4}\,\frac{\theta_V^4}{4} = O(1),$$

and

$$\theta_V = O(1/(\bar{k}l_x)^{1/2}). \quad (26)$$

Scattering in the transverse horizontal direction is investigated by choosing $y = R\sin\theta\sin\varphi = 0$ or $\cos\varphi = 1$. Equation (24) is then

$$\{\hat{I}_S^C(\mathbf{x}_H)\} = \frac{\sigma_0 \bar{k}^4 \hat{I} V}{(4\pi)^2 R^2} l_x^2 l_y \exp[-\tfrac{1}{2}\bar{k}^2 l_x^2 (1-\cos\theta)]. \quad (27)$$

This is similar to case (A) and yields

$$\theta_H = O(1/\bar{k}l_x). \quad (28)$$

Examination of the results of case (C) shows that for single scattering we may expect that the horizontal angular spread will be of order $1/\bar{k}l_x$ while the vertical spread will be of order $(1/\bar{k}l_x)^{1/2}$. The same type of analysis holds in general for an arbitrary function $\sigma(\mathbf{u})$. We only require that $\bar{k}l_{ym} \ll 1$ and $\bar{k}l_{xm} \gg 1$.

The statement that

$$\theta = O(1/\bar{k}l_{xm})$$

means that scattering from the smallest scale fluctuation gives $\theta = O(1/\bar{k}l_{xm})$. The largest scales give $\theta = O(1/\bar{k}l_{xM})$. Since, however, $\bar{k}l_{xM} > \bar{k}l_{xM}$, we usually use only the order of magnitude associated with $\bar{k}l_{xm}$.

For the case $\bar{k}l_{yM} \ll [(\bar{k}l_{xM})^{1/2}(l_{xm}/l_{xM})^{1/2}]$ similar analysis shows that θ is $O(1/(\bar{k}l_{xm})^{1/2})$. In general we find that when $\bar{k}l_{xm} \geq 1$, θ is no greater than $O(1/\bar{k}l_{xm})^{1/2})$ for all values of $\bar{k}l_{yM}$.

The results given in Eqs. (26) and (28) are only valid for single scattering, but we shall see in the next section how these results motivate approximations that allow us to solve the multiple scattering problem.

3. MULTIPLE SCATTER SOLUTION

In this section we study the mutual coherence function, i.e., $\{\hat{\Gamma}(\mathbf{x}_1, \mathbf{x}_2, \bar{\nu})\}$; defined on the complex pressure field according to

$$\{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\bar{\nu})\}=\{\hat{p}(\mathbf{x}_1,\bar{\nu})\hat{p}^*(\mathbf{x}_2,\bar{\nu})\}. \tag{29}$$

We use this definition here in the interest of simplicity of presentation. See Refs. 1 and 3 for a proper definition for finite band stationary signals. The random medium is confined to the half-space $z>0$. Homogeneous statistics are assumed. The problem of a plane wave radiation field incident from $z\to-\infty$ is considered. Using the results of the last section we neglect any backscattered radiation which enables our writing the following expression for $\{\hat{\Gamma}\}$ for two points in the half-space $z<0$,

$$\{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2,\bar{\nu})\}=I\exp[i\bar{k}(z_1-z_2)],\quad z_1,z_2<0. \tag{30}$$

By virtue of the restriction to homogeneous statistics, the general solution in the half-space $z>0$ is of the form

$$\{\hat{\Gamma}(x_{12},y_{12},z_{12},z,\bar{\nu})\},$$

where $x_{12}=x_2-x_1$, $y_{12}=y_2-y_1$, $z_{12}=z_2-z_1$, and $z=z_1$. We are not interested in the general solution, however, and shall only determine the mutual coherence function measured at two points in the same plane normal to the original plane wave direction, i.e., $z_{12}=0$. It is this quantity that gives a measure of the resolution limitation resulting from the presence of the random medium.

A single scatter solution of the posed problem is immediately afforded by Eq. (13). It is well appreciated, however, that the validity of the solution so obtained has a range dependence and that a procedure for incorporating multiple scattering effects is necessary for longer propagation distances. In this section we use the procedure of Beran.[1] In an appendix we present a derivation based on a Bethe–Salpeter formalism. Both procedures are seen to lead to identical results.

In carrying out the solution the region between 0 and z is divided by a series of M-1 infinite planes located at the coordinates $\Delta z, 2\Delta z,\cdots$, where $\Delta z=z/M$. The number M is chosen large enough so that in the interval $j\Delta z$ and $(j+1)\Delta z$, $\{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2)\}$ (we suppress the $\bar{\nu}$ argument in $\{\hat{\Gamma}\}$) can be obtained from its value measured on $z_1=z_2=j\Delta z$ using a single scatter approximation. Thus, one can show that in this interval

$$\{\hat{\Gamma}_S(\mathbf{x}_1,\mathbf{x}_2)\}=\frac{\bar{k}^4}{(4\pi)^2}\int\int\frac{\exp\{i\bar{k}[r(\mathbf{x}_1,\mathbf{x}')-r(\mathbf{x}_2,\mathbf{x}'')]\}}{r(\mathbf{x}_1,\mathbf{x}')r(\mathbf{x}_2,\mathbf{x}'')}\times\sigma(\mathbf{x}',\mathbf{x}'')\{\hat{\Gamma}_{j\Delta z}(\mathbf{x}',\mathbf{x}'')\}\,d\mathbf{x}'\,d\mathbf{x}''. \tag{31}$$

In this equation $\{\hat{\Gamma}_S(\mathbf{x}_1,\mathbf{x}_2)\}$ is the scattered portion of $\{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}_2)\}$, and $\{\hat{\Gamma}_{j\Delta z}(\mathbf{x}_j',\mathbf{x}'')\}$ is the mutual coherence function that exists in the interval $j\Delta z<z<(j+1)\Delta z$ when there is no scattering in this interval. The integrations are over the region between the planes located $z=j\Delta z$ and $z=(j+1)\Delta z$. Equation (31) is equivalent to Eq. (7) of Beran.[1] For convenience we shall denote equations in that paper with a capital B, i.e., Eq. (B7).

In writing Eq. (31) we have assumed that $\mu(\mathbf{x}')\mu(\mathbf{x}'')$ and $\{\hat{\Gamma}_{j\Delta z}(\mathbf{x}',\mathbf{x}'')\}$ are uncorrelated in the interval $j\Delta z<z<(j+1)\Delta z$. This assumption is valid if we require that $\Delta z\gg l_{zM}$. If $\Delta z\gg l_{zM}$ then over most of the interval $\mu(\mathbf{x}')$ and $\mu(\mathbf{x}'')$, $z',z''>j\Delta z$, are uncorrelated to $\mu(\mathbf{x})$, $z<j\Delta z$. On the other hand $\hat{\Gamma}_{j\Delta z}(\mathbf{x}',\mathbf{x}'')$ is only dependent on $\mu(\mathbf{x})$, $z<j\Delta z$, since the scattering is in the forward direction. Therefore to a good approximation $\mu(\mathbf{x}')\mu(\mathbf{x}'')$ and $\hat{\Gamma}_{j\Delta z}(\mathbf{x}',\mathbf{x}'')$ may be assumed to be uncorrelated in the interval $j\Delta z<z<(j+1)\Delta z$.

It is desired to simplify the rhs of Eq. (31) based on the knowledge that the scattering is restricted to small angles. The procedure is the same as for the isotropic case, only here, since $\theta_y=O(1/(\bar{k}l_{zm})^{1/2})$ rather than $O(1/\bar{k}l_{zm})$, the small angle approximation is weaker and we must impose more stringent conditions than those given in Eq. (3).

An expansion and truncation of the expressions for $r(\mathbf{x}_1,\mathbf{x}')$ and $r(\mathbf{x}_2,\mathbf{x}'')$ yield simplified expressions that can be validly used provided we can show that the neglected terms are small. Thus, we approximate $r(\mathbf{x}_1,\mathbf{x}')$ in the exponent, by

$$r(\mathbf{x},\mathbf{x}')\approx(z_1-z')+\frac{1}{2}\frac{[(x_1-x')^2+(y_1-y')^2]}{z_1-z'} \tag{32}$$

and, in the denominator by the single term, z_1-z'. This is valid provided

$$\frac{(x_1-x')^2+(y_1-y')^2}{(z_1-z')^2}\ll 1 \tag{33}$$

and

$$\frac{\bar{k}[(x_1-x')^4+(y_1-y')^4]}{(z_1-z')^3}\ll 1. \tag{34}$$

The condition required by Eq. (33) is satisfied if

$$\theta_x^2\ll 1\quad\text{and}\quad\theta_y^2\ll 1, \tag{35}$$

where θ_x and θ_y are the angular spreads measured at a generic point in the interval in the x and y directions, respectively. The results of the last section justify the assumption of small angle scattering incorporated in this condition. The condition required by Eq. (34) is satisfied for all points in the interval $j\Delta z<z_1<(j+1)\Delta z$ if

$$\bar{k}(\Delta z)\theta_x^4\ll 1\quad\text{and}\quad\bar{k}(\Delta z)\theta_y^4\ll 1. \tag{36}$$

This condition can be interpreted as an upper bound limitation on the interval size (Δz). The results of the last section demonstrate that it is the second of the two conditions given that will be the most difficult to satisfy. By using the result that

$$\theta_y=O(1/(kl_{zm})^{1/2}), \tag{26}$$

Eq. (30) requires that

$$\Delta z/\bar{k}l_{zm}^2\ll 1. \tag{37}$$

This condition together with the condition that $\Delta z\gg l_{zM}$ jointly require that

$$l_{zM}/\bar{k}l_{zm}^2\ll 1. \tag{38}$$

We shall accept Eq. (38) as a restriction on the theory being developed.

Introducing the simplified expressions for $r(\mathbf{x}_1,\mathbf{x}')$ and $r(\mathbf{x}_2,\mathbf{x}'')$ into Eq. (31) and introducing the transformed coordinates

$$\mathbf{s}=\mathbf{x}''-\mathbf{x}',\quad\mathbf{p}=\mathbf{x}' \tag{39}$$

yields the following expression for $\{\hat{\Gamma}_S(\mathbf{x}_1,\mathbf{x}_2)\}$:

$$\begin{aligned}\{\hat{\Gamma}_S(\mathbf{x}_1,\mathbf{x}_2)\}\\ &= \frac{\bar{k}^4}{(4\pi)^2}\exp[i\bar{k}(z_1-z_2)]\int\int\frac{\exp(i\bar{k}s_z)}{(z_1-p_z)(z_2-s_z-p_z)}\\ &\times\exp\left[\frac{i\bar{k}}{2}\left(\frac{(x_1-p_x)^2+(y_1-p_y)^2}{(z_1-p_z)}\right.\right.\\ &-\left.\left.\frac{(x_2-s_x-p_x)^2+(y_2-s_y-p_y)^2}{(z_2-s_z-p_z)}\right)\right]\\ &\times\sigma(\mathbf{s})\{\hat{\Gamma}_{j\Delta z}(\mathbf{s},p_z)\}\,d\mathbf{s}\,d\mathbf{p}.\end{aligned} \tag{40}$$

We next wish to introduce the following simplified expressions into the exponent appearing in Eq. (40).

$$\frac{(x_2-s_x-p_x)^2}{(z_2-s_z-p_z)}\approx\frac{(x_2-s_x-p_x)^2}{(z_2-p_z)} \tag{41}$$

and

$$\frac{(y_2-s_y-p_y)^2}{(z_2-s_z-p_z)}\approx\frac{(y_2-s_y-p_y)^2}{(z_2-p_z)}+\frac{s_z(y_2-s_y-p_y)^2}{(z_2-p_z)^2}. \tag{42}$$

Again looking to the first term neglected, the condition given by Eq. (41) requires, in addition to $l_{zM}\ll\Delta z$,

$$\bar{k}s_z(x_2-s_x-p_x)^2/(z_2-p_z)^2\ll 1,$$

which is satisfied if

$$\bar{k}l_{zM}\theta_x^2\ll 1. \tag{43}$$

By using the result of Section 2 that

$$\theta_x=O(1/\bar{k}l_{zm}),$$

Eq. (43) leads to the condition already accepted, i.e., Eq. (38).

To justify Eq. (42), we must show that

$$\bar{k}s_z^2(y_2-s_y-p_y)^2/(z_2-p_y)^3\ll 1.$$

This, in turn, requires

$$\bar{k}l_{zM}^2\theta_y^2/\Delta z\ll 1,$$

which is a more severe restriction on the formalism than $\Delta z\gg l_{zM}$. Using the order of magnitude estimate of θ_y, we write

$$l_{zM}^2/(\Delta z)l_{zm}\ll 1. \tag{44}$$

This, together with Eq. (37), leads to a more severe restriction than Eq. (38) namely

$$l_{zM}^2/\bar{k}l_{zm}^3\ll 1. \tag{45}$$

We shall accept Eqs. (44) and (45) as restrictions on the theory.

Equations (41) and (42) are now introduced into Eq. (41). In addition, we replace $z_2-s_z-p_z$ in the denominator by z_2-s_z, which is consistent with all of the approximations already introduced. The result is written

$$\begin{aligned}\{\hat{\Gamma}_S(\mathbf{x}_1,\mathbf{x}_2)\} &= \frac{\bar{k}^4}{(4\pi)^2}\exp[i\bar{k}(z_1-z_2)]\int\int\frac{\exp(i\bar{k}s_z)}{(z_1-p_z)(z_2-p_z)}\\ &\times\exp\left\{\frac{i\bar{k}}{2}\left[\left(\frac{(x_1-p_x)^2}{(z_1-p_z)}-\frac{(x_2-p_x-s_x)^2}{(z_2-p_z)}\right)\right.\right.\\ &+\left(\frac{(y_1-p_y)^2}{(z_1-p_z)}-\frac{(y_2-p_y-s_y)^2}{(z_2-p_z)}\right.\\ &-\left.\left.\left.\frac{s_z(y_2-p_y-s_y)^2}{(z_2-p_z)^2}\right)\right]\right\}\sigma(\mathbf{s})\{\Gamma_{j\Delta z}(\mathbf{s},\mathbf{p})\}\,d\mathbf{s}\,d\mathbf{p}.\end{aligned} \tag{46}$$

We next set $z_1=z_2=z$. The integral over p_x can then be readily performed since the p_x^2 terms cancel. We find for this contribution

$$(2\pi/\bar{k})(z-p_z)\delta(s_x-x_{12}).$$

The integral over p_y is more complex, but after some manipulation we find the contribution

$$(2\pi)^{1/2}\frac{(z-p_z)}{(\bar{k}|s_z|)^{1/2}}\exp\left\{i\bar{k}\left[\frac{(y_{12}-s_y)^2}{(z_2-p_z)}\pm\left(\frac{(y_{12}-s_y)^2}{2|s_z|}-\frac{\pi}{4\bar{k}}\right)\right]\right\}.$$

(Here the upper sign corresponds to $s_z>0$ and the lower sign to $s_z<0$).

Using the above expressions, Eq. (B13) is replaced by

$$\begin{aligned}\{\hat{\Gamma}_S(x_{12},y_{12},z)\} &= \left(\frac{2}{\pi}\right)^{1/2}\frac{\bar{k}^3}{8}\\ &\times\int\int\int\exp\left\{i\left[\bar{k}(y_{12}-s_y)^2\left(\frac{1}{(z_2-p_z)}\right.\right.\right.\\ &\pm\left.\left.\left.\frac{1}{2|s_z|}\right)\right]\right\}\frac{\sigma(x_{12},s_y,s_z)}{(\bar{k}|s_z|)^{1/2}}\exp[i(\bar{k}s_z\mp\tfrac{1}{4}\pi)]\\ &\times\{\hat{\Gamma}_{j\Delta z}(x_{12},s_y,s_z,p_z)\}\,ds_y\,ds_z\,dp_z.\end{aligned} \tag{47}$$

It is possible to introduce a further simplification and carry out the integration over the s_y coordinate by making use of the smallness of the maximum eddy size measured in the y (i.e., the depth) direction. First, we notice that the characteristic spread of $\{\Gamma_{j\Delta z}\}$ in the s_y direction is of order $(l_{zm}/\bar{k})^{1/2}$ (or $1/\bar{k}\theta_y$). [We shall see from the solution given in Eq. (62) that it is consistent to assume that the order of magnitude of θ_y is the same in the multiple scatter region as in the single scatter region.] Thus, if we require that this distance be much greater than l_{yM} the integration over s_y may be performed upon replacing the s_y argument in $\{\hat{\Gamma}_{j\Delta z}\}$ by zero. In addition s_y may be set equal to zero in the exponential terms if we have the somewhat stronger condition:

$$\bar{k}l_{yM}\ll(\bar{k}l_{zm})^{1/2}(l_{zm}/l_{zM})^{1/2}. \tag{48}$$

[We note that we have already required that $(kl_{zm}^2)/l_{zM}\gg 1$, i.e., see Eq. (38).] This same restriction also enables us to approximate

$$\exp[i\bar{k}y_{12}^2/(z_2-p_z)]\approx 1.$$

Carrying out the integration over s_y as a result of these three approximations leads to

$$\begin{aligned}\{\hat{\Gamma}_S(x_{12},y_{12},z)\}\\ &= \left(\frac{2}{\pi}\right)^{1/2}\frac{\bar{k}^3}{8}\int\int\exp\left[\pm i\left(\frac{\bar{k}y_{12}^2}{2|s_z|}-\frac{\pi}{4}\right)\right]\\ &\times\frac{\sigma_2(x_{12},s_z)}{(\bar{k}|s_z|)^{1/2}}\exp(i\bar{k}s_z)\{\hat{\Gamma}_{j\Delta z}(x_{12},0,s_z,p_z)\,ds_z\,dp_z,\end{aligned} \tag{49}$$

where

$$\sigma_2(x_{12},s_z)=\int_{-\infty}^{\infty}\sigma(x_{12},s_y,s_z)\,ds_y.$$

In this paper we accept the restriction given by Eq. (48) and make use of Eq. (49). We note, however, that a theory could be developed that would be valid for arbitrary $\bar{k}l_{yM}$ by retaining Eq. (47). We would then find that $\{\Gamma\}$

is governed by an integral equation, the integration being over the variable s_y.

As a final observation we note that for narrow angle propagation we can write

$$\exp(i\bar{k}s_z)\{\hat{\Gamma}_{j\Delta z}(x_{12},0,s_z,p_z)\}\approx\{\hat{\Gamma}_{j\Delta z}(x_{12},0,0,0)\}$$

for $|s_z|<l_M$ and $p_z<\Delta z$. This allows us to carry out the integration over s_z and p_z leading to

$$\{\hat{\Gamma}_S(x_{12},y_{12},z)\}=\bar{\sigma}_2(x_{12},y_{12})z'\{\hat{\Gamma}_{j\Delta z}(x_{12},0,j\Delta z)\}, \tag{50}$$

where

$$\bar{\sigma}_2(x_{12},y_{12})=\left(\frac{2}{\pi}\right)^{1/2}\frac{\bar{k}^3}{4}\int_0^{-\infty}\cos\left(\frac{\bar{k}y_{12}^2}{2s_z}-\frac{\pi}{4}\right)(ks_z)^{-1/2}\times\sigma_2(x_{12},s_z)\,ds_z, \tag{51}$$

and $z'=z-j\Delta z$.

The condition for the validity of the single scattering approximation (which is a perturbation approximation) is readily seen to be

$$|\bar{\sigma}_2(x_{12},y_{12})|\Delta z\ll 1. \tag{52}$$

The intensity of the scattered radiation is

$$\{\hat{I}_S(z)\}=\{\hat{\Gamma}_S(0,0,z)\}=\bar{\sigma}_2(0,0)\hat{I}z', \tag{53}$$

where $\hat{I}$ is the intensity of the initial radiation. For small angle scattering the intensity remains a constant independent of z. Thus, the intensity of the unscattered radiation must be

$$\{\hat{I}_U(z)\}=[1-\bar{\sigma}_2(0,0)z']\hat{I}. \tag{54}$$

In this statistically homogeneous problem $\{\hat{\Gamma}_{j\Delta z}(x_{12},y_{12},j\Delta z)\}$ may be derived by considering the superposition of an angular spectrum of plane waves. For small angle scattering the power in each plane wave is reduced by the same amount and thus

$$\{\hat{\Gamma}_U(x_{12},y_{12},z)\}=\{\hat{\Gamma}_{j\Delta z}(x_{12},y_{12},j\Delta z)\}[1-\bar{\sigma}_2(0,0)z']. \tag{55}$$

The mutual coherence function is now given as the sum of the scattered and unscattered parts since these parts are uncorrelated. (The lack of correlation may be proven by a direct calculation.) Therefore, we write

$$\{\hat{\Gamma}(x_{12},y_{12},z)\}=\{\hat{\Gamma}_{j\Delta z}(x_{12},y_{12},j\Delta z)\}[1-\bar{\sigma}_2(0,0)z'\}+\{\hat{\Gamma}_{j\Delta z}(x_{12},0,j\Delta z)\}\bar{\sigma}_2(x_{12},y_{12})z', \tag{56}$$

which leads to

$$\{\hat{\Gamma}_{(j+1)\Delta z}(x_{12},y_{12},(j+1)\Delta z)\}=\{\hat{\Gamma}_{j\Delta z}(x_{12},y_{12},j\Delta z)\}[1-\bar{\sigma}_2(0,0)\Delta z]+\{\hat{\Gamma}_{j\Delta z}(x_{12},0,j\Delta z)\}\bar{\sigma}_2(x_{12},y_{12})\Delta z. \tag{57}$$

The difference equation can be approximated by the following differential equation

$$\frac{d\{\hat{\Gamma}(x_{12},y_{12},z)\}}{dz}=-\{\hat{\Gamma}(x_{12},y_{12},z)\}\bar{\sigma}_2(0,0)+\{\hat{\Gamma}(x_{12},0,z)\}\bar{\sigma}_2(x_{12},y_{12}). \tag{58}$$

We note that Eq. (58) follows exactly from Eq. (57) in the limit of $\Delta z\to 0$. In our treatment, however, this step must be taken as an approximation since in deriving Eq. (57) we introduced the restriction that $\Delta z\gg l_{zM}$. The nature of the approximation is similar to that used in continuum fluid mechanics where we allow the elemental volume size, ΔV to approach zero even though it must satisfy the restriction $(\Delta V)^{1/3}\gg l_p$, where l_p is the molecular mean free path.

The solution of Eq. (55) for an initial plane wave is

$$\{\hat{\Gamma}(x_{12},y_{12},z)\}=I\left[\frac{\bar{\sigma}_2(x_{12},y_{12})}{\bar{\sigma}_2(x_{12},0)}\exp\{-[\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0)]z\}+\left(1-\frac{\bar{\sigma}_2(x_{12},y_{12})}{\bar{\sigma}_2(x_{12},0)}\right)\exp[-\bar{\sigma}_2(0,0)z]\right]. \tag{59}$$

To study the coherence function $\{\hat{\Gamma}(x_{12},y_{12},z)\}$ we thus only require a knowledge of the function $\bar{\sigma}_2(x_{12},y_{12})$.

Two special cases of Eq. (59) are of interest. They are

$$\{\hat{\Gamma}(x_{12},0,z)\}=\hat{I}\exp\{-[\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0)]z\}, \tag{60}$$

and

$$\{\hat{\Gamma}(0,y_{12},z)\}=\hat{I}\left[\frac{\bar{\sigma}_2(0,y_{12})}{\bar{\sigma}_2(0,0)}+\left(1-\frac{\bar{\sigma}_2(0,y_{12})}{\bar{\sigma}_2(0,0)}\right)\exp[-\bar{\sigma}_2(0,0)z]\right]. \tag{61}$$

The function $\{\hat{\Gamma}(x_{12},0,z)\}$, for example, allows us to determine horizontal resolution of an aperture system while $\{\hat{\Gamma}(0,y_{12},z)\}$ allows us to determine vertical resolution. As $z\to\infty$, Eq. (61) approaches the simple limit

$$\{\hat{\Gamma}(0,y_{12},z\to\infty)\}=\hat{I}[\bar{\sigma}_2(0,y_{12})/\bar{\sigma}_2(0,0)]. \tag{62}$$

We synopsize here the conditions to be satisfied for the validity of Eq. (58):

$$\bar{k}l_{yM}\ll(kl_{zm})^{1/2}(l_{zm}/l_{zM})^{1/2} \tag{63a}$$

and

$$\bar{\sigma}_2(0,0)\Delta z\ll 1, \tag{63b}$$

where Δz is a distance that satisfies the inequalities

$$l_{zM}^2/l_{zm}\ll\Delta z\ll 1/\bar{k}\theta_x^2 \tag{64}$$

An alternate derivation of Eq. (58) is given in the Appendix.

4. RESOLUTION LIMITATIONS CAUSED BY SCATTERING

Equations (60) and (61) can be used to investigate the resolution limitations that are a consequence of scattering by an anisotropic random medium. Since our primary interest is in horizontal resolution limitations that result from the scattering of acoustic signals by the ocean temperature microstructure, we shall restrict attention to Eq. (60). From this equation we see that $\{\hat{\Gamma}(x_{12},0,z)\}$ decays exponentially with increasing separation distance x_{12} ($z=$ fixed) according to the following functional form for the exponent:

$$[\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0)]z.$$

It is well known that the horizontal resolution limitation can be related to a characteristic distance defined by this decay. Further, it is easily demonstrated that the maximum useful length of a horizontal array, for the coherent summing of signals, is directly related to such a characteristic decay distance.

In this section we study the function $\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0)$ and obtain a qualitative discription of the loss of spatial coherence due to scattering. We also define, somewhat arbitrarily, the characteristic decay distance l_Γ as the $1/e$ distance, i.e., by the condition

$$[\bar{\sigma}_2(0,0)-\bar{\sigma}_2(l_\Gamma,0)]z=1. \tag{65}$$

We obtain algebraic expressions for l_Γ for media with index of refraction fluctuations that have horizontal power spectra that follow a simple power law. The extension of these efforts to power spectra that are given by a linear sum of simple power laws is readily achieved. Experimental data and theoretical predictions indicate that the ocean temperature microstructure gives rise to a random medium that might suitably be described by a combination of power laws. The larger size scale temperature fluctuations ($\approx 1-25$km) are thought to result due to the presence of randomly phased internal waves. Internal waves result in a p^{-2} power spectrum, where p denotes the inverse space coordinate (Phillips[4]). The smaller scale temperature fluctuations are a result of ocean turbulence. For length scales of the order of ≈ 300m to ≈ 0.1m, the spectrum might be described by a Kolmogorov spectrum, i.e., $p^{-5/3}$. A transition range, in which bouyancy forces play a role, appears to separate these two regions. In the transition region a p^{-3} power law is indicated (Moseley and Del Balzo[5]).

We write the expression for $\bar{\sigma}_2(x_{12},0)$:

$$\bar{\sigma}_2(x_{12},0)=\frac{\bar{k}^3}{4\sqrt{\pi}}\int_0^\infty \frac{\sigma_2(x_{12},s_z)}{(\bar{k}s_z)^{1/2}}\,ds_z, \tag{66}$$

where

$$\sigma_2(x_{12},s_z)=\int_{-\infty}^\infty \sigma(x_{12},s_y,s_z)\,ds_y.$$

For the intended application of the present study, experimental data of index of refraction fluctuations with depth is largely lacking. Almost all reported data is based on horizontal measurements, i.e., information on $\sigma(x_{12},0,s_z)$. Although a simple relationship may not exist between $\sigma_2(x_{12},s_z)$ and $\sigma(x_{12},0,s_z)$, we shall assume that one does and, in particular, assume that

$$\sigma_2(x_{12},s_z)=l_{yM}\sigma(x_{12},0,s_z). \tag{67}$$

Further, we shall restrict attention to media that have isotropic statistics for measurements taken in a horizontal plane, i.e., $\sigma(x_{12},0,s_z)=\sigma((x_{12}^2+s_z^2)^{1/2},0)$. Thus,

$$\bar{\sigma}_2(x_{12},0)=\frac{\bar{k}^3 l_{yM}}{8\sqrt{\pi}}\int_{-\infty}^\infty \frac{\sigma((x_{12}^2+s_z^2)^{1/2},0)}{(\bar{k}s_z)^{1/2}}\,ds_z. \tag{68}$$

It is convenient to discuss the horizontal fluctuations in the index of refraction in terms of the one-dimensional spectrum $\Phi_1(p)$, given by

$$\Phi_1(p)=\frac{1}{2\pi}\int_{-\infty}^\infty \sigma(q,0)\exp(ipq)\,dq, \tag{69}$$

where

$$q=(x_{12}^2+s_z^2)^{1/2}$$

Introducing the inverse of Eq. (69) into Eq. (68) yields

$$\bar{\sigma}_2(x_{12},0)=\frac{\bar{k}^3 l_{yM}}{8\sqrt{\pi}}\int_{-\infty}^\infty \frac{ds_z}{(ks_z)^{1/2}} \times\int_{-\infty}^\infty \Phi_1(p)\exp[-ip(x_{12}^2+s_z^2)^{1/2}]\,dp. \tag{70}$$

Assuming that the orders of integrations may be interchanged the integration over s_z may be accomplished which leads to the following result:

$$\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0) =\frac{1}{2^{3/2}}(\bar{k}l_{yM})(\bar{k}x_{12})^{1/2}\bar{k} \times\int_0^\infty\left[\frac{1}{(px_{12})^{1/2}}-\frac{\Gamma(\frac{1}{4})}{2^{3/4}}J_{-3/4}(px_{12})(px_{12})^{1/4}\right]\Phi_1(p)\,dp. \tag{71}$$

Here, $\Gamma(\frac{1}{4})$ denotes the gamma function and $J_{-3/4}(\eta)$ denotes the Bessel function. Thus, for a specified value for $x_{12}, \bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0)$ is a linear functional of the one-dimensional power spectrum $\Phi_1(p)$. The kernel of the functional is given by

$$F(px_{12})=\frac{1}{(px_{12})^{1/2}}-\frac{\Gamma(\frac{1}{4})}{2^{3/2}}(px_{12})^{1/4}J_{-3/4}(px_{12}). \tag{72}$$

A graphical representation of $F(px_{12})$ is given in Fig. 2. Changing the value of x_{12} amounts to a changing of the scale of the abscissa when viewing F in p space.

Using Eq. (71) and Fig. 2, we can construct a qualitative description of the dependence of the coherence function on x_{12} and its relationship to the one-dimensional power spectrum. To do so, it is necessary to have a visualization of the power spectrum. A schematic of a typical spectrum of interest is illustrated in Fig. 3. We note that the spectrum is band limited ranging between a low wavenumber cutoff, denoted by p_M and a high wavenumber cutoff, denoted by p_m. The value of p_m is typically several orders of magnitude greater than that of p_M. The spectrum rises very rapidly to a maximum value in the vicinity of p_M and then decreases monotonically with increasing p. The range of values of $\Phi_1(p)$ between its maximum and the value at its high wavenumber cutoff is, typically, several orders of magnitude. We consider the dependence of the coherence function on x_{12} for x_{12} very large, i.e., of the order of p_M^{-1}. Large x_{12} corresponds to a much compressed abscissa for viewing $F(px_{12})$ in p space. The falloff of $F(px_{12})$ and the

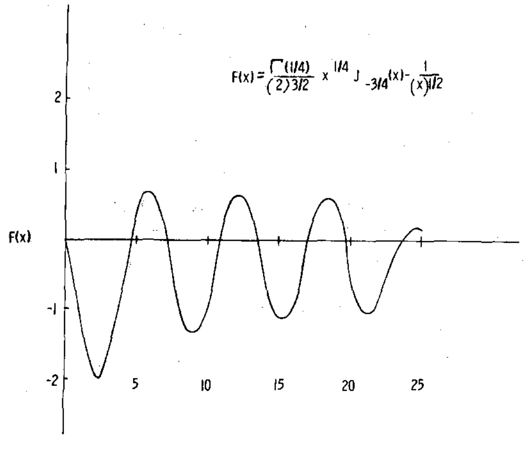


FIG. 2.

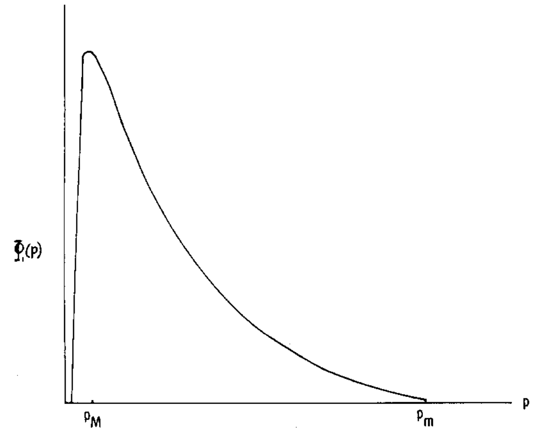


FIG. 3.

very rapid falloff of $\Phi_1(p)$ with increasing p combine to suppress the contribution of the high wavenumber portion of the index of refraction fluctuations spectrum. Decreasing x_{12} results in a stretching of the abscissa for viewing $F(px_{12})$ in p space thus reducing the rate of falloff with increasing p. This leads to an increase in the maximum wavenumber for which $\Phi_1(p)$ offers a significant contribution to the mutual coherence function. Further, we note the low wavenumber decrease of $F(px_{12})$. For values of $x_{12} \ll p_M^{-1}$, this decrease results in a suppression of the contribution of the low wavenumber portion of the spectrum. Thus, we have a picture of a limited portion of the total spectrum contributing significantly to the coherence function of the acoustic signal for a specified value of the separation distance x_{12}. The specific portion of the spectrum that contributes ranges from low wavenumber, i.e., large correlation lengths, to high wavenumber, small correlation lengths, with decreasing separation distance. This fact is important in discussing the applicability of the theory to a given physical problem, e.g., scattering of acoustic signals by the temperature microstructure. The limitations on the correlation lengths listed in Eq. (63) and Eq. (64) should be interpreted in terms of the portion of the spectrum that significantly contributes to the prediction problem of interest. Thus, applicability can only be discussed in terms of an *a* posteori comparison of the important portion of the spectrum and the theory limitations.

We note from Eq. (65) that l_Γ decreases with increasing range, z. Thus, we conclude that the dominant portion of the index of refraction fluctuations spectrum, for determining the maximum coherence length, changes from low wavenumbers to high wavenumbers with increasing range. This is in agreement with our physical intuition. At short ranges the limitation of resolution is controlled by the higher energy large scale fluctuations. At longer ranges, multiple scatter effects enter and the smaller scale fluctuations become of importance.

To proceed further requires the introduction of a specific functional form for $\Phi_1(p)$. We choose, here, a simple power law to illustrate the required calculations. That is, we choose

$$\begin{aligned}\Phi_1(p) &= A_n^2/(p^2+p_M^2)^{n/2}, \quad p<p_m,\\ &=0, \qquad\qquad\qquad\quad p>p_m.\end{aligned} \tag{73}$$

We note the low and high wavenumber cutoffs in Eq. (73). Since experimental data of $\Phi_1(p)$ will be limited to a finite range of values of p, these cutoffs represent extrapolations of the measured data to all ranges of p. We do not simply extrapolate p^{-n} since this would not result in covergent integrals for all values of x_{12} for the values of n of interest. Further, we know from physical considerations that cutoffs do exist for spectra that could represent the temperature fluctuations in the ocean. The spectrum of Eq. (73) achieves its maximum value at $p=0$, which is to be compared with the rapid fall of the typical spectrum illustrated schematically in Fig. 3. This difference will be significant only for x_{12} of the order of p_M^{-1}. For $x_{12} \ll p_M^{-1}$, the form of $F(px_{12})$ will suppress the difference. We note that the low wavenumber cutoff used in Eq. (73) will enable the analytic evaluation of the integral in Eq. (71).

Introducing the assumed form of $\Phi_1(p)$ into Eq. (71), the integration can be carried out provided we restrict attention to values of $x_{12} \gg p_m^{-1}$, in which case we can allow p_m to approach infinity. (See, for example, Gradshteyn and Ryzhik.[6]) Thus we write

$$\begin{aligned}\overline{\sigma}_2(0,0)-\overline{\sigma}_2(x_{12},0) &= \frac{\Gamma(\frac{1}{4})}{2^{5/2}\Gamma(n/2)}(\overline{k}l_{yM})\overline{k}^{3/2}A_n^2\\ &\quad\times\left[\frac{\Gamma((2n-1)/4)}{p_M^{(2n-1)/2}}-\frac{1}{2^{(2n-5)/4}}\right.\\ &\quad\left.\times\left(\frac{|x_{12}|}{p_M}\right)^{(2n-1)/4}K_{(2n-1)/4}(p_M|x_{12}|)\right]\end{aligned} \tag{74}$$

Here, $K_{(2n-1)/4}(p_M|x_{12}|)$ denotes a modified Bessel function. Equation (74) can be considerably simplified for the range of values $x_{12} \ll p_M^{-1}$. In this case the Bessel function can be expanded and truncated. The form of the single term approximation depends on the value of n. For $n<\frac{5}{2}$,

$$\begin{aligned}&\overline{\sigma}_2(0,0)-\overline{\sigma}_2(x_{12},0)\\ &\quad=\left(\frac{\pi\Gamma(\frac{1}{4})}{2^{n+2}\Gamma(n/2)\Gamma((2n+3)/4)\sin[(2n-1)/4\pi]}\right)\\ &\qquad\times A_n^2\overline{k}^{5/2}l_{yM}|x_{12}|^{(2n-1)/2}\end{aligned} \tag{75}$$

and, for $n>\frac{5}{2}$,

$$\begin{aligned}&\overline{\sigma}_2(0,0)-\overline{\sigma}_2(x_{12},0)\\ &\quad=\left(\frac{-\pi\Gamma(\frac{1}{4})}{2^{9/4}\Gamma(n/2)\Gamma((9-2n)/4)\sin[(2n-1)/4\pi]}\right)\\ &\qquad\times\frac{A_n^2\overline{k}^{5/2}l_{yM}x_{12}^2}{p_M^{(2n-5)/2}}.\end{aligned} \tag{76}$$

The most important distinction between these two expressions is that for $n>\frac{5}{2}$ the rate of fall of the simple power law is rapid enough that the value of the coherence function for $x_{12} \ll p_M^{-1}$ depends on the value of p_M. Substituting Eq. (75) or Eq. (76) into Eq. (65) leads to simple algebraic expressions for l_Γ. We present the following expressions for $n=\frac{5}{3}$, $n=2$, and $n=3$, the three most commonly discussed values for the ocean temperature microstructure:

$$n=\tfrac{5}{3}, \quad l_\Gamma=1.07A_{5/3}^{-12/7}\overline{k}^{-15/7}l_{yM}^{-6/7}z^{-6/7},$$

$$n=2,\quad l_\Gamma = 0.94A_2^{-4/3}\bar{k}^{-5/3}l_{yM}^{-2/3}z^{-2/3}, \tag{77}$$
$$n=3,\quad l_\Gamma = 0.56A_3^{-1}\bar{k}^{-5/4}p_M^{1/4}l_{yM}^{-1/2}z^{-1/2}.$$

These expressions are valid provided $p_M^{-1} \gg l_\Gamma \gg p_m^{-1}$.

For values of x_{12} of the order of p_m^{-1} the form of the spectrum in the vicinity of the high wavenumber cutoff becomes important. A simplified expression is not possible for $\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0)$ in this range. For $x_{12} \ll p_m^{-1}$, one can again obtain a simplified expression by expanding and truncating the Bessel function. Thus, we have

$$\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0) = \frac{\Gamma(\frac{1}{4})}{2^{9/4}\Gamma(\frac{5}{4})}(\bar{k}l_{yM})\bar{k}^{3/2}x_{12}^2\int_0^{p_m} p^{3/2}\Phi_1(p)\,dp. \tag{78}$$

Equation (78) shows that $\{\hat{\Gamma}(x_{12},0,z)\}$ has a Gaussian form, with respect to x_{12}, independent of the form of $\Phi_1(p)$ for $x_{12} \ll p_m^{-1}$. For long enough ranges $\{\hat{\Gamma}(x_{12},0,z)\}$ will be nonzero only for such values of x_{12}. Thus, the coherence function approaches this Gaussian form in the limit.

The high wavenumber portion of the $\Phi_1(p)$ spectrum will dominate the form of $\{\hat{\Gamma}\}$ for $x_{12} \ll p_m^{-1}$. Thus, we substitute a truncated Kolmogorov spectrum into Eq. (78). This gives

$$\bar{\sigma}_2(0,0)-\bar{\sigma}_2(x_{12},0) = 1.01A^2(\bar{k}l_{ym})\bar{k}^{3/2}p_m^{5/6}x_{12}^2. \tag{79}$$

Thus, as $z \to \infty$, we have

$$\{\hat{\Gamma}(x_{12},0,z)\} = \hat{I}\exp[-1.01A^2(\bar{k}l_{ym})\bar{k}^{3/2}p^{5/6}x_{12}^2 z] \tag{80}$$

5. SUMMARY

In this paper we treated the problem of plane wave scattering by an anistropic medium subject to the conditions given in Eqs. (63) and (64). The general solution is given in Eq. (59). An expression which considers scattering in the horizontal direction (where $\bar{k}l_{xm} \gg 1$) is given by Eq. (60). Explicit calculations are given for $\Phi_1(p)$ given by Eq. (73).

The governing equation given in this paper was derived using the solution obtained in any interval $z, z+\Delta z$. An alternate approach, based on the Bethe–Salpeter equation, is given in the Appendix. In this derivation Eq. (48) is replaced by the condition $\bar{k}l_{ym} \ll 1$.

ACKNOWLEDGMENTS

The reported work was conducted for the Large Aperture Systems Branch, Acoustics Division, Naval Research Laboratory, Washington, D.C. and supported by the Office of Naval Research. The authors acknowledge this support and helpful suggestions of Dr. B.B. Adams, Branch chief, Dr. William Mosely, J. Griffin, and D. Del Balzo.

APPENDIX: ALTERNATE DEVELOPMENT OF THE THEORY

In this appendix we present an alternate derivation of Eq. (58), which is basic to all of the calculations and conclusions achieved in this report. This alternate derivation is based on a Bethe–Salpeter equation that can be written for the mutual coherence function. Simplifications of the fundamental equation that are asymptotically valid for limiting values of $\bar{k}l_i$, l_i being a measure of an eddysize measured in some direction, are then shown to lead to the desired result.

The Bethe–Salpeter equation for the mutual coherence function can be derived in either of two ways. One approach amounts to a partial summation of a Neumann perturbation series for the mutual coherence function. (See, for example, Frisch.[7]) The second approach is based on a perturbation procedure termed the method of smoothing. (See McCoy.[8]) Either approach lends to the following integral form of the Bethe–Salpeter equation:

$$\begin{aligned}\{\hat{\Gamma}_0(\mathbf{x}_1,\mathbf{x}_2)\} &= \{\hat{\Gamma}_0(\mathbf{x}_1,\mathbf{x}_2)\} \\ &\quad + \bar{k}^4\Big[\int\int G(\mathbf{x},\mathbf{x}')G(\mathbf{x}',\mathbf{x}'')\sigma(\mathbf{x}',\mathbf{x}'')\{\hat{\Gamma}(\mathbf{x}'',\mathbf{x}_2)\}\,d\mathbf{x}''\,d\mathbf{x}' \\ &\quad + \int\int G^*(\mathbf{x}_2,\mathbf{x}')G^*(\mathbf{x}',\mathbf{x}'')\sigma(\mathbf{x}',\mathbf{x}'')\{\hat{\Gamma}(\mathbf{x}_1,\mathbf{x}'')\}\,d\mathbf{x}''\,d\mathbf{x}' \\ &\quad + \int\int G(\mathbf{x}_1,\mathbf{x}')G^*(\mathbf{x}_2,\mathbf{x}'')\sigma(\mathbf{x}',\mathbf{x}'')\{\hat{\Gamma}(\mathbf{x}',\mathbf{x}'')\}\,d\mathbf{x}'\,d\mathbf{x}''\Big].\end{aligned} \tag{A1}$$

Here $\{\hat{\Gamma}_0(\mathbf{x}_1,\mathbf{x}_2)\}$ is the mutual coherence function in the absence of any scattering, i.e., for the problem posed,

$$\{\hat{\Gamma}_0(\mathbf{x}_1,\mathbf{x}_2)\} = I\exp[i\bar{k}(z_1-z_2)]. \tag{A2}$$

The integrations are over the scattering region, i.e., for the problem posed, the half-space $z>0$, and $G(\mathbf{x},\mathbf{x}')$ is the Green's function for the homogeneous mean medium, i.e.,

$$G(\mathbf{x},\mathbf{x}') = \exp[i\bar{k}r(\mathbf{x},\mathbf{x}')]/-4\pi r(\mathbf{x},\mathbf{x}'). \tag{A3}$$

The correlation function $\sigma(\mathbf{x}',\mathbf{x}'')$ is that defined in the main body of the report.

Specifically, the objective is to introduce simplifications that will enable our accomplishing most of the integrations required by Eq. (A1). The validity of the simplifications will be seen to be asymptotic in the double limit of $\bar{k}l_{Hm}$ increasing without bound and $\bar{k}l_{ym}$ becoming vanishingly small. Here, l_{Hm} is the minimum correlation length measured in a horizontal plane, i.e., an x-z plane, and l_{yM} is the maximum correlation length measured in the depth direction. We note that the presence of the mutual coherence function in the integrands necessitates our introducing, *a priori*, assumptions of the behavior of this function over a distance equal to a wavelength (i.e., $2\pi/\bar{k}$) or a correlation length. One can investigate the self-consistency of these assumptions by an *a posteriori* comparison of the assumption and the predictions made by the approximate theory that results from the assumptions. The solutions presented in the main body of the report support the assumptions. We also note that in this appendix we do not investigate questions pertaining to the relative rates at which the limits are to be taken nor do we consider specific numerical measures as to the validity of the limiting form as an approximation to be applied for finite values of $\bar{k}l_{Hm}$ and $\bar{k}l_{yM}$. These investigations would require *a posteriori* reasoning of the type presented in the main body of the report and would, in affect, reproduce the same arguments.

To accomplish the simplifications, we introduce a Cartesian coordinate system and consider the integrals required by Eq. (A1) for one coordinate axis at a time.

In this way we encounter a series of integrals that have a generic form

$$I_1(x)=\int_{-\infty}^{\infty} F(x,x')\exp[i\bar{k}\varphi(x')]\,dx', \tag{A4}$$

where $F(\mathbf{x},x')$ is nonzero only if the pair of points located by x and x' lie within a common region of length given by L_1, and $\varphi(x')$ is independent of $\bar{k}$. Two different approximations are to be considered depending on the relative lengths defined by $F(x,x')$, by $\bar{\lambda}=2\pi/\bar{k}$ and by $\varphi(x')$. In the first we require $L_1\varphi'/\bar{\lambda}\ll 1$. This allows our ignoring, to a first approximation, any variations of $\varphi(x')$ over the region in which $F(x,x')$ is nonzero. The approximation is written

$$I_1(x)\approx\left[\int_{-\infty}^{\infty} F(\xi)\,d\xi\right]\exp[i\bar{k}\varphi(x)]. \tag{A5}$$

In the second approximation, we require λ to be small relative to all lengths defined by $F(x,x')$ and $\varphi(x')$. This is just the condition required for a stationary phase approximation of the integral. The stationary phase approximation is based on the observation that, for $\bar{k}$ very large, the major contributor to an integral of the form given by Eq. (5) is the immediate vicinity of $x'=x'_s$, where x'_s satisfies

$$\frac{d\varphi(x'_s)}{dx'_s}=0. \tag{A6}$$

The rapid oscillations that arise from the exponential with $\bar{k}$ large will lead to a cancellation of contributions from points outside this region. Within the region of interest we approximate

$$F(x,x')\approx F(x,x'_S)$$

and

$$\varphi(x')\approx\varphi(x'_S)+\frac{1}{2}\frac{d^2\varphi(x'_S)}{dx'^2_S}(x'-x'_S)^2,$$

which leads to

$$I_1(x)\approx F(x,x'_S)\exp[i\bar{k}\varphi(x'_S)]\times\int_{\Delta}\exp\left[\frac{i\bar{k}}{2}\frac{d^2\varphi(x'_S)}{dx'^2_S}(x'-x'_S)^2\right]dx',$$

where Δdenotes the immediate vicinity of x'_S. Finally, we allow Δ to become unboundedly large since the contributions from the added region will cancel as $\bar{k}$ becomes large relative to the linear extent of Δ. The resulting integral can be evaluated, which leads to

$$I_1(x)\approx\left(\frac{2\pi}{\bar{k}}\right)^{1/2}F(x,x'_S)\left(\left|\frac{d^2\varphi(x'_S)}{dx'^2_S}\right|\right)^{-1/2}\times\exp i\left[\bar{k}\varphi(x'_S)+\frac{\pi}{4}\operatorname{sgn}\left(\frac{d^2\varphi(x'_S)}{dx'^2_S}\right)\right]. \tag{A7}$$

We consider the first integral on the rhs and introduce Cartesian coordinates. The integration over x'' is considered first and the integration over x' next. In both cases we resort to a stationary phase approximation because of the restriction that $l_{xm}\gg\bar{\lambda}$. [We note the assumption that variations of $\{\hat{\Gamma}(\mathbf{x}'',\mathbf{x}_2)\}$ with changes of x'' of the order of $\bar{\lambda}$ may be neglected.] The result is written

$$I_1(\mathbf{x}_1,\mathbf{x}_2)\approx\frac{i}{8\pi\bar{k}}\int\int\int\int\frac{\exp(i\bar{k}\{[(y_1-y')^2+(z_1-z')^2]^{1/2}+[(y'-y'')^2+(z'-z'')^2]^{1/2}\})}{[(y_1-y')^2+(z_1-z')^2]^{1/4}[(y'-y'')^2+(z'-z'')^2]^{1/4}}\times\sigma(0,y'-y'',z'-z'')\{\hat{\Gamma}(x_1,y'',z'',\mathbf{x}_2)\}\,dy''\,dy'\,dz''\,dz'.$$

Next, we consider the integral over y''. In this case we resort to the approximation expressed by Eq. (A5) because of the restriction that $l_{yM}\ll\lambda$. [We note the assumption that variations of $\{\Gamma(\mathbf{x}'',\mathbf{x}_2)\}$ with changes of y'' of the order of l_{yM} may be neglected.] The result is written

$$I(\mathbf{x}_1,\mathbf{x}_2)\approx\frac{i}{8\pi\bar{k}}\times\int\int\int\frac{\exp(i\bar{k}\{[(y_1-y')^2+(z_1-z')^2]^{1/2}+|z'-z''|\})}{[(y_1-y')^2+(z_1-z')^2]^{1/4}|z'-z''|^{1/2}}\times\sigma_2(0,z'-z'')\{\hat{\Gamma}(x_1,y',z'',\mathbf{x}_2)\}\,dy'\,dz''\,dz',$$

where

$$\sigma_2(0,z'-z'')=\int_{-\infty}^{\infty}\sigma(0,\xi,z'-z'')\,d\xi.$$

The integration over y' is considered next and the assumption is made that variations of $\{\hat{\Gamma}(x_1,y',z'',\mathbf{x}_2)\}$ with changes of y' of the order of $\bar{\lambda}$ may be neglected. Thus, we can again make use of stationary phase to write

$$I(\mathbf{x}_1,\mathbf{x}_2)\approx\frac{\exp(3i\pi/4)}{4(2\pi)^{1/2}\bar{k}^{3/2}}\int\int\sigma_2(0,z'-z'')\times\{\hat{\Gamma}(x_1,y_1,z'',\mathbf{x}_2)\}\frac{\exp[i\bar{k}(|z_1-z'|+|z'-z''|)]}{|z'-z''|^{1/2}}\,dz''\,dz'.$$

We interchange orders of integration and accomplish the integration over z', i.e.,

$$\int\sigma_2(0,z'-z'')\frac{\exp[i\bar{k}(|z_1-z'|+|z'-z''|)]}{|z'-z''|^{1/2}}\,dz'.$$

Subject to the restriction that $l_{zm}>\bar{\lambda}$, this integral is approximately given by

$$\int_0^{\infty}\frac{\sigma_2(0,s_z)}{s_z^{1/2}}\,ds_z\exp(i\bar{k}|z''-z_1|)$$

except for $|z''-z_1|<l_{zM}$. Within this layer, a more complex expression is required. We are interested in problems in which $z\gg l_{zM}$; hence the layer in which the approximation is not valid is very thin relative to the total region of integration over z''. Thus, we are justified in ignoring this layer. We introduce

$$\bar{\sigma}_2(0,0)=\frac{\bar{k}^3}{4(\pi)^{1/2}}\int_0^{\infty}\frac{\sigma_2(0,s_z)}{(ks_z)^{1/2}}\,ds_z$$

and write

$$I_1(\mathbf{x}_1,\mathbf{x}_2)\approx\frac{\exp(3i\pi/4)}{\sqrt{2}\,\bar{k}^4}\,\bar{\sigma}_2(0,0)\times\int\{\hat{\Gamma}(x_1,y_1,z'',\mathbf{x}_2)\}\exp(i\bar{k}|z''-z_1|)\,dz''. \tag{A8}$$

Turning to the second integral on the rhs of Eq. (A1), we perform a similar set of calculations which yields

$$I_2(\mathbf{x}_1,\mathbf{x}_2)\approx\frac{\exp(-3i\pi/4)}{\sqrt{2}\,\bar{k}^4}\,\bar{\sigma}_2(0,0)$$
$$\times\int\{\hat{\Gamma}(\mathbf{x}_1,x_2,y_2,z'')\}\exp(-i\bar{k}|z''-z_2|)\,dz''. \tag{A9}$$

We consider the last integral on the rhs of Eq. (1). Upon introducing Cartesian coordinates, we integrate over x' and x'' using the stationary phase approximation. Subject to the same restrictions and assumptions encountered in treating $I_1(\mathbf{x}_1,\mathbf{x}_2)$, we can now write

$$I_3(\mathbf{x}_1,\mathbf{x}_2)\approx\frac{1}{8\pi\bar{k}}\int\int\int\int$$
$$\frac{\exp(i\bar{k}\{[(y_1-y')^2+(z_1-z')^2]^{1/2}-[(y_2-y'')^2+(z_2-z'')^2]^{1/2}\})}{[(y_1-y')^2+(z_1-z')^2]^{1/4}[(y_2-y'')^2+(z_2-z'')^2]^{1/4}}$$
$$\times\sigma(x_{12},y'-y'',z'-z'')\{\hat{\Gamma}(x_1,y',z',x_2,y'',z'')\}\,dy''\,dy'\,dz''\,dz'.$$

The integration over y'' is next carried out using the approximation expressed by Eq. (A5). The result is written

$$I_3(\mathbf{x}_1,\mathbf{x}_2)\approx\frac{1}{8\pi\bar{k}}\int\int\int$$
$$\frac{\exp(i\bar{k}\{[(y_1-y')^2+(z_1-z')^2]^{1/2}-[(y_2-y')^2+(z_2-z'')^2]^{1/2}\})}{[(y_1-y')^2+(z_1-z')^2]^{1/4}[(y_2-y')^2+(z_2-z'')^2]^{1/4}}$$
$$\times\sigma_2(x_{12},z'-z'')\{\hat{\Gamma}(x_1,y',z',x_2,y',z'')\}\,dy'\,dz''\,dz',$$

where $\sigma_2(x_{12},z'-z'')$ is the obvious generalization of $\sigma_2(0,z'-z'')$. The integration over y' is now to be carried out using the stationary phase approximation. In this case the details are a bit more complex than they were in the preceding calculations. We note that the stationary phase point is located by

$$y'_S=(|\eta|y_2-|\xi|y_1)/(|\eta|-|\xi|),$$

where

$$\eta=z_2-z''$$

and

$$\xi=z_1-z'.$$

As usual, $|\ |$ denotes the absolute value. By direct substitution,

$$I_3(\mathbf{x}_1,\mathbf{x}_2)$$
$$\approx\frac{1}{4(2\pi)^{1/2}\bar{k}^{3/2}}\int\int\sigma_2(x_{12},z'-z'')$$
$$\times\{\hat{\Gamma}(x_1,y'_S,z',x_2,y'_S,z'')\}\frac{[y_{12}^2+(|\eta|-|\xi|)^2]^{1/4}}{||\eta|-|\xi||}$$
$$\times\exp(i\,\mathrm{sgn}(|\xi|-|\eta|)\{\bar{k}[y_{12}^2+(|\eta|-|\xi|)^2]^{1/2}-\tfrac{1}{4}\pi\})\,dz''\,dz', \tag{A10}$$

where sgn is the signum function. To proceed further, it is necessary for us to assume that

$$\{\hat{\Gamma}(x_1,y'_S,z',x_2,y'_S,z'')\}=\{\hat{\Gamma}(x_1,0,z',x_2,0,z'\}$$
$$\times\exp[-i\bar{k}(z''-z')]$$

for $|z''-z'|<l_{zM}$. In writing this expression we have made use of the statistical homogeneity of the radiation field in a vertical plane. Upon substitution of this approximation into the expression for $I_3(\mathbf{x}_1,\mathbf{x}_2)$, we can accomplish the integration over z''. A simplified expression can be obtained only if we restrict the two points $(\mathbf{x}_1,\mathbf{x}_2)$ to lie in the same vertical plane, i.e., $z_1=z_2=z$. Introducing this restriction, the integration to be carried out over z'' is written

$$\int\sigma_2(x_{12},z'-z'')\{[y_{12}^2+(|z-z'|-|z-z''|)^2]^{1/4}/||z-z'|$$
$$-|z-z''||\}\exp[i(\bar{k}\{\mathrm{sgn}(|z-z'|-|z-z''|)[y_{12}^2$$
$$+(|z-z'|-|z-z''|)^2]^{1/2}-(z''-z')\}$$
$$-\mathrm{sgn}(|z-z'|-|z-z''|)\pi/4)]\,dz''. \tag{A11}$$

To approximate this integral we treat two cases separately; i.e., $z'<z$ and $z'>z$. For $z'<z$, the integral may be written as

$$\int_0^z\sigma_2(x_{12},z'-z'')\frac{[y_{12}^2+(z'-z'')^2]^{1/4}}{|z'-z''|}$$
$$\times\exp[i(\bar{k}\{\mathrm{sgn}(z''-z')[y_{12}^2+(z''-z')^2]^{1/2}-(z''-z')\}$$
$$+\mathrm{sgn}(z'-z'')\pi/4)]\,dz''$$
$$+\int_z^\infty\sigma_2(x_{12},z'-z'')\frac{[y_{12}^2+(2z-z'-z'')^2]^{1/4}}{|2z-z'-z''|}$$
$$\times\exp[i(\bar{k}\{\mathrm{sgn}(2z-z'-z'')[y_{12}^2+(2z-z'-z'')^2]^{1/2}$$
$$-(z''-z')\}+\mathrm{sgn}(z'+z''-2z)\pi/4)]\,dz''.$$

We are interested in the restricted problem in which $l_{zm}\gg\bar{\lambda}$. By using the same reasoning that is the basis of the stationary phase approximation, the conclusion can be reached that the major contribution to the first integral is the region of z'' space that satisfies the inequality that $|z'-z''|\gg|y_{12}|$ provided we can insure that $l_{zM}\gg|y_{12}|$. We accept the limitation required by this restriction, which enables us to approximate the first integral by

$$\int_0^z\frac{\sigma_2(x_{12},z'-z'')}{|z'-z''|^{1/2}}\exp\left[i\left(\frac{\bar{k}y_{12}^2}{2(z'-z'')}+\mathrm{sgn}(z'-z'')\frac{\pi}{4}\right)\right]dz''.$$

This integral may, in turn, be approximated by

$$2\int_0^\infty\frac{\sigma_2(x_{12},s_z)}{s_z^{1/2}}\cos\left(\frac{\bar{k}y_{12}^2}{2s_z}-\frac{\pi}{4}\right)ds_z \tag{A13}$$

provided we ignore values of z'' and z' that fall with neighborhoods of $z=0$, $z=z$. This is consistent with a similar neglect discussed with respect to $I_1(\mathbf{x}_1,\mathbf{x}_2)$. Applying the stationary phase reasoning to the second of the two integrals in Eq. (A12), we can see that this integral is of higher order in terms of $(\bar{k}l_{zm})^{-1}$ than is the first integral and, hence, may be neglected. A similar conclusion is reached for all integrals encountered for the case $z'>z$. By substitution of Eq. (A13) into Eq. (10), we write

$$I_3(\mathbf{x}_{1T},\mathbf{x}_{2T},z)\approx\frac{\bar{\sigma}_2(x_{12},y_{12})}{\bar{k}^4}\int_0^z\{\hat{\Gamma}(x_1,0,z',x_2,0,z')\}\,dz', \tag{A14}$$

where

$$\bar{\sigma}_2(x_{12},y_{12})$$

$$= \frac{\sqrt{2}\,\bar{k}^3}{4\sqrt{\pi}} \int_0^\infty \sigma_2(x_{12}, s_z) \cos\left[\frac{k y_{12}^2}{2 s_z} - \frac{\pi}{4}\right] (k s_z)^{-1/2}\, ds_z. \tag{51}$$

The notation $\mathbf{x}_{1T}$ refers to a two dimensional vector in the x—y plane.

Substituting Eqs. (A8), (A9), and (A14) into Eq. (A1) yields

$$\begin{aligned}\{\hat{\Gamma}(\mathbf{x}_{1T}, z; \mathbf{x}_{2T}, z)\} &= I + \frac{\exp(3i\pi/4)}{\sqrt{2}}\,\bar{\sigma}_2(0,0) \\ &\quad\times \int_0^\infty \{\hat{\Gamma}(\mathbf{x}_{1T}, z'; \mathbf{x}_{2T}, z)\} \exp(i\bar{k}\,|z - z'|)\, dz' \\ &\quad+ \frac{\exp(-3i\pi/4)}{\sqrt{2}}\,\bar{\sigma}_2(0,0) \\ &\quad\times \int_0^\infty \{\hat{\Gamma}(\mathbf{x}_{1T}, z; \mathbf{x}_{2T}, z')\} \exp(-ik\,|z - z'|)\, dz' \\ &\quad+ \bar{\sigma}_2(x_{12}, y_{12}) \int_0^z \{\hat{\Gamma}(x_1, 0, z'; x_2, 0, z')\}\, dz\ .\end{aligned} \tag{A15}$$

We note that we have set $z_1 = z_2 = z$ in Eq. (A15). We can obtain a differential equation by differentiating Eq. (A15) with respect to z. In carrying out the required manipulations we introduce the assumptions

$$\frac{\partial}{\partial z}\{\hat{\Gamma}(\mathbf{x}_{1T}, z'; \mathbf{x}_{2T}, z)\} = -ik\{\hat{\Gamma}(\mathbf{x}_{1T}, z'; \mathbf{x}_{2T}, z)\}$$

$$\frac{\partial}{\partial z}\{\hat{\Gamma}(\mathbf{x}_{1T}, z; \mathbf{x}_{2T}, z')\} = +ik\{\hat{\Gamma}(\mathbf{x}_{1T}, z; \mathbf{x}_{2T}, z')\},$$

and the assumption that $\{\hat{\Gamma}(\mathbf{x}_{1T}, z; \mathbf{x}_{2T}, z)\}$ varies little with changes of z over a distance of the order of $\bar{\lambda}$. All three assumptions will be valid so long as the radiation field can be represented by a narrow angled spectrum of plane waves in the limit $\bar{k} \rightarrow \infty$. The assumptions lead to the conclusion that the derivatives of the first two integrals on the rhs of Eq. (15) are equal to $\{\hat{\Gamma}(\mathbf{x}_{1T}, z; \mathbf{x}_{2T}, z)\}$. Thus, the differential form of the equation is written

$$\begin{aligned}\frac{d}{dz}\{\hat{\Gamma}(x_{12}, y_{12}, z)\} &= \bar{\sigma}_2(x_{12}, y_{12})\{\hat{\Gamma}(x_{12}, 0, z)\} \\ &\quad- \bar{\sigma}_2(0,0)\{\hat{\Gamma}(x_{12}, y_{12}, z)\}.\end{aligned} \tag{58}$$

This result is identical to that obtained in Sec. 2.

[*] Also, Consultants for Large Aperture Systems Branch, Acoustics Division, Naval Research Laboratories, Washington, D.C.

[1] M.J. Beran, J. Opt. Soc. Am. 56, 1475 (1966).
[2] L. Chernov, *Wave Propagation in a Random Medium* (McGraw-Hill, New York, 1960).
[3] M.J. Beran and G.B. Parrent, *Theory of Partial Coherence* (Prentice-Hall, New York, 1964).
[4] O.M. Phillips, Dynamics of the Upper Ocean (Cambridge U.P., Cambridge, 1966).
[5] W. Moseley and D. Del Balzo, "Oceanic Horizontal Random Temperature Structure," submitted for publication.
[6] I. Gradshteyn and I. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1965), p. 686, No. 4.
[7] U. Frisch, *In Probabilitic Methods in Applied Mathematics*, edited by A.T. Bharucha-Reid (Academic, New York, 1968).
[8] J.J. McCoy, J. Opt. Soc. Am. 62, 30 (1972).

The plane-wave expansion method

P. Griffin, P. Nagel*, and R. D. Koshel*

Physics Department, Ohio University, Athens, Ohio 45701

(Received 1 July 1974)

A proof of the uniform convergence of the plane-wave expansion method is given. Also, an alternative method to obtain the expansion coefficients of the plane-wave expansion is derived through the use of the Rayleigh–Ritz variational method.

I. INTRODUCTION

Recently Robson and Koshel[1] introduced the plane-wave expansion method for the purpose of simplifying the evaluation of transition matrix elements which appear in the theory of direct nuclear reactions. This method has proven to be extremely useful[2,3]; however, the mathematical properties of this expansion method have not been investigated. It is our purpose to present the results of such an investigation, in particular, we shall demonstrate the convergence properties of the expansion. We also present another method for the evaluation of the expansion coefficients and discuss some of the difficulties which arise in the numerical determination of these coefficients.

In Sec. II we shall review the plane-wave expansion method and show the connection between this method and the method of least squares. We also discuss in this section the numerical difficulties that arise when one determines the coefficients by matrix inversion. We show the connection between the plane-wave expansion method and Fourier series expansions in Sec. III. In this section we also prove the uniform convergence of the method. In Sec. IV, we present another method of evaluating the expansion coefficients. This is a variational method. In this section we also show the connection between the variational method and the boundary condition method introduced by Bloch.[4] Finally, in Sec. V we present a summary and our conclusions.

II. THE PLANE-WAVE EXPANSION METHOD

In the paper by Robson and Koshel[1] two forms of the plane-wave expansion method are introduced. These are the partial-wave decomposition and angle forms of the expansion. The discussion in this paper is limited to the partial-wave form of the series.

We can write the optical model wave functions for the relative motion of two particles as

$$\chi(\mathbf{k},\mathbf{r}) = 4\pi \sum_{lm} i^l \chi_l(kr)\, Y_{lm}(\hat{\mathbf{r}})\, Y^*_{lm}(\hat{\mathbf{k}}). \tag{1}$$

We shall neglect spin in our discussion as it only makes the equations more complicated and its inclusion does not alter the results. The assumption of the plane-wave expansion method is that we can write the radial wave function as

$$\chi_l(kr) = \sum_n a^l_n j_l(k_n r). \tag{2}$$

It is then easy to show that

$$4\pi i^l \chi_l(kr)\, Y_{lm}(\hat{\mathbf{r}}) = \sum_n a^l_n \int d\hat{\mathbf{k}}_n \exp(i\mathbf{k}_n \cdot \mathbf{r})\, Y_{lm}(\hat{\mathbf{k}}_n). \tag{3}$$

Thus, Eq. (1) becomes

$$\chi(\mathbf{k},\mathbf{r}) = \sum_{lm} Y^*_{lm}(\hat{\mathbf{k}}) \sum_n a^l_n \int d\hat{\mathbf{k}}_n \exp(i\mathbf{k}_n \cdot \mathbf{r})\, Y_{lm}(\hat{\mathbf{k}}_n). \tag{4}$$

Use of this expression in DWBA transition matrix elements considerably simplifies the calculation when finite-range and recoil effects are included. This is due to the property of the exponential function, namely, that $\exp(x+y) = e^x e^y$. This considerably simplifies the multidimensional integrals which appear in DWBA matrix elements when finite range effects are included. It is obvious that the method even allows for more simplification if the set of k_n used is the same for all l. Of course all of these conclusions depend upon the validity of the expansion given by Eq. (2). It is the purpose of this paper to explore the properties of this expansion.

If we form the quantity

$$C = \int_0^R r^2\, dr \,\Big|\, \chi_l(kr) - \sum_n a^l_n j_l(k_n r) \Big|^2 \tag{5}$$

and minimize the result with respect to the expansion coefficients a^l_n, we find that these coefficients are given by

$$\underline{a}^l = (\underline{A}^l)^{-1} \underline{B}^l, \tag{6}$$

where we have made use of matrix notation. The matrices $\underline{A}^l$ and $\underline{B}^l$ are given by

$$(\underline{A})^l_{n'n} = \int_0^R r^2\, dr\, j_l(k_{n'} r)\, j_l(k_n r)$$

and

$$\underline{B}^l_n = \int_0^R r^2\, dr\, j_l(k_n r)\, \chi_l(kr).$$

The appearance of the integration limit is easily explained. When a calculation of a DWBA matrix element is performed an upper limit on the radial integration is introduced because the form factor appearing in the expression can usually be assumed to be zero at this point. This limit is our R. Thus, we only represent the optical model wavefunction in a finite region of space.

The method by which Eq. (6) was obtained, i.e., the minimization of Eq. (5), is of course the method of least squares when an upper limit is placed on the sum. Equation (6) is identical to the expression found in Ref. 1, which was obtained through the use of nonorthogonal expansions. The use of the method of least squares for nonorthogonal expansions has also been discussed by Garside and Tobocman[5] in their work on the extended R-matrix theory of nuclear reactions. Because this method makes use of a matrix inversion we shall henceforth designate it as the matrix inversion (MI) method.

We have developed the computer program EXPAC which computes the coefficients by this method. EXPAC automatically chooses the set of k_n for the particular problem it is dealing with and proceeds to calculate the expansion coefficients for all l values. The program has been tested for a variety of cases and in all these cases it gave an excellent representation of the radial optical function. The criteria used to give a goodness of fit value is the weighted mean square χ^2.[6] This is defined by

$$\chi^2=\frac{1}{N}\sum_{i=1}^{N} \mid \psi_t(x_i) \mid^2 \mid \psi_t(x_i)-\psi_r(x_i) \mid^2, \qquad (7)$$

where $\psi_t(x_i)$ is the wavefunction we are representing by means of the expansion and $\psi_r(x_i)$ is the spherical Bessel function representation. These are evaluated at the points x_i and N is the total number of points. We found this to be a most suitable method to determine the number of states needed to give a good representation for any particular l value.

However, in some cases that were studied it was found that the matrix that is inverted, $\underline{A}^l$, became ill-conditioned. In particular, when $\Delta k \equiv k_{n+1}-k_n$ became too small many of the matrix elements became similar. The ill-conditioning also occurred if the maximum k_n value became large. As a result the program EXPAC has certain limits on it with regard to the problems for which it can give reliable results. This is strictly due to numerical processes and can be a function of the type of computer that is used. For EXPAC, which we run on either an IBM 360 or 370 computer in double precision mode this limit is set at $kR=200$, where k is the incoming wave number and R is the upper integration limit described previously. For larger values of kR special methods to deal with the ill-conditioning can be applied; however, this has not been attempted yet.

Table I shows some results for the representation of the optical model wavefunctions for the elastic scattering of protons from ^{12}C for three energies. This table shows the number of states N_l needed for representative l values and the χ^2 value. It indicates the expected feature that the number of states needed to represent the optical model wavefunction decreases as l increases, i.e., one expects the optical model wavefunction to approach a spherical Bessel function in appearance as the height of the centrifugal barrier increases. One should not make a strict comparison of the coefficients as a function of energy as different optical model parameters were used for the different energies. The optical model parameters were taken from Haybron and McManus.[7]

Extensive numerical testing of the convergence properties of the plane-wave expansion has also been performed by Robson and Charlton[3,8] for a wide range of problems. They have also found very good agreement between the representation and the optical model wavefunctions.

III. CONVERGENCE PROPERTIES

In the matrix inversion method described above we take $k_n=n\beta$, where β is some determined value. For $l=0$, we then have

$$\chi_0(kr)=\frac{u_0(kr)}{r}=\sum_n a_n^0 j_0(n\beta r). \qquad (8)$$

This is a Fourier series for u_0 if β is chosen in an appropriate manner. Thus, at least, for $l=0$ there is a connection between the plane-wave expansion method and Fourier series.

TABLE I. Representation of optical model wavefunctions.

			^{12}C$(p,p)^{12}$C					
46 MeV			90 MeV			155 MeV		
l	N_l	χ^2	l	N_l	χ^2	l	N_l	χ^2
0	17	10^{-8}	0	17	10^{-9}	0	18	10^{-9}
3	14	10^{-10}	3	13	10^{-11}	3	14	10^{-10}
6	10	10^{-10}	6	9	10^{-10}	6	10	10^{-9}
9	3	10^{-10}	9	5	10^{-10}	9	7	10^{-9}
12	3	10^{-10}	12	3	10^{-10}	12	5	10^{-9}

Watson[9] defines a series of the form

$$f(x)=\sum_n a_n J_\nu(nx). \qquad (9)$$

Convergence of this series is discussed for $|\nu| \leq \frac{1}{2}$. The Bessel functions are not orthogonal over the interval of definition. These expansions are called Schlömilch series. Because of the relationship between the ordinary Bessel functions and the spherical Bessel functions we wish to extend Watson's results for $\nu > \frac{1}{2}$ and to give the convergence properties of Schlömilch type series as given by Eq. (9) with J_ν replaced by j_l.

This is easily done if we extend the work by Pennel[10] on the use of fractional integration and differentiation to obtain certain expansion of functions in terms of ordinary Bessel functions. Many of the relationships used below are from this work and will not be separately cited.

Let us say that we have a given function $f(x)$ defined over some interval which includes the origin as one end-point. We obtain a function $\phi(x)$ from this function by use of the relationship

$$\phi(x^{1/2}) \equiv 2^n p^n x^{(n+1)/2} f(x^{1/2}). \qquad (10)$$

We use $x^{1/2}$ instead of x as the argument because it simplifies the use of fractional integration and differentiation with the Heaviside operator p. These fractional integration and differentiation operators are defined by

$$p^\nu g(x)=\frac{d^b}{dx^b}\int_0^x \frac{(x-\lambda)^{c-1}}{\Gamma(c)}\, g(\lambda)\, d\lambda, \qquad (11)$$

and

$$p^{-\nu} g(x)=\int_0^x \frac{(x-\lambda)^{\nu-1}}{\Gamma(\nu)}\, g(\lambda)\, d\lambda, \qquad (12)$$

where $\nu>0$, $0<c\leq 1$, b is a positive integer, and $\nu=b-c$. When $\nu=0$, $p^0\equiv 1$. When ν is an integer $p^\nu=d^\nu/dx^\nu$, i.e., $c=1$ so that $b=\nu+1$.

In our own work we take n, which appears in Eq. (10), to be an integer. Let us also assume that the quantity $x^{(n+1)/2}f(x^{1/2})$ is such that $\phi(x)$ exists, i.e., the function $x^{(n+1)/2}f(x^{1/2})$ is n-times differentiable. Let us also assume $\phi(x)$ is piecewise very smooth in the interval of interest, i.e., if the interval can be divided into a finite number of subintervals, ϕ its first and second derivatives are continuous within each of these subintervals. Then $\phi(x)$ can be expanded in a Fourier series in the interval and this series converges uniformly to $\phi(x)$ within any subinterval which does not contain a discontinuity.[11] The same is true for $\phi(x^{1/2})$.

Let us assume $\phi(0)=0$, so that we can expand ϕ as a Fourier sine series. Thus, we have

$$\phi(x^{1/2})=\sum_{s=1}^{\infty} A_s \sin(s\beta x^{1/2}),$$

where β is determined by the interval. We then have

$$A_s=\frac{2\beta}{\pi}\int_0^{\pi/\beta}\phi(x)\sin(s\beta x)\,dx.$$

Since the series for ϕ is uniformly convergent in the interval we can apply the operator p^{-n} to it and integrate term by term. The resulting series is also uniformly convergent since p^{-n} is essentially integration. If we make use of the results of Pennel, we have

$$f(x^{1/2})=2^{-n}x^{-[(n+1)/2]}p^{-n}\phi(x^{1/2})$$
$$=\sum_s A_s(s\beta)^{1-n}j_n(s\beta x^{1/2}),$$

or

$$f(x)=\sum_{s=1}^{\infty}A_s(s\beta)^{1-n}j_n(s\beta x).$$

This we denote as the generalized Schlömilch expansion for the function $f(x)$. The series is uniformly convergent in the interval except possibly at the endpoints.

We now take $n=l$, the orbital angular momentum quantum number and $f(x)=\chi_l(kr)$. For the type of potentials we deal with $\chi_l(kr)$ is an analytic function throughout the interval of interest. Therefore, all derivatives exist and are continuous. Therefore, ϕ exists for all l and it as well as its first and second derivatives are continuous except, perhaps, at the origin. Why the problem at the origin? If we take $l=1$, we see why. We have

$$\phi_1(r^{1/2})=2\frac{d}{dr}[(r^{1/2})^2\chi_1(r^{1/2})]$$
$$=\frac{1}{r}\frac{d}{dy}[y^2\chi_1(y)],$$

where we have $y=r^{1/2}$. This shows that there may be a singularity at $r=0$. To show that this is not the case let us investigate the behavior at the origin more closely. We have

$$\phi_l(r^{1/2})=2^l\frac{d^l}{dr^l}[r^{(l+1)/2}\chi_l(r^{1/2})].$$

For the potentials we deal with the centrifugal term will dominate at the origin and in the limit of r approaching zero we have $\chi_l(r)\rightarrow r^l$. Thus $r^{(l+1)/2}\chi_l(r^{1/2})\rightarrow r^{l+1/2}$ plus higher order terms. Thus,

$$\phi_l(r^{1/2})\rightarrow 2^l\frac{d^l}{dr^l}[r^{l+1/2}+\cdots],$$

and we see that ϕ and it as well as its first and second derivatives are continuous. We also see that $\phi_l(r=0)=0$. Thus, $\phi_l(r)$ can be expanded as a Fourier sine series as was discussed above.

Thus we have

$$\chi_l(kr)=\sum_{s=1}^{\infty}A_s(s\beta)^{1-l}j_l(s\beta r), \tag{13}$$

with

$$A_s=\frac{2\beta}{\pi}\int_0^{\pi/\beta}\phi_l(r)\sin(s\beta r)\,dr, \tag{14}$$

and

$$\phi_l(r^{1/2})=2^l p^l[r^{(l+1)/2}\chi_l(kr^{1/2})]. \tag{15}$$

TABLE II. Comparison of coefficients obtained by the matrix inversion method and Fourier series method.

	Real coefficients		Imaginary coefficients	
n	MI	FS	MI	FS
1	0.0036	0.0036	−0.0205	−0.0205
2	−0.0073	−0.0073	0.0512	0.0511
3	0.0103	0.0103	−0.0694	−0.0691
4	−0.0459	−0.0458	0.1481	0.1481
5	0.0014	0.0017	−0.6409	−0.6389
6	0.4324	0.4318	1.0203	1.0199
7	−0.1136	−0.1136	0.9057	0.9044
8	−0.0231	−0.0231	0.0669	0.0667
9	−0.0547	−0.0546	0.2237	0.2232
10	0.0043	0.0042	−0.0907	−0.0906
11	−0.0172	−0.0172	0.0841	0.0839
12	0.0132	0.0132	−0.0816	−0.0814
13	−0.0102	−0.0102	0.0658	0.0657
14	0.0103	0.0102	−0.0606	−0.0605
15	−0.0090	−0.0090	0.0554	0.0554
16	0.0082	0.0082	−0.0503	−0.0503
17	−0.0077	−0.0077	0.0467	0.0467
18	0.0072	0.0071	−0.0434	−0.0435

If we compare Eq. (13) with Eq. (2) we see that $A_s(s\beta)^{1-l}=a_s^l$. We have made a comparison between the method described just above, which we denote as the Fourier series (FS) method and the MI method described previously. A comparison between the two methods for 100 MeV protons incident on ^{12}C and for $l=1$ is shown in Table II. The agreement is extremely good.

IV. THE VARIATIONAL METHOD

We now ask the question: Can we directly solve the radial differential equation with the plane-wave expansion? If we insert the expansion given by Eq. (2) into the differential equation for the lth partial wave radial wavefunction and perform the necessary overlaps, we find that we have a homogeneous system of equations for the expansion coefficients. This, as is well known, does not lead to a unique solution. However, we have not taken into account the boundary conditions at $r=R$.

In order to introduce the boundary condition and also make use of the expansion method we shall make use of the Rayleigh–Ritz variational method.[12] This method allows us to find a solution of our differential equation for u_l in the interval 0 to R by minimizing the functional

$$J=\int_0^R\left[\left(\frac{du_l}{dr}\right)^2+u_l(r)\left(\frac{l(l+1)}{r^2}+U(r)-k^2\right)\right]dr, \tag{16}$$

where $u_l(r)=r\chi_l(r)$. Insertion of the plane-wave expansion into this equation would again give a homogeneous set of equations for the expansion coefficients. We must introduce the boundary conditions as a constraint. Let us put

$$G=u_l(R)-c=0. \tag{17}$$

We now set

$$u_l(r)=\sum_{n=1}^{N}a_n^l\phi_n^l(r),$$

where $\phi_n^l(r)=rj_l(k_nr)$. We see that the function and the series satisfies the boundary condition $u_l(0)=0$. If we insert this expansion into Eq. (16), we have

$$J=\sum_{nm} a_n^l a_m^l \left(\int_0^R \frac{d\phi_n^l}{dr}\frac{d\phi_m^l}{dr}\,dr + A_{nm}\right), \tag{18}$$

where

$$A_{nm}=\int_0^R \phi_n^l \left(\frac{l(l+1)}{r^2}+U(r)-k^2\right)\phi_m^l\,dr. \tag{19}$$

If we insert the expansion into Eq. (17), we have

$$G=\sum_n a_n^l \phi_n^l(R)-c=0. \tag{20}$$

If we now introduce the Lagrange multiplier λ, the variational problem we have to solve is

$$\frac{\partial J}{\partial a_j^l}+\lambda\frac{\partial G}{\partial a_j^l}=0. \tag{21}$$

This yields

$$\sum_{m=1}^{N} a_m^l\left(\int_0^R \frac{d\phi_j^l}{dr}\frac{d\phi_m^l}{dr}\,dr + A_{jm}\right)+\frac{\lambda}{2}\phi_j^l(R)=0. \tag{22}$$

If we take this set of N equations plus Eq. (20), we have a set of $N+1$ equations with $N+1$ unknowns, namely the N coefficients a_n^l plus λ. It is no longer a homogeneous set of equations because of the boundary condition constraint and as such the coefficients can be found by matrix inversion provided the matrix is not singular. The matrix equation we must solve is

$$\begin{pmatrix} b_{11} & b_{12}\cdots & b_{1,N+1}\\ b_{21} & b_{22}\cdots & b_{2,N+1}\\ \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot\\ b_{N+1,1} & b_{N+2,2}\cdots & 0 \end{pmatrix}\begin{pmatrix} a_1^l\\ a_2^l\\ \cdot\\ \cdot\\ \cdot\\ \lambda/2\end{pmatrix}=\begin{pmatrix}0\\0\\ \cdot\\ \cdot\\ \cdot\\ c\end{pmatrix}. \tag{23}$$

If we let c be some arbitrary constant, e.g. one, we can obtain a set of coefficients a_n^l which differ from the correct value by a constant. If we then make use of the relation

$$D\sum_{n=1}^{N} a_n^l \phi_n^l(r)=F_l(r)+C_l[G_l(r)+iF_l(r)],$$

at two points beyond the range of the nuclear force it is possible to find D, the constant coefficient needed to give the correct expansion coefficients, and C_l, the scattering amplitude. Here F_l and G_l are, respectively, the regular and irregular solutions of the differential equations without the nuclear potential.

A similar procedure was employed by Tobocman and Nagarajan[13] in their boundary condition constraint method to obtain shell model states to be used in R-matrix calculations. They, however, in addition to finding the expansion coefficients found the eigenvalues of the shell model states instead of the scattering amplitudes C_l as is done here.

We can calculate λ directly by solving Eq. (22). If we perform an integration by parts we have

$$\frac{\lambda}{2}\phi_j^l(R)=-\phi_j^l(R)\sum_m a_m^l\frac{d\phi_m^l}{dr}+\int_0^R dr\,\phi_j^l(r)$$
$$\times\left(\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}-U(r)+k^2\right)\sum_m a_m^l\phi_m^l(r).$$

If we let the last term in this expression vanish because $\sum_m a_m^l\phi_m^l(r)$ is a solution to the differential equation we have

$$\lambda=-2\sum_m a_m^l\frac{d\phi_m^l(R)}{dr}. \tag{24}$$

Normally the constraint equation for the Rayleigh–Ritz variational method is written in integral form. This can easily be done in this case. We have

$$G=\int_0^R dr\,\delta(r-R)\,u_l(r)-c=0.$$

This equation as well as the work by Lane and Robson[14] on stationary principles for scattering problems and the relationship with the Bloch operator suggests that there is a connection between the results obtained here and the boundary condition operator method introduced by Bloch.[4] We shall now show that the two methods lead to identical results and as such provides us with a check on our results.

If we follow the prescription given by Bloch and define $\mathcal{L}(r)=\delta(r-R)d/dr$ and insert it into the radial differential equation, we have

$$\left[\frac{d^2}{dr^2}-\frac{l(l+1)}{r^2}-U(r)+k^2-\mathcal{L}(r)\right]u_l(r)=-\mathcal{L}(r)\,u_l(r). \tag{25}$$

If we insert our expansion for u_l, multiply by $\phi_j^l(r)$, and integrate, we find

$$\sum_m a_m^l\left(\int_0^R \frac{d\phi_j^l}{dr}\frac{d\phi_m^l}{dr}\,dr+A_{jm}\right)=\phi_j^l(R)\sum_m a_m^l\frac{d\phi_m^l}{dr}. \tag{26}$$

This expression is identical with Eq. (22) with λ given by Eq. (24).

Calculations were performed using this method. The results obtained agreed with those obtained by the matrix inversion method discussed in Sec. II.

V. SUMMARY AND CONCLUSIONS

The present work has demonstrated the uniform convergence of the plane-wave expansion method and has also provided an alternative method for the derivation of the expansion coefficients. The proof of uniform convergence should dispel any questions on convergence and uniqueness of the expansion. This coupled with the success of calculations using this method indicate the usefulness of this method not only for studies of nuclear reactions but also for studies of atomic and molecular collisions.

Presently, work is underway which extend this approach to coupled- channel Born approximation (CCBA) and coupled-channel mass transfer (CCMT) calculations. In particular, the method described in Sec. IV seems to be extremely useful in the description of CCMT calculations. We will report on this work in the near future.

The authors wish to thank Professor D. Robson for a critical reading of an early version of this manuscript. His critical comments were much appreciated.

*Supported in part by the National Science Foundation, Grant No. NSF-GP-43798.

[1] D. Robson and R. D. Koshel, Phys. Rev. C 6, 1125 (1972).

[2] R. D. Koshel and P. Nagel, Bull. Amer. Phys. Soc. 18, 1400 (1973).

[3] L. A. Charlton, Phys. Rev. C 8, 146 (1973).

[4] C. Bloch, Nucl. Phys. 4, 503 (1957).

[5] L. Garside and W. Tobocman, Phys. Rev. 173, 1047 (1968).

[6] S. G. Mikhlin, *Variational Methods in Mathematical Physics* (Macmillan, New York, 1964), p. 48.

[7] R. B. Haybron and H. McManus, Phys. Rev. 136, B 1730 (1964).

[8] L. A. Charlton and D. Robson, Bull. Am. Phys. Soc. 17, 508 (1972).

[9] G. N. Watson, *A Treatise of the Theory of Bessel Functions* (Cambridge U. P., Cambridge, 1944), 2nd ed., p. 618.

[10] W. O. Pennel, Bull. Amer. Math. Soc. 38, 115 (1932).

[11] W. Kaplan, *Advanced Calculus* (Addison-Wesley, Reading, Mass., 1952), p. 391.

[12] E. Butkov, *Mathematical Physics* (Addison-Wesley, Reading, Mass., 1968), p. 565.

[13] W. Tobocman and M. A. Nagarajan, Phys. Rev. 138, B 1351 (1965).

[14] A. M. Lane and D. Robson, Phys. Rev. 178, 1715 (1969).

Multiplicative stochastic processes, Fokker–Planck equations, and a possible dynamical mechanism for critical behavior

Ronald Forrest Fox

School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332

(Received 21 December 1973; revised manuscript received 24 January 1974)

A derivation of the Fokker-Planck equations for additive stochastic processes is given which involves treating the continuity equation in the configuration space representation of the additive stochastic process as a multiplicative stochastic process. The average of the continuity equation becomes the Fokker–Planck equation. A presentation of the "multiplicative stochastic, Markov approximation" follows. This approximation is applied to the analysis of the dynamics of a heavy particle in a molecular fluid as described by Hamilton's equations. The nonperturbative approximation technique leads to the Fokker–Planck equation for simple Brownian motion. As part of the analysis, "intrinsic diffusion" is discovered and used to show ergodicity for the autocorrelation formula which appears during the Brownian motion calculation. An account of how these methods might be used to study the dynamical origins of critical behavior is given.

INTRODUCTION

The purpose of this paper is to illustrate connections between additive and multiplicative stochastic processes. The distinction between additive and multiplicative stochastic processes was explained in an earlier paper.[1] In several subsequent papers,[2–5] application was made of multiplicative stochastic processes to the description of quantum mechanical phenomena. Such applications may be thought of as the generalization for quantum mechanical processes of Brownian motion, an additive stochastic process, and Kubo[6] has referred to such applications as quantum mechanical Brownian motion. A general theory for additive stochastic processes has been presented[7,8] within the framework of stationary, Gaussian, Markov processes. The domain of applicability for additive stochastic processes includes Brownian motion,[9–11] irreversible thermodynamics,[12] fluctuating hydrodynamics,[13] and light scattering from simple fluids and fluid mixtures,[14–16] as well as other topics. It would appear that additive stochastic processes are of relevance for classical physics rather than for quantum physics. This separation of applicability of additive or multiplicative stochastic processes to classical or quantal physics, respectively, is not of a fundamental nature: For most purposes it is a natural separation. However, Haken[17] has reviewed the use of additive stochastic processes in quantum mechanical contexts, and this paper will exhibit the utility of multiplicative stochastic processes in classical physics.

The primary connection to be described involves a technique for the derivation of the Fokker—Planck equation for an additive stochastic process. The technique, however, uses a multiplicative stochastic process which is generated by the original additive stochastic process. In this way, the relationship between the Markov nature of the process and the properties of the second moments of the fluctuating forces is highlighted.

The technique just referred to suggests a method for the derivation of the Fokker—Planck equation for Brownian motion which connects the dissipative, friction constant of Brownian motion with interparticle potentials, in a model of Brownian motion in which the fluid is treated as N interacting fluid molecules obeying conservative classical mechanics. This problem is worked out in detail and involves an unusual type of diffusive behavior characterized by a Green's function

$$\left(2\pi\frac{K_BT}{m}t^2\right)^{-1/2}\exp\left[-\frac{(q-q')^2}{2(K_BT/m)t^2}\right]$$

in one dimension. This peculiar diffusion provides for a proof that the autocorrelation formula, for the dissipative, friction constant which is derived, is ergodic.

The conclusion of this paper considers certain infinite series which arise in the Brownian motion problem. The possibility of summing the series is raised. In addition, it is suggested that the series provides the possibility for understanding the origin of critical behavior in terms of a dynamical mechanism.

FOKKER-PLANCK EQUATIONS

The complete stochastic description of an additive stochastic process is given by the solution to its associated Fokker—Planck equation. For the case of Brownian motion, in one dimension, the additive stochastic process is given by Langevin's equation

$$M\frac{d}{dt}u(t)=-\alpha u(t)+\tilde{F}(t) \tag{1}$$

where $\alpha>0$, and $\tilde{F}(t)$ is a purely random, stationary, Gaussian force satisfying

$$\langle\tilde{F}(t)\rangle=0 \quad\text{and}\quad \langle\tilde{F}(t)\tilde{F}(s)\rangle=2\lambda\delta(t-s) \tag{2}$$

where $\langle\cdots\rangle$ denotes stochastic averaging.[7,10,11] The delta function in (2) implies the Markov property for the stochastic process $u(t)$,[11] as must be proved using the definition of a Markov process, which definition involves specific behavior by the higher than second-order correlation distributions for $u(t)$.[11] The Markov property is then used to imply the validity of the Smoluchowski equation, which for Eq. (1) is written as[11]

$$P(u(0)\,|\,u,t+\Delta t)=\int P(u(0)\,|\,u',t)P(u'\,|\,u,\Delta t)\,du' \tag{3}$$

where $P(u(0)\,|\,u,t)\,du$ is the conditional probability that the value of $u(t)$ and time t will be between u and $u+du$ given that at time $t=0$, $u(t)$ had the value $u(0)$. The Smoluchowski equation is used to derive the Fokker—Planck equation which in this example becomes

$$\frac{\partial}{\partial t}P(u(0)\,|\,u,t) = -\frac{\partial}{\partial u}\left(-\frac{\alpha}{M}uP(u(0)\,|\,u,t)\right) + \frac{\lambda}{M^2}\frac{\partial^2}{\partial u^2}P(u(0)\,|\,u,t) \tag{4}$$

The solution to Eq. (4) is subject to the initial condition at $t=0$ that $P(u(0)\,|\,u,0)=\delta(u-u(0))$. The solution to the Fokker–Planck equation provides the conditional probability distribution which for a Markov process contains all possible information about the time evolution of the process. The solution to (4) is

$$P(u(0)\,|\,u,t) = \left(2\pi\frac{K_BT}{M}(1-\rho^2(t))\right)^{-1/2}\exp-\left(\frac{M(u-\rho(t)u(0))^2}{2K_BT(1-\rho^2(t))}\right) \tag{5}$$

where $\rho(t)\equiv\exp[-(\alpha/M)t]$, and we have used the relation

$$\lambda = K_BT\alpha \tag{6}$$

which is Einstein's relation within the context of Langevin's equation. Relation (6) follows from the requirement that the solution to (4) satisfy the asymptotic limit

$$\lim_{t\to\infty}P(u(0)\,|\,u,t) = \left(2\pi\frac{K_BT}{M}\right)^{-1/2}\exp-\left(\frac{Mu^2}{2K_BT}\right) \tag{7}$$

where the right-hand side of (7) is the Maxwell distribution. In Einstein's treatment of Brownian motion,[18] the analog of (6) is derived

$$D = K_BT/\alpha \tag{8}$$

where D is the diffusion constant. Both relations (6) and (8) are referred to as Einstein's relation in the literature. Relation (8) is the original relation which follows from the diffusion equation, whereas relation (6) is the prototype for fluctuation–dissipation theorems.[7]

Multicomponent generalizations of Langevin's equation lead to the general theory of stationary, Gaussian, Markov processes[7,19] which are described by the equations

$$\frac{d}{dt}a_\alpha(t) = A_{\alpha\alpha'}a_{\alpha'}(t) + S_{\alpha\alpha'}a_{\alpha'}(t) + \tilde{F}_\alpha(t) \tag{9}$$

where $\alpha=1,2,\ldots,N$, $A_{\alpha\alpha'}$ is an antisymmetric matrix, $S_{\alpha\alpha'}$ is a symmetric matrix with nonpositive eigenvalues, and $\tilde{F}_\alpha(t)$ is an n-component, purely random, Gaussian fluctuating force satisfying

$$\langle\tilde{F}_\alpha(t)\rangle = 0 \quad \text{for each } \alpha, \tag{10}$$

$$\langle\tilde{F}_\alpha(t)\tilde{F}_\beta(s)\rangle = 2Q_{\alpha\beta}\delta(t-s), \tag{11}$$

where $Q_{\alpha\beta}$ is a symmetric matrix with nonnegative eigenvalues. The corresponding Fokker–Planck equation is given by

$$\frac{\partial}{\partial t}P(\mathbf{a}(0)\,|\,\mathbf{a},t) = -\frac{\partial}{\partial a_\alpha}[(A_{\alpha\alpha'}a_{\alpha'}+S_{\alpha\alpha'}a_{\alpha'})P(\mathbf{a}(0)\,|\,\mathbf{a},t)] + \frac{\partial^2}{\partial a_\alpha\,\partial a_\beta}[Q_{\alpha\beta}P(\mathbf{a}(0)\,|\,\mathbf{a},t)] \tag{12}$$

with the initial condition $P(\mathbf{a}(0)\,|\,\mathbf{a},0)=\delta(\mathbf{a}-\mathbf{a}(0))$. The solution to (12) is given by[7,19]

$$P(\mathbf{a}(0)\,|\,\mathbf{a},t) = \left[\frac{\|\mathbf{C}(t)\|}{(2\pi)^n}\right]^{1/2}\exp[-\tfrac{1}{2}(a_\alpha - D_{\alpha\alpha'}(t)a_{\alpha'}(0))C_{\alpha\beta}(t) \times (a_\beta - D_{\beta\beta'}(t)a_{\beta'}(0))] \tag{13}$$

where $\|\mathbf{C}(t)\|$ denotes the determinant of $\mathbf{C}(t)$, and $D_{\alpha\alpha'}(t)$ is defined by

$$\mathbf{D}(t)\equiv\exp[(\mathbf{A}+\mathbf{S})t] \tag{14}$$

and $C_{\alpha\beta}(t)$ is defined by

$$\mathbf{C}^{-1}(t)\equiv\mathbf{E}^{-1}-\mathbf{D}(t)\mathbf{E}^{-1}\mathbf{D}^{+}(t) \tag{15}$$

in which the matrix $\mathbf{E}$ is determined by the entropy $S(t)$, given by

$$S(t) = S_0 - \tfrac{1}{2}K_Ba_\alpha(t)E_{\alpha\beta}a_\beta(t), \tag{16}$$

so that $\mathbf{E}$ is symmetric with positive eigenvalues. To get (13) from (12) we have used the generalized fluctuation–dissipation theorem[7,19]

$$2Q_{\alpha\beta} = -(A_{\alpha\theta}+S_{\alpha\theta})E^{-1}_{\theta\beta} + E^{-1}_{\alpha\theta}(A_{\theta\beta}-S_{\theta\beta}) \tag{17}$$

which follows from the requirement that

$$\lim_{t\to\infty}P(\mathbf{a}(0)\,|\,\mathbf{a}t) = W_0\exp(-\tfrac{1}{2}a_\alpha E_{\alpha\beta}a_\beta) \tag{18}$$

where $W_0=[\|\mathbf{E}\|/(2\pi)^n]^{1/2}$, and the right-hand side of (18) is seen to be the Boltzmann–Planck formula because the entropy is given by (16). The derivation of (12) requires proving that (9) with (11) describes a Markov process. Then a generalization to n components of Smoluchowski's equation leads to (12).[11]

A further generalization is possible in which one deals with continuous, variable indices. Analogs to (9), (11), and (17) result in

$$\frac{\partial}{\partial t}a_i(\mathbf{r},t) = -\int A_{ij}(\mathbf{r},\mathbf{r}')a_j(\mathbf{r}',t)\,d\mathbf{r}' - \int S_{ij}(\mathbf{r},\mathbf{r}')a_j(\mathbf{r}',t)\,d\mathbf{r}' + \tilde{F}_i(\mathbf{r},t), \tag{19}$$

$$\langle\tilde{F}_i(\mathbf{r},t)\tilde{F}_j(\mathbf{r}',s)\rangle = 2Q_{ij}(\mathbf{r},\mathbf{r}')\delta(t-s), \tag{20}$$

$$2Q_{ij}(\mathbf{r},\mathbf{r}') = \int[(A_{il}(\mathbf{r},\mathbf{r}'')+S_{il}(\mathbf{r},\mathbf{r}''))E^{-1}_{lj}(\mathbf{r}'',\mathbf{r}') + E^{-1}_{il}(\mathbf{r},\mathbf{r}'')(-A_{lj}(\mathbf{r}'',\mathbf{r}')+S_{lj}(\mathbf{r}'',\mathbf{r}'))]\,d\mathbf{r}'' \tag{21}$$

where $E_{lK}(\mathbf{r},\mathbf{r}')$ is the entropy matrix analog of (16)

$$S(t) = S_0 - \tfrac{1}{2}K_B\int\int a_i(\mathbf{r},t)E_{ij}(\mathbf{r},\mathbf{r}')a_j(\mathbf{r}'t)\,d\mathbf{r}\,d\mathbf{r}'. \tag{22}$$

This form of the general theory of stationary, Gaussian, Markov processes leads to hydrodynamical applications,[7,14–16] as well as to applications in other field dependent problems.[8,20] There is also in this case a corresponding Fokker–Planck equation which will involve utilization of the techniques of functional differentiation. We shall not write down this equation.

At this point a partial review of the structure and significance of the Fokker–Planck equation for additive stochastic processes has been presented. It is necessary, before proceeding further with Fokker–Planck equations, to review the basic theorem in the theory of multiplicative stochastic processes.

Consider an equation of the form

$$\frac{d}{dt}a_\alpha(t) = A_{\alpha\alpha'}a_{\alpha'}(t) + \tilde{A}_{\alpha\alpha'}(t)a_{\alpha'}(t) \tag{23}$$

where $\alpha=1,2,\ldots,N$, $A_{\alpha\alpha'}=-A_{\alpha'\alpha}$, $\tilde{A}_{\alpha\alpha'}(t)=-\tilde{A}_{\alpha'\alpha}(t)$, and the matrix $\tilde{A}_{\alpha\alpha'}(t)$ is a purely random, Gaussian matrix with mean value zero, which implies that[1]

$$\langle \tilde{A}_{\alpha\alpha'}(t)\rangle = 0 \quad \text{and} \quad \langle \tilde{A}_{\alpha\alpha'}(t)\tilde{A}_{\beta\beta'}(s)\rangle = 2Q_{\alpha\alpha'\beta\beta'}\delta(t-s) \tag{24}$$

and that the higher-order averaged moments are given by

$$\langle \tilde{A}_{\alpha_1\alpha_1'}(t_1)\cdots A_{\alpha_{2n-1}\alpha_{2n-1}'}(t_{2n-1})\rangle = 0, \tag{25}$$

$$\langle A_{\alpha_1\alpha_1'}(t_1)\cdots A_{\alpha_{2n}\alpha_{2n}'}(t_{2n})\rangle = \frac{1}{2^n n!}\sum_{p\in S_{2n}}\prod_{j=1}^{n}\langle A_{\alpha_{p(2j-1)}\alpha'_{p(2j-1)}}(t_{p(2j-1)}A_{\alpha_{p(2j)}\alpha'_{p(2j)}}(t_{p(2j)})\rangle \tag{26}$$

where $\sum_{p\in S_{2n}}$ denotes the sum over all permutations of the symmetric group of order $(2n)!$. Equations (25) and (26) are consequences of the Gaussian property of $\tilde{A}(t)$. It has been proved that these properties of $\tilde{A}(t)$ lead to the averaged equation[1]

$$\frac{d}{dt}\langle a_\alpha(t)\rangle = A_{\alpha\alpha'}\langle a_{\alpha'}(t)\rangle + Q_{\alpha\theta\theta\alpha'}\langle a_{\alpha'}(t)\rangle \tag{27}$$

in which the matrix $Q_{\alpha\theta\theta\alpha'}$ is determined by (24) which implies that it is symmetric with nonpositive eigenvalues. Whereas (23) describes fluctuating oscillations, (27) describes damped oscillations.

The proof of (27) from (23) using the properties of (24), (25), and (26) goes through unchanged if the $A_{\alpha\alpha'}$ in (23) is augmented by the addition of a symmetric matrix $S_{\alpha\alpha'}$ which has nonpositive eigenvalues. That is,

$$\frac{d}{dt}a_\alpha(t) = A_{\alpha\alpha'}a_{\alpha'}(t) + S_{\alpha\alpha'}a_{\alpha'}(t) + \tilde{A}_{\alpha\alpha'}(t)a_{\alpha'}(t) \tag{28}$$

leads to

$$\frac{d}{dt}\langle a_\alpha(t)\rangle = A_{\alpha\alpha'}\langle a_{\alpha'}(t)\rangle + S_{\alpha\alpha'}\langle a_{\alpha'}(t)\rangle + Q_{\alpha\theta\theta\alpha'}\langle a_{\alpha'}(t)\rangle \tag{29}$$

upon performing the average of (28).

In addition, this result can be generalized to infinite matrices and to continuous indices. Therefore, we shall also use the theorem in the form

$$\frac{\partial}{\partial t}f(q,t) = L(q)f(q,t) + \tilde{L}(q,t)f(q,t) \tag{30}$$

leads to

$$\frac{\partial}{\partial t}\langle f(q,t)\rangle = L(q)\langle f(q,t)\rangle + D(q)\langle f(q,t)\rangle \tag{31}$$

where $L(q)$ is a linear differential or integral operator, $\tilde{L}(q,t)$ is a purely random, Gaussian, fluctuating linear differential or integral operator with mean value zero, and $D(q)$ is a linear differential or integral operator given by

$$\langle \tilde{L}(q,t)\tilde{L}(q,t')\rangle = 2D(q)\delta(t-t'). \tag{32}$$

In (30), it will be assumed that $\tilde{L}(q,t)$ is an antisymmetric operator, which means that if $\{\varphi_I(q)\}$ is a complete set of real orthonormal functions of the coordinates q, then

$$[\tilde{L}(t)]_{IK} = -[\tilde{L}(t)]_{KI} \tag{33}$$

where $[\tilde{L}(t)]_{IK} \equiv \int \varphi_I(q)\tilde{L}(q,t)\varphi_K(q)\,dq$. Consequently, $D(q)$ is a symmetric operator, which means that

$$[D]_{IK} = [D]_{KI} \tag{34}$$

where $[D]_{IK} \equiv \int \varphi_I(q)D(q)\varphi_K(q)\,dq$ because

$$[D]_{IK} = \int_0^t \langle [L(t)]_{Ij}[L(s)]_{jK}\rangle\,ds. \tag{35}$$

From (35) it is also seen that the eigenspectrum of $D(q)$ is entirely nonpositive. The operator $L(q)$ in (30) may be either antisymmetric or of mixed symmetry in analogy with (23) and (28), respectively. Specific examples of these formal expressions will occur in the following paragraphs.

The q in (30) and (31) may represent a single variable, or it may represent a set of variables. Suppose we explicitly have N q's: $q_1, q_2, \ldots, q_N$. Consider the configuration space corresponding with these q_j's and let $f(q_1q_2\cdots q_Nt)$ be the distribution function in configuration space. If $N=2m$ and m of the q_j's are coordinates and m of the q_j's are momenta, then the associated configuration space becomes phase space and f becomes the phase space distribution. However, in several contexts it is not phase space which will interest us and, therefore, we use the notation q_j's for a generalized configuration space of N variables.

It is always the case that the continuity equation is true in configuration space. In our notation this becomes

$$\frac{\partial}{\partial t}f(q_1\cdots q_Nt) = -\frac{\partial}{\partial q_j}(q_jf(\dot{q}_jf(q_1\cdots q_Nt)) \tag{36}$$

where $\dot{q}_j \equiv d/dt\,q_j(t)$. If we are actually dealing with phase space and a conservative system for which Hamilton's equations are valid, then (36) leads to Liouville's equation.[21] However, in general we will not necessarily be in phase space or be dealing with a conservative system, so that Liouville's theorem is not generally valid although (36) is. In the general case, some relationship between the $\dot{q}_j$'s and the q_j's must hold if (36) is to lead to anything useful.

Returning to (9), we will take for the q_j's the a_α's. Therefore, (36) is seen to be given by

$$\frac{\partial}{\partial t}f(a_1\cdots a_Nt) = -\frac{\partial}{\partial a_\alpha}(\dot{a}_\alpha f(a_1\cdots a_Nt)) \tag{37}$$

For the a_α's we use (9) which converts (37) into

$$\frac{\partial}{\partial t}f(a_1\cdots a_Nt) = -\frac{\partial}{\partial a_\alpha}[(A_{\alpha\alpha'}a_{\alpha'} + S_{\alpha\alpha'}a_{\alpha'} + \tilde{F}_\alpha(t))f(a_1\cdots a_Nt)] \tag{38}$$

which is clearly a multiplicative stochastic process in the form of (30) if we identify

$$L(q) \to -\frac{\partial}{\partial a_\alpha}(A_{\alpha\alpha'}a_{\alpha'} + S_{\alpha\alpha'}a_{\alpha'})x$$

and (39)

$$\tilde{L}(q,t) \to -\frac{\partial}{\partial a_\alpha}\tilde{F}_\alpha(t)x = -\tilde{F}_\alpha(t)\frac{\partial}{\partial a_\alpha}.$$

The analog of (31) is then

$$\frac{\partial}{\partial t}\langle f(a_1\cdots a_Nt)\rangle = -\frac{\partial}{\partial a_\alpha}[A_{\alpha\alpha'}a_{\alpha'} + S_{\alpha\alpha'}a_{\alpha'})\langle f(a_1\cdots a_Nt)\rangle]$$

$$+\frac{\partial^2}{\partial a_\alpha \partial a_\beta}[Q_{\alpha\beta}\langle f(a_1 \cdots a_N t)\rangle]. \tag{40}$$

This is precisely the Fokker—Planck equation (12) if we identify

$$P(\mathbf{a}(0)\,|\,\mathbf{a}, t) \to \langle f(a_1 \cdots a_N t)\rangle. \tag{41}$$

Consequently, given an additive stochastic process, we can write out its associated continuity equation in its configuration space and we arrive at a multiplicative stochastic process, the average of which is the Fokker—Planck equation of the original additive stochastic process.

Because the connection between (38) and (40) requires that $\tilde{F}_\alpha(t)$ is purely random, we see the Markov property and its connection with the Fokker—Planck equation without having to proceed via the Smoluchowski equation. As a special case of (38) and (40), we momentarily return to (1) from which the analogs of (37) and (38) follow:

$$\frac{\partial}{\partial t}f(u,t) = -\frac{\partial}{\partial u}(\dot{u}f(u,t)), \tag{42}$$

$$\frac{\partial}{\partial t}f(u,t) = -\frac{\partial}{\partial u}\left[\left(-\frac{\alpha}{M}u + \frac{\tilde{F}(t)}{M}\right)f(u,t)\right].$$

The average of (42) is

$$\frac{\partial}{\partial t}\langle f(u,t)\rangle = \frac{\partial}{\partial u}\left(\frac{\alpha}{M}u\langle f(u,t)\rangle\right) + \frac{\lambda}{M^2}\frac{\partial^2}{\partial u^2}\langle f(u,t)\rangle \tag{43}$$

which is identical with the Fokker—Planck equation (4).

THE MULTIPLICATIVE STOCHASTIC, MARKOV APPROXIMATION

In this section we will discuss an approximation procedure which will be referred to as the multiplicative stochastic, Markov approximation. Its connection with the preceeding section will be illustrated, and it will provide the background necessary for the analysis of Brownian motion which follows in the next section. The approximation procedure introduced here is nonperturbative.

Suppose we have a particle in a fluid. It is, on the average, at rest, although it does execute a fluctuating motion. We could describe this motion of the particle phenomenologically by the equation

$$M\frac{dx}{dt} = \tilde{p}(t) \tag{44}$$

in which the fluctuating momentum $\tilde{p}(t)$ is assumed to be a purely random, Gaussian stochastic process with average value zero and a second moment given by

$$\langle \tilde{p}(t)\tilde{p}(s)\rangle = 2M^2 D\delta(t-s) \tag{45}$$

in which M is the mass, as in (44), and D is a constant. The continuity equation in the configuration space description of (44) is

$$\begin{aligned}\frac{\partial}{\partial t}f(x,t) &= -\frac{\partial}{\partial x}(\dot{x}f(x,t)) \\ &= -\frac{\partial}{\partial x}\frac{\tilde{p}(t)}{M}f(x,t) \\ &= -\frac{\tilde{p}(t)}{M}\frac{\partial}{\partial x}f(x,t).\end{aligned} \tag{46}$$

This is again a multiplicative stochastic process with a purely random, Gaussian stochastic operator. The average equation is, therefore,

$$\frac{\partial}{\partial t}\langle f(x,t)\rangle = D\frac{\partial^2}{\partial x^2}\langle f(x,t)\rangle \tag{47}$$

which is recognized as the diffusion equation with diffusion constant D. Equations (46) and (47) are special cases of (37), (38), and (40).

It may be objected that from Brownian motion we know that (45) is not so, but that instead

$$(\langle \tilde{p}(t)\tilde{p}(s)\rangle) = MK_BT\exp\left(-\frac{\alpha}{M}|t-s|\right) \tag{48}$$

which follows from (5), and in which the rounded $(\cdots)$ denote an average over the initial value distribution in addition to the stochastic average denote by $\langle\cdots\rangle$. The second average makes (48) a stationary expression depending upon $|t-s|$.[10] Using (48) in (46) leads again to

$$\frac{\partial}{\partial t}f(x,t) = -\frac{\tilde{p}(t)}{M}\frac{\partial}{\partial x}f(x,t) \tag{49}$$

which is still a multiplicative stochastic process, but the stochastic operator is now no longer purely random, so that we cannot use our theorem for averaged multiplicative stochastic processes. However, the $\tilde{p}(t)$ in (48) and (49) is still a Gaussian process because the $u(t)$ in (1) inherits the Gaussian property from the $\tilde{F}(t)$ in (1) since (1) is a linear equation. In the Appendix it is shown that the Gaussianness of $\tilde{p}(t)$ leads to the exact result

$$\frac{\partial}{\partial t}(\langle f(xt)\rangle) = \int_0^t \frac{1}{M^2}(\langle \tilde{p}(t)\tilde{p}(s)\rangle)\,ds\frac{\partial^2}{\partial x^2}(\langle f(xt)\rangle). \tag{50}$$

Using (48) in (50) permits performance of the integration giving

$$\frac{\partial}{\partial t}(\langle f(xt)\rangle) = \frac{K_BT}{\alpha}\left[1-\exp\left(-\frac{\alpha}{M}t\right)\right]\frac{\partial^2}{\partial x^2}(\langle f(xt)\rangle) \tag{51}$$

which differs from (47) by the presence of a time dependent diffusion constant: $(K_BT/\alpha)\{1-\exp[-(\alpha/M)t]\}$. For times long compared with M/α we can neglect the exponential term, and this constitutes the multiplicative stochastic, Markov approximation. The integral in (50) can be used to define a diffusion constant when we neglect the exponential part in (51). We get

$$D' = \lim_{t\to\infty}\frac{1}{M^2}\int_0^t (\langle \tilde{p}(t)\tilde{p}(s)\rangle)\,ds \tag{52}$$

where D' will be used in an expression like (45). Because of the infinite decay tail in (48), we have taken the limit $t\to\infty$ in (52), although the greatest part of the integral comes early. Using (48) in (52) gives

$$D' = K_BT/\alpha. \tag{53}$$

Note that (52) also gives the strength of the second moment for $M^{-1}\tilde{p}(t)$ as expressed by (45) since

$$D = \lim_{t\to\infty}\frac{1}{M^2}\int_0^t \langle \tilde{p}(t)\tilde{p}(s)\rangle\,ds. \tag{54}$$

Therefore, if $D \equiv D'$ we have the multiplicative stochastic, Markov approximation for (48) and (49) given by (45) and (46). The average equation (47) or (51) is the diffusion equation, and the connection between diffusion and Brownian motion given by

$$D = K_B T/\alpha \tag{55}$$

is Einstein's original formula (8).[18] The self-consistency of the multiplicative stochastic, Markov approximation requires that the D' we get from (52) not be too large so that diffusion is slow compared with the relaxation of (48) which is governed by the magnitude of α. Indeed, (48) decays faster for larger α which via (53) results in a smaller D' which implies a slower diffusion. Therefore, in this example, the multiplicative stochastic, Markov approximation is seen to be intrinsically self-consistent. The reciprocal relation between D' and α generalizes to the more general case of multicomponent processes.

In the next section, analogs of (52) will arise in which a stochastic quantity has a non-purely-random second moment correlation formula from which we are calculating its strength. A formula like (52) will be used to calculate the strength to be used in the replacement correlation formula which is purely random. This will be analogous to the replacement of (48) by (45) where the D in (45) satisfies (55). The presence of the infinite time limit is formal, and it will be shown that the correlation integrand actually decays significantly in a very short time interval, as was the case with (52) with (48) in the integrand.

Before proceeding further, it is worth remarking that while we can actually solve (49), given (48), by writing (50), because of the Gaussian property, it is not a generally valid procedure for situations in which the stochastic operator does not commute with itself at different times.[22] Here, commutivity is guaranteed by the simple form of (49). When there is noncommutivity, then the Gaussian property alone is not sufficient for the reduction of the averaged equation to a workable form, and the multiplicative stochastic, Markov approximation becomes essential.[22]

MICROSCOPIC MODEL OF BROWNIAN MOTION

In this section we shall bring to bear the techniques of the preceeding sections as we attempt a derivation of the Fokker–Planck equation (4), for a Brownian particle, starting from a description involving a heavy particle interacting with N fluid molecules according to Hamilton's equations of motion. This is not a new program as far as its objective is concerned. Others, using other contexts and techniques, have also made this objective their goal.[23-27] Our present context and techniques were suggested by Kubo[26] in his remarkable pioneering work on multiplicative stochastic processes. The results in this paper differ somewhat with Kubo's results because he used some simplifying assumptions which we have found unnecessary. It will be seen that the analysis presented here goes beyond that of any of the other references cited in terms of the detailed account of what is happening dynamically. Two main consequences accrue: (1) A new kind of "intrinsic diffusion" is discovered which enables us to show ergodicity for the correlation formulae which appear, and (2) a connection with the dynamical origins of critical phenomena is seen. The first point will be emphasized as it comes up later in this section, while the second point will be discussed in the section following the next section.

Our starting point is the Hamiltonian

$$H = \frac{|\mathbf{P}|^2}{2M} + \sum_{j=1}^{N} \frac{|\mathbf{p}_j|^2}{2m} + \sum_{j=1}^{N} \varphi(\mathbf{R}, \mathbf{r}_j) + \frac{1}{2} \sum_{j \neq k}^{N} U(\mathbf{r}_j, \mathbf{r}_k) \tag{56}$$

where the heavy particle has mass M, position $\mathbf{R}$, and momentum $\mathbf{P}$, and the fluid molecules have mass m, position $\mathbf{r}_j$, and momenta $\mathbf{p}_j$ where $j = 1, 2, \ldots, N$. $\varphi(\mathbf{R}, \mathbf{r}_j)$ is the interaction potential between fluid molecule j and the heavy particle, and $U(\mathbf{r}_j, \mathbf{r}_k)$ is the interaction potential for the fluid molecules. Because our description is that of a conservative system we may use Liouville's theorem to express the continuity equation corresponding with the phase space picture for the system:

$$\frac{\partial}{\partial t} f = -i\mathbf{L} f \tag{57}$$

where $f \equiv f(\mathbf{R}\mathbf{P}\mathbf{r}_1\mathbf{p}_1 \cdots \mathbf{r}_N\mathbf{p}_N t)$ and $-i\mathbf{L}$ is defined by

$$-i\mathbf{L} \equiv -\frac{\mathbf{P}}{M} \cdot \nabla_{\mathbf{R}} - \sum_{j=1}^{N} \frac{\mathbf{p}_j}{m} \cdot \nabla_{\mathbf{r}_j} + \sum_{j=1}^{N} \nabla_{\mathbf{R}} \varphi(\mathbf{R}, \mathbf{r}_j) \cdot (\nabla_{\mathbf{P}} - \nabla_{\mathbf{p}_j})$$
$$+ \sum_{j \neq k} \nabla_{\mathbf{r}_j} U(\mathbf{r}_j, \mathbf{r}_k) \cdot \nabla_{\mathbf{p}_j}. \tag{58}$$

The Liouville operator defined by (58) will be separated into two parts:

$$-i\mathbf{L}_B \equiv -\frac{\mathbf{P}}{M} \cdot \nabla_{\mathbf{R}} + \sum_{j=1}^{N} \nabla_{\mathbf{R}} \varphi(\mathbf{R}, \mathbf{r}_j) \cdot \nabla_{\mathbf{P}}, \tag{59}$$

$$-i\mathbf{L}_R \equiv -\sum_{j=1}^{N} \frac{\mathbf{p}_j}{m} \cdot \nabla_{\mathbf{r}_j} + \sum_{j=1}^{N} \nabla_{\mathbf{r}_j} \varphi(\mathbf{R}, \mathbf{r}_j) \cdot \nabla_{\mathbf{p}_j}$$
$$+ \sum_{j \neq k}^{N} \nabla_{\mathbf{r}_j} U(\mathbf{r}_i, \mathbf{r}_k) \cdot \nabla_{\mathbf{p}_j}. \tag{60}$$

Now, define $\hat{f}$ by

$$f \equiv \exp(-it\mathbf{L}_R)\hat{f} \tag{61}$$

where $\hat{f} \equiv \hat{f}(\mathbf{R}\mathbf{P}\mathbf{r}_1\mathbf{p}_1 \cdots \mathbf{r}_N\mathbf{p}_N t)$. Using (57) through (61) leads to

$$\frac{\partial}{\partial t} \hat{f} = -i\tilde{\mathbf{L}}_B(t)\hat{f} \tag{62}$$

where $\tilde{\mathbf{L}}_B(t)$ is defined by

$$-i\tilde{\mathbf{L}}_B(t) \equiv \exp(it\mathbf{L}_R)(-i\mathbf{L}_B)\exp(-it\mathbf{L}_R). \tag{63}$$

Our notation suggests that $\tilde{\mathbf{L}}_B(t)$ is a stochastic operator. Of course, it is clearly not a stochastic operator as is explicitly evident if (59) and (60) are inserted into (63). However, it acts like a stochastic operator because the noncommutivity of $\mathbf{L}_B$ and $\mathbf{L}_R$ results in extremely rapid variations in $\tilde{\mathbf{L}}_B(t)$ if N is sufficiently large. Moreover, (59) and (60) may be used to exhibit $\mathbf{L}_B(t)$ as a sum of N similar terms. This suggests that to treat $\tilde{\mathbf{L}}_B(t)$ as Gaussian for large N is not unreasonable. Therefore, we shall invoke the nonperturbative, multiplicative stochastic, Markov approximation while we treat $\tilde{\mathbf{L}}_B(t)$ as stochastic.

It shall be assumed that the average of $\hat{f}$, $\langle\hat{f}\rangle$, factors for all time as follows:

$$\langle\hat{f}\rangle = g_B(Pt) W_R^{\mathrm{eq}}(\mathbf{R}\mathbf{r}_1\mathbf{p}_1 \cdots \mathbf{r}_N\mathbf{p}_N) \tag{64}$$

where $W_R^{\mathrm{eq}}(\mathbf{R}_1\mathbf{r}_1\mathbf{p}_1 \cdots \mathbf{r}_N\mathbf{p}_N)$ is the canonical equilibrium distribution for the fluid molecules in the presence of the heavy particle at **R**, and is given by

$$W_R^{\text{eq}}(\mathbf{R}\mathbf{r}_1\mathbf{p}_1\cdots\mathbf{r}_N\mathbf{p}_N)\equiv\frac{1}{Q_N}\exp\left[-\beta\left(\sum_{j=1}^{N}\frac{|\mathbf{p}_j|^2}{2m}+\sum_{j=1}^{N}\varphi(\mathbf{R},\mathbf{r}_j)+\frac{1}{2}\sum_{j\neq k}^{N}U(\mathbf{r}_j,\mathbf{r}_k)\right)\right] \tag{65}$$

where Q_N the normalization constant which satisfies

$$\int\cdots\int W_R^{\text{eq}}(\mathbf{R}\mathbf{r}_1\mathbf{p}_1\cdots\mathbf{r}_N\mathbf{p}_N)\,d\mathbf{R}\,d\mathbf{r}_1\cdots d\mathbf{r}_N\,d\mathbf{p}_1\cdots d\mathbf{p}_N=1. \tag{66}$$

The heavy particle coordinates **R** appear in (65) because the inertia of the heavy particle is so large that the fluid molecules achieve thermal equilibrium relative to the instantaneous position **R** of the heavy particle very rapidly compared with the rate of change of the position **R**. This situation is analogous to the technique used to derive the Langevin equation starting with a heavy particle in a fluctuating fluid, in which case the fluid fluctuations must be computed subject to boundary conditions representing the presence of the heavy particle.[7,28,29] The response of the fluid to the presence of the heavy particle must appear in the computation if sensible results are desired.

Returning to (62) and using (64) we get, on the average,

$$\frac{\partial}{\partial t}g_B(\mathbf{P}t)=-\int_0^\infty\int\cdots N\cdots\int\int\cdots N+1\cdots\int\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)\times W_R^{\text{eq}}(\mathbf{R}\mathbf{r}^N\mathbf{p}^N)\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\,ds\,g_B(\mathbf{P}t), \tag{67}$$

where we have integrated over the fluid variables and **R**, and the quantity

$$-\int_0^\infty\int\cdots N\cdots\int\int\cdots N+1\cdots\int\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)\,W_R^{\text{eq}}(\mathbf{R}\mathbf{r}^N\mathbf{p}^N)\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\,ds \tag{68}$$

is almost the analog of the correlation strength given in (52). The minus sign comes from $(i)^2$ and $\mathbf{r}^N$ denotes $\mathbf{r}_1\cdots\mathbf{r}_N$ whereas $d^N\mathbf{r}$ denotes $d\mathbf{r}_1\cdots d\mathbf{r}_N$, etc. The integration over the variables in W_R^{eq} corresponds with the round brackets average in (48) and provides the analog of stationarity. Note that (68) is not exactly the analog of (52) because it is still an operator. Equation (67) is what would result if $\tilde{\mathbf{L}}_B(t)$ were really a stochastic, Gaussian process for which the Markov property holds. We have implicitly assumed that the average of $\tilde{\mathbf{L}}_B(t)$ is zero in writing (67). We shall digress for a moment and explicitly demonstrate that $\tilde{\mathbf{L}}_B(t)$ has a zero average, and that (68) is stationary. The average of $\tilde{\mathbf{L}}_B(t)$ is given by

$$\begin{aligned}\langle\langle\tilde{\mathbf{L}}_B(t)\rangle\rangle&\equiv\int\cdots\int\tilde{\mathbf{L}}_B(t)W_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\\&=\int\cdots\int\exp(it\mathbf{L}_R)\mathbf{L}_B\exp(-it\mathbf{L}_R)W_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}.\end{aligned} \tag{69}$$

The operator given by (69) acts upon functions of **P**. Using (60) and (65) it follows that

$$\exp(-it\mathbf{L}_R)(W_R^{\text{eq}}\psi(\mathbf{P}))=W_R^{\text{eq}}\exp(-it\mathbf{L}_R)(\psi(\mathbf{P}))=W_R^{\text{eq}}\psi(\mathbf{P}) \tag{70}$$

where $\psi(\mathbf{P})$ is an arbitrary function of **P**, because $\mathbf{L}_R(W_R^{\text{eq}})=0$ and $\mathbf{L}_R(W_R^{\text{eq}}\psi(\mathbf{P}))=\mathbf{L}_R(W_R^{\text{eq}})\psi(\mathbf{P})+W_R^{\text{eq}}\mathbf{L}_R(\psi(\mathbf{P}))$, and $\mathbf{L}_R(\psi(\mathbf{P}))=0$ since $\mathbf{L}_R$ does not involve **P**. Therefore, we have

$$\langle\langle\mathbf{L}_B(t)\rangle\rangle=\int\cdots\int\exp(it\mathbf{L}_R)\mathbf{L}_BW_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}. \tag{71}$$

Expanding $\exp(it\mathbf{L}_R)$ in a power series and integrating each term by parts leads to

$$\langle\langle\tilde{\mathbf{L}}_B(t)\rangle\rangle=\int\cdots\int\mathbf{L}_BW_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p} \tag{72}$$

since all but the first term in the power series give zero because W_R^{eq} vanishes at the boundaries of integration. Using (59) and (65) gives

$$\begin{aligned}i\langle\langle\tilde{\mathbf{L}}_B(t)\rangle\rangle&=\int\cdots\int\left[\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}(W_R^{\text{eq}})+\nabla_{\mathbf{R}}(W_R^{\text{eq}})\cdot\frac{1}{\beta}\nabla_{\mathbf{P}}\right]d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\\&=0\end{aligned} \tag{73}$$

where the second equality follows from integration by parts with respect to the coordinates **R**. The stationarity relation follows from similar arguments beginning with

$$\begin{aligned}\langle\langle\tilde{\mathbf{L}}_B(t)\tilde{\mathbf{L}}_B(s)\rangle\rangle&\equiv\int\cdots\int\tilde{\mathbf{L}}_B(t)\tilde{\mathbf{L}}_B(s)W_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\\&=\int\cdots\int\exp(it\mathbf{L}_R)\mathbf{L}_B\exp(-it\mathbf{L}_R)\\&\quad\times\exp(is\mathbf{L}_R)\mathbf{L}_B\exp(-is\mathbf{L}_R)W_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}.\end{aligned} \tag{74}$$

Using an argument which is like that used in going from (70) to (71), we can go in the reverse direction and get

$$\begin{aligned}\langle\langle\mathbf{L}_B(t)\mathbf{L}_B(s)\rangle\rangle&=\int\cdots\int\exp(it\mathbf{L}_R)\mathbf{L}_B\exp(-it\mathbf{L}_R)\\&\quad\times\exp(is\mathbf{L}_R)\mathbf{L}_B\exp(-is\mathbf{L}_R)\exp(it\mathbf{L}_R)\\&\quad\times W_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\\&=\int\cdots\int\mathbf{L}_B\exp[i(s-t)\mathbf{L}_R]\mathbf{L}_B\\&\quad\times\exp[-i(s-t)\mathbf{L}_R]W_R^{\text{eq}}\,d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\end{aligned} \tag{75}$$

where the second equality follows from an argument identical with that used to get from (71) to (72). Therefore, the quantity given by (68) is in general

$$\begin{aligned}\int_t^\infty\langle\langle\tilde{\mathbf{L}}_B(t)\tilde{\mathbf{L}}_B(s)\rangle\rangle\,ds&=\int_t^\infty\langle\langle\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s-t)\rangle\rangle\,ds\\&=\int_0^\infty\langle\langle L_B(0)L_B(s')\rangle\rangle\,ds'\end{aligned} \tag{76}$$

where the first equality follows from (75) and the second equality follows from the change of variables $s'=s-t$. This ends our digression. The expression in (67) is an approximation to the exact behavior described by (62), and we shall analyze the detailed behavior of (68) regorously from here on, to the end of this section.

It is very convenient to get the W_R^{eq} term in (68) as far to the left as possible before attempting to perform integrations. This requires letting $\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)$ act on W_R^{eq} as is appropriate. Using (63) gives

$$-\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)W_R^{\text{eq}}g_B=-\mathbf{L}_B\exp(is\mathbf{L}_R)\mathbf{L}_B\exp(-is\mathbf{L}_R)W_R^{\text{eq}}g_B. \tag{77}$$

As is indicated in (67) there will be an integration over **R**. Using (59) for $\mathbf{L}_B$ and integrating over **R** by parts shows that

$$\int\cdots\int\left(-\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}+\sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}}\right)\exp(is\mathbf{L}_R)\mathbf{L}_B\times\exp(-is\mathbf{L}_R)W_R^{\text{eq}}g_B\,d\mathbf{R}$$

$$=\int\cdots\int\sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}}(\exp(is\mathbf{L}_R)\mathbf{L}_B \\ \times\exp(-is\mathbf{L}_R)W_R^{\text{eq}}g_B)\,d\mathbf{R} \tag{78}$$

where the integrated term at the boundaries vanishes because W_R^{eq} vanishes there. Equation (78) shows that the $\mathbf{L}_B$ on the left in (77) acts only through the potential term

$$-i\mathbf{L}_B'=\sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}}. \tag{79}$$

Therefore, in effect (77) reduces to

$$-\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)W_R^{\text{eq}}g_B\to-\mathbf{L}_B'\exp(is\mathbf{L}_R)\mathbf{L}_B\exp(-is\mathbf{L}_R)W_R^{\text{eq}}g_B, \tag{80}$$

at least after the $\mathbf{R}$ integration is performed. Returning to the justification of (70), we see that

$$\exp(-is\mathbf{L}_R)W_R^{\text{eq}}g_B=W_R^{\text{eq}}g_B. \tag{81}$$

Using (59) and (65), it is seen that

$$-i\mathbf{L}_BW_R^{\text{eq}}g_B=W_R^{\text{eq}}\Big(-\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}+\sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}} \\ +\beta\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}\sum_{j=1}^{N}\varphi(\mathbf{R},\mathbf{r}_j)\Big)g_B. \tag{82}$$

Furthermore, we have

$$\exp(is\mathbf{L}_R)W_R^{\text{eq}}h=W_R^{\text{eq}}\exp(is\mathbf{L}_R)h, \tag{83}$$

where h is any function of the variables $\mathbf{P}$, $\mathbf{R}$, $\mathbf{r}_j$, and $\mathbf{p}_j$ for $j=1,2,\ldots,N$. This relation follows from the reasons given in justification of (70). Therefore, (80)—(83) give

$$-\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)W_R^{\text{eq}}g_B \\ \to-\mathbf{L}_B'W_R^{\text{eq}}\exp(is\mathbf{L}_R)\Big(-\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}+\sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}} \\ +\beta\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}\sum_{j=1}^{N}\varphi(\mathbf{R},\mathbf{r}_j)\Big)g_B \\ =W_R^{\text{eq}}\ \sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}}\ \ \exp(is\mathbf{L}_R) \\ \times\Big(-\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}+\sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}}+\beta\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}\sum_{j=1}^{N}\varphi(\mathbf{R},\mathbf{r}_j)\Big)g_B \tag{84}$$

where the last equality follows from a result analogous to (82) which depends upon (79):

$$-i\mathbf{L}_B'W_R^{\text{eq}}h=W_R^{\text{eq}}(-i\mathbf{L}_B')h. \tag{85}$$

The expression in (84) becomes an equality when the $\mathbf{R}$ integration is performed. Note also that the $\mathbf{P}/M\cdot\nabla_{\mathbf{R}}$ operator in (84) acts on g_B only, and because g_B depends upon $\mathbf{P}$ only the effect is zero. Therefore, (84) becomes

$$-\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)W_R^{\text{eq}}g_B \\ \to W_R^{\text{eq}}\Big(\sum_{j=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_j)\cdot\nabla_{\mathbf{P}}\Big)\exp(is\mathbf{L}_R) \\ \times\Big(\sum_{k=1}^{N}\nabla_{\mathbf{R}}\varphi(\mathbf{R},\mathbf{r}_k)\cdot\nabla_{\mathbf{P}}+\beta\frac{\mathbf{P}}{M}\cdot\nabla_{\mathbf{R}}\sum_{k=1}^{N}\varphi(\mathbf{R},\mathbf{r}_k)\Big)g_B. \tag{86}$$

Finally, since $\mathbf{L}_R$ does not contain $\mathbf{P}$, we may combine terms as follows:

$$-\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)W_R^{\text{eq}}g_B \\ \to W_R^{\text{eq}}\Big(\sum_{j=1}^{N}\frac{\partial}{\partial R_\mu}\varphi(\mathbf{R},\mathbf{r}_j)\Big)\exp(is\mathbf{L}_R) \\ \times\Big(\sum_{k=1}^{N}\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}_k)\Big)\frac{\partial}{\partial P_\mu}\Big(\beta\frac{P_\nu}{M}+\frac{\partial}{\partial P_\nu}\Big)g_B. \tag{87}$$

Returning to (68), we have found that

$$-\int_0^\infty\int\cdots N\cdots\int\int\cdots N+1\cdots\int\tilde{\mathbf{L}}_B(0)\tilde{\mathbf{L}}_B(s)W_R^{\text{eq}} \\ \times d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\,g_B \\ =\int\int\cdots N\cdots\int\int\cdots N+1\cdots\int W_R^{\text{eq}} \\ \times\Big(\sum_{k=1}^{N}\frac{\partial}{\partial R_\mu}\varphi(\mathbf{R},\mathbf{r}_j)\Big)\exp(is\mathbf{L}_R) \\ \times\Big(\sum_{k=1}^{N}\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}_k)\Big)\frac{\partial}{\partial P_\mu}\Big(\beta\frac{P_\nu}{M}+\frac{\partial}{\partial P_\nu}\Big) \\ \times d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\,ds\,g_B. \tag{88}$$

If for a moment we specialize the above results to one dimension, and use the fact that $P=Mu$, then we have

$$\frac{\partial}{\partial P_\mu}\Big(\beta\frac{P_\nu}{M}+\frac{\partial}{\partial P_\nu}\Big)\to\frac{\partial}{\partial u}\Big(\beta\frac{u}{M}+\frac{1}{M^2}\frac{\partial}{\partial u}\Big) \tag{89}$$

and

$$\int_0^\infty\int\cdots\int W_R^{\text{eq}}\Big(\sum_{j=1}^{N}\frac{\partial}{\partial R_\mu}\varphi(\mathbf{R},\mathbf{r}_j)\Big)\exp(is\mathbf{L}_R) \\ \times\Big(\sum_{k=1}^{N}\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}_k)\Big)d^N\mathbf{r}\,d\mathbf{R}\,d^N\mathbf{p}\,ds \\ \to\int_0^\infty\int\cdots\int W_R^{\text{eq}}\Big(\sum_{j=1}^{N}\frac{\partial}{\partial R}\varphi(R,r_j)\Big)\exp(isL_R) \\ \times\Big(\sum_{k=1}^{N}\frac{\partial}{\partial R}\varphi(R,r_k)\Big)d^Nr\,dR\,d^Np\,ds, \tag{90}$$

where W_R^{eq} and L_R are one-dimensional expressions in the last expression above. Comparing these results with (4), we see that we have the Fokker—Planck equation for a Brownian particle in (67), if we use the one-dimensional analogs given by (89) and (90). Using (6), it is seen that we have the identity

$$\lambda=\int_0^\infty\int\cdots\int W_R^{\text{eq}}\ \sum_{j=1}^{N}\frac{\partial}{\partial R}\varphi(R,r_j)\ \ \exp(isL_R) \\ \times\sum_{k=1}^{N}\frac{\partial}{\partial R}\varphi(R,r_k)\ \ d^Nr\,dR\,d^Np\,ds. \tag{91}$$

This is all reasonable since we can readily associate

$$\sum_{j=1}^{N}\frac{\partial}{\partial R}\varphi(R,r_j)\to-\tilde{F}(0) \\ \text{and} \\ \exp(isL_R)\sum_{k=1}^{N}\frac{\partial}{\partial R}\varphi(R,r_k)\to-\tilde{F}(s). \tag{92}$$

This second association is perhaps even clearer if we reinsert $\exp(-isL_R)$ on the right of the integrand of (91), which changes nothing as was seen in going from (69) to (71). The strength of the stationary correlation, $\langle\langle\tilde{F}(0)\tilde{F}(s)\rangle\rangle$, follows from (2) and a formula like (52) which reads

$$\lambda = \int_0^\infty \langle\langle F(0)F(s)\rangle\rangle\, ds. \tag{93}$$

In Eq. (88), we actually have the three-dimensional generalization of (4) which gives

$$\lambda_{\mu\nu} \equiv \int_0^\infty\int \cdots \int W_R^{\text{eq}}\left(\sum_{j=1}^{N} \frac{\partial}{\partial R_\mu}\varphi(\mathbf{R}, \mathbf{r}_j)\right) \exp(isL_R) \times\left(\sum_{k=1}^{N} \frac{\partial}{\partial R_\nu}\varphi(\mathbf{R}, \mathbf{r}_k)\right) d^N\mathbf{r}\, d\mathbf{R}\, d^N\mathbf{p}\, ds \tag{94}$$

where $\lambda_{\mu\nu}$ is the friction tensor,[7] and it is useful in describing rotational Brownian motion if the potential φ is not spherically symmetric.[29]

It is to be especially noted that $\exp(is\mathbf{L}_R)$ appears in (94) and not the full Liouville evolution operator $\exp(is\mathbf{L})$. This point has been emphasized by Mori,[30,31] who also gets this type of result from a different approach. Other treatments involve $\exp(is\mathbf{L})$ in analogous expressions, and lead to certain technical difficulties, sometimes referred to as the "plateau value problem."[23,24,27] The relationship of (94) to Kubo's transport coefficient formula may be found in the references just cited.

ERGODICITY

A phase space function $f(q_1 \cdots q_N p_1 \cdots p_N t)$ is said to be ergodic if

$$\lim_{s\to\infty} \langle f(q_1 \cdots q_N p_1 \cdots p_N 0) f(q_1 \cdots q_N p_1 \cdots p_N s)\rangle = 0 \tag{95}$$

where $\langle \cdots \rangle$ denotes the canonical, initial value average.[32] We would like to show that the force on the Brownian particle $\sum_{j=1}^N \partial/\partial R_\nu\, \varphi(R, r_j)$ in Eq. (94) is ergodic in the sense of (95). Furthermore, if the vanishing of the correlation function in (94) is fast enough, then the integral over s will be finite. To show ergodicity it is necessary to analyze the detailed structure of $\exp(is\mathbf{L}_R)$. In doing so, we will discover a phenomenon which we shall call "intrinsic diffusion." This intrinsic diffusion has interest in its own right, independently of our particular context here.[33]

The time ordered exponential for a time dependent operator which does not commute with itself at different times is defined by

$$\underset{\leftarrow}{T}\exp\left(\int_0^s \mathrm{O}(s')\, ds'\right) \equiv \sum_{n=1}^{\infty} \int_0^s \int_0^{s_1} \int_0^{s_2} \cdots \int_0^{s_{n-2}} \int_0^{s_{n-1}} \mathrm{O}(s_1)\mathrm{O}(s_2) \cdots \mathrm{O}(s_{n-1})\mathrm{O}(s_n) \times ds_n\, ds_{n-1} \cdots ds_2\, ds_1. \tag{96}$$

Given an operator of the form $\exp[is(\mathbf{A}+\mathbf{B})]$ where $\mathbf{A}$ and $\mathbf{B}$ are noncommuting differential operators, we have the following disentanglement theorem:

$$\exp[is(\mathbf{A}+\mathbf{B})] = \exp(is\mathbf{A})\underset{\leftarrow}{T}\exp\left[i\int_0^s \exp(-is'\mathbf{A})\mathbf{B} \times \exp(is'\mathbf{A})\, ds'\right]. \tag{97}$$

Both sides of (97) are clearly equal for $s=0$, and differentiation of both sides for $s \neq 0$ gives identical results, proving the validity of (97).

When (97) is applied to $\exp(is\mathbf{L}_R)$, where we use (60), we get

$$\begin{aligned}\exp(is\mathbf{L}_R) &= \exp\left(s\sum_{j=1}^{N} \frac{\mathbf{p}_j}{m}\cdot\nabla_{\mathbf{r}_j}\right)\underset{\leftarrow}{T}\exp\left[-\int_0^s \exp\left(-s\sum_{k=1}^{N}\frac{\mathbf{p}_k}{m}\cdot\nabla_{\mathbf{r}_k}\right)\right.\\ &\quad\times\left(\sum_{l=1}^{N}\nabla_{\mathbf{r}_l}\varphi(\mathbf{R}, \mathbf{r}_l)\cdot\nabla_{\mathbf{p}_l} + \sum_{l\neq i}^{N}\nabla_{\mathbf{r}_l}U(\mathbf{r}_l, \mathbf{r}_i)\cdot\nabla_{\mathbf{p}_l}\right)\\ &\quad\left.\times\exp\left(s'\sum_{k'=1}^{N}\frac{\mathbf{p}_{k'}}{m}\cdot\nabla_{\mathbf{r}_{k'}}\right) ds'\right].\end{aligned} \tag{98}$$

The expression in (98) will be studied in detail in this paper with respect to the leading term only. This term is

$$\exp\left(s\sum_{j=1}^{N}\frac{\mathbf{p}_j}{m}\cdot\nabla_{\mathbf{r}_j}\right). \tag{99}$$

The detailed analysis of this term will indicate how the analysis goes for higher-order terms, although in this paper no explicit expressions for the higher-order terms will be offered.

Returning to (94), we see that we need to calculate the canonical average of expression (99). We shall perform the momenta integrations first. The leading term in (99) gives

$$\begin{aligned}&(2\pi m K_B T)^{-3N/2}\int\cdots\int \exp\left(-\beta\sum_{k=1}^{N}\frac{|\mathbf{p}_j|^2}{2m}\right)\times\exp\left(s\sum_{j=1}^{N}\frac{\mathbf{p}_j}{m}\cdot\nabla_{\mathbf{r}_j}\right) d^N\mathbf{p}\\ &= (2\pi m K_B T)^{-3N/2}\prod_{j=1}^{N}\int\int\int \exp\left(-\beta\frac{|\mathbf{p}_j|^2}{2m}\right) \exp\left(s\frac{\mathbf{p}_j}{m}\cdot\nabla_{\mathbf{r}_j}\right) d\mathbf{p}_j.\end{aligned} \tag{100}$$

For a fixed j, we shall now consider the integral over the x component of $\mathbf{p}_j$ because the other components and the other momenta integrals all work out similarly. We have, therefore, as a typical integral

$$(2\pi m K_B T)^{-1/2}\int_{\infty}^{-\infty} \exp\left(-\beta\frac{p_{jx}^2}{2m}\right)\exp\left(s\frac{p_{jx}}{\partial x}\frac{\partial}{\partial x}\right) dp_{jx} = \sum_{n=0}^{\infty}\frac{[(s/m)(\partial/\partial x)]^n}{n!}\langle p_{jx}^n\rangle \tag{101}$$

where

$$\langle p_{jx}^n\rangle \equiv (2\pi m B K_B T)^{-1/2}\int_{-\infty}^{\infty} p_{jx}^n \exp\left(-\beta\frac{p_{jx}^2}{2m}\right) dp_{jx}.$$

These Gaussian integrals are well known and are given by

$$\langle p_{jx}^n\rangle = 0 \qquad \text{for } n = 2m+1 \text{ where } m = 0, 1, 2, \cdots$$

and (102)

$$\langle p_{jx}^n\rangle = \frac{(2m)!}{2^m m!}\langle p_{jx}^2\rangle^m \quad \text{for } n = 2m \text{ where } m = 0, 1, 2, \cdots .$$

Inserting these averages into (101) gives

$$(2\pi m K_B T)^{-1/2}\int_{-\infty}^{\infty}\exp\left(-\beta\frac{p_{jx}^2}{2m}\right)\exp\left(s\frac{p_{jx}}{m}\frac{\partial}{\partial x}\right) dp_{jx} = \exp\left(s^2\frac{K_B T}{2m}\frac{\partial^2}{\partial x^2}\right) \tag{103}$$

because $\langle p_{jx}^2\rangle = mK_BT$. Returning to (100), we get

$$(2\pi m K_B T)^{-3N/2}\prod_{j=1}^{N}\iiint \exp\left(-\beta\frac{|\mathbf{p}_j|^2}{2m}\right)\exp\left(s\frac{\mathbf{p}_j}{m}\cdot\nabla_{\mathbf{r}_j}\right)d\mathbf{p}_j$$
$$=\prod_{j=1}^{N}\exp\left(s^2\frac{K_BT}{2m}\nabla^2_{\mathbf{r}_j}\right). \qquad (104)$$

If only the leading term in (99) is used in expression (94), then the canonical momenta averages give

$$\lambda^0_{\mu\nu}=\int_0^\infty\int\cdots\int W^{\text{eq}}_*(\mathbf{R},\mathbf{r}_1\cdots\mathbf{r}_N)\left(\sum_{j=1}^{N}\frac{\partial}{\partial R_\mu}\varphi(\mathbf{R},\mathbf{r}_j)\right)$$
$$\times\prod_{l=1}^{N}\exp\left(s^2\frac{K_BT}{2m}\nabla^2_{\mathbf{r}_l}\right)\left(\sum_{k=1}^{N}\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}_k)\right)d^N\mathbf{r}\,d\mathbf{R}\,ds \qquad (105)$$

where $W^{\text{eq}}_*(\mathbf{R},\mathbf{r}_1\cdots\mathbf{r}_N)$ is the coordinate dependent factor in W^{eq}_R, and the superscript on $\lambda^0_{\mu\nu}$ indicates the lowest order truncation of the series implicit in (98) by virtue of (96). A typical term in (105) has an integrand factor such as

$$\exp\left(s^2\frac{K_BT}{2m}\nabla^2_{\mathbf{r}_l}\right)\left(\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}_l)\right). \qquad (106)$$

We shall express (106) in a different form which is suggested by an analogy with diffusion. It is worthwhile to digress for a moment in order to make the analogy clear.

The diffusion equation in three dimensions is

$$\frac{\partial}{\partial t}D(\mathbf{r},t)=D\nabla^2_{\mathbf{r}}D(\mathbf{r},t) \qquad (107)$$

where D is the diffusion constant. A formal solution to (107) may be written as

$$D(\mathbf{r},t)=\exp(tD\nabla^2_{\mathbf{r}})D(\mathbf{r},0) \qquad (108)$$

as is readily verified by differentiation. In addition, the solution to (107) in an infinite volume, with the initial condition $D(\mathbf{r},0)=\delta(\mathbf{r}-\mathbf{r}^0)$, is given by

$$D_G(\mathbf{r},t)=(4\pi Dt)^{-3/2}\exp-\frac{|\mathbf{r}-\mathbf{r}^0|^2}{4Dt} \qquad (109)$$

where the subscript G indicates that this solution can be used as a Green's function, which enables us to conclude from (108) and (109) that

$$\exp(tD\nabla^2_{\mathbf{r}})D(\mathbf{r},0)=(4\pi Dt)^{-3/2}\iiint$$
$$\exp-\frac{|\mathbf{r}-\mathbf{r}^0|^2}{4Dt}\;D(\mathbf{r}^0,0)\,d\mathbf{r}^0 \qquad (110)$$

for arbitrary initial distrubutions $D(\mathbf{r}^0,0)$.

The analogy between these results for diffusion and expression (106) should be clear. The important differences are $K_BT/2m$ in (106) where D is in (108), and s^2 in (106) where t is in (108). However, remarkably enough, there is a Green's function which goes with (106) given by

$$\exp\left(s^2\frac{K_BT}{2m}\nabla^2_{\mathbf{r}_l}\right)\left(\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}_l)\right)$$
$$=\left(4\pi\frac{K_BT}{2m}s^2\right)^{-3/2}\iiint\exp\left(-\frac{|\mathbf{r}_l-\mathbf{r}'_l|^2}{4(K_BT/2m)s^2}\right)$$
$$\times\left(\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}'_l)\right)d\mathbf{r}'_l. \qquad (111)$$

The corresponding differential equation follows from (106) and is

$$\frac{\partial}{\partial s}D(\mathbf{r},s)=2s\frac{K_BT}{2m}\nabla^2_{\mathbf{r}}(\mathbf{r},s) \qquad (112)$$

which is analogous with (107). One may check all this by showing that the Green's function in (111) satisfies (112) with initial condition $D(\mathbf{r},0)=\delta(\mathbf{r}-\mathbf{r}^0)$, in an infinite volume. We shall refer to the behavior exhibited by (111) as "intrinsic diffusion" to distinguish it from true diffusion as exhibited by (110) while suggesting the marked similarities. Notice in particular, that "intrinsic diffusion" acts like true diffusion with a diffusion coefficient which grows linearly with time: $D\to(K_BT/2m)s$. Therefore, "intrinsic diffusion" smooths out coordinate dependent functions increasingly rapidly compared with true diffusion. Moreover, it increases the rate of smoothing proportionately with temperature and inverse fluid molecule mass.

Returning to (105), and using (111), we get

$$\lambda^0_{\mu\nu}=\sum_{j=1}^{N}\sum_{k=1}^{N}\int_0^\infty\int\cdots\int W^{\text{eq}}_*\left(\frac{\partial}{\partial R_\mu}\varphi(\mathbf{R},\mathbf{r}_j)\right)\left(4\pi\frac{K_BT}{2m}s^2\right)^{-3/2}$$
$$\times\exp\left(-\frac{|\mathbf{r}_k-\mathbf{r}'_k|^2}{4(K_BT/2m)s^2}\right)\left(\frac{\partial}{\partial R_\nu}\varphi(\mathbf{R},\mathbf{r}'_k)\right)d^N\mathbf{r}'\,d^N\mathbf{r}\,d\mathbf{R}\,ds.$$
$$(113)$$

It is convenient to use a Fourier transformation representation of the potentials in evaluating further expression (113). Define $\hat{\varphi}(\rho)$ by

$$\varphi(\mathbf{R},\mathbf{r}_k)\equiv\frac{1}{2\pi}{}^{3}\iiint\exp[i\boldsymbol{\rho}\cdot(\mathbf{R}-\mathbf{r}_k)]\hat{\varphi}(\boldsymbol{\rho})\,d\boldsymbol{\rho} \qquad (114)$$

using (111) in (113) with (114) gives

$$\lambda^0_{\mu\nu}=\sum_{j=1}^{N}\sum_{k=1}^{N}\int_0^\infty\int\cdots\int W^{\text{eq}}_*\left(\frac{1}{2\pi}\right)^6(i\rho_\mu)\exp[i\boldsymbol{\rho}\cdot(\mathbf{R}-\mathbf{r}_j)]\hat{\varphi}(\boldsymbol{\rho})$$
$$\times\exp\left(-s^2\frac{K_BT}{2m}|\boldsymbol{\rho}'|^2\right)\exp[i\boldsymbol{\rho}'\cdot(\mathbf{R}-\mathbf{r}_k)]\hat{\varphi}(\boldsymbol{\rho}')$$
$$\times d\boldsymbol{\rho}'\,d\boldsymbol{\rho}\,d^N\mathbf{r}\,d\mathbf{R}\,ds. \qquad (115)$$

Equation (115) also follows directly from (105) using (114). In getting (115) we have used $\exp[s^2(K_BT/2m)\nabla^2_{\mathbf{r}}]\times\exp(i\boldsymbol{\rho}'\cdot\mathbf{r})=\exp[-s^2(K_BT/2m)|\boldsymbol{\rho}'|^2]$. Either (113) or (115) shows that to this lowest order of truncation we have ergodic behavior. Note that in (115) the $\boldsymbol{\rho}'=0$ term, which does not decay, is multiplied by $i\rho'_\nu$, which is zero. In (113) we see that qualitatively, "intrinsic diffusion" smoothes out the force at time s, for large s, so that we are left with the canonical coordinate average of the initial force times essentially a constant:

$$\lim_{s\to\infty}(\text{time integrand of }\lambda^0_{\mu\nu})$$
$$\approx\sum_{j=1}^{N}\sum_{k=1}^{N}\int\cdots\int W^{\text{eq}}_*\left(\frac{\partial}{\partial R_\mu}\varphi(R,r_j)\right)C\,d^N\mathbf{r}\,d\mathbf{R}$$
$$=0 \qquad (116)$$

where C is the constant corresponding with the smoothed force at time s, and the second equality follows from a result similar to (73). In addition, (115) may be integrated over s explicitly giving

$$\lambda^0_{\mu\nu}=\sum_{j=1}^{N}\sum_{k=1}^{N}\int\cdots\int W_*^{\text{eq}}\left(\frac{1}{2\pi}\right)^6(i\rho_\mu)\exp[i\boldsymbol{\rho}\cdot(\mathbf{R}-\mathbf{r}_j)]\hat{\varphi}(\boldsymbol{\rho})$$
$$\times\left(\frac{\pi m}{2K_BT|\boldsymbol{\rho}'|^2}\right)^{1/2}(i\rho'_\nu)\exp[i\boldsymbol{\rho}'\cdot(\mathbf{R}-\mathbf{r}_k)]\,\hat{\varphi}(\boldsymbol{\rho}')$$
$$\times d\boldsymbol{\rho}'\,d\boldsymbol{\rho}\,d^N\mathbf{r}\,d\mathbf{R}. \tag{117}$$

As long as $\hat{\varphi}(0)$ is finite, the $(|\boldsymbol{\rho}'|^2)^{1/2}$ denominator in (117) does not create a divergence because of the factor $(i\rho'_\nu)$. If $\varphi(\mathbf{R},\mathbf{r}_k)$ is spherically symmetric, (114) may be inverted to show that $\hat{\varphi}(0)=4\pi\int_0^\infty r^2\varphi(r)\,dr$, which is finite for a large class of physically reasonable potentials, which do not possess too terribly singular hard cores, and which tail off rapidly with increased separation of the interacting particles. The presence of W_*^{eq} in (117) makes it effectively impossible to perform the coordinate integrals in general. We will end our present discussion of $\lambda^0_{\mu\nu}$ by approximating the coordinate integrals for high temperatures.

At sufficiently high temperatures we shall approximate W_*^{eq} by $V^{-(N+1)}$, where V is the volume of our system, which volume we have been taking to be essentially infinite during much of the preceeding analysis. Although (113) is valid only for an infinite V, since the Green's function in its integrand obtains only for an infinite V, (117) is valid for finite V since it follows from (105) and (114) without any restriction on V. The approximation for W_*^{eq} by $V^{-(N+1)}$ requires that the potential have a hard core repulsion which, while very large, is finite. The limit $V\to\infty$ is accompanied by $N\to\infty$ while N/V remains equal to a constant $\hat{n}$. This is the thermodynamic limit. With the thermodynamic limit the coordinate integrals in (117) can be performed, and we get

$$\lambda^0_{\mu\nu}\approx\lim_{\text{thermo}}V^{-(N+1)}\sum_{j=1}^{N}\sum_{k=1}^{N}\int\cdots\int\left(\frac{1}{2\pi}\right)^6(i\rho_\mu)$$
$$\times\exp[i\boldsymbol{\rho}\cdot(\mathbf{R}-\mathbf{r}_j)]\hat{\varphi}(\boldsymbol{\rho})\left(\frac{\pi m}{2K_BT}\right)^{1/2}\frac{1}{|\boldsymbol{\rho}'|}(i\rho'_\nu)$$
$$\times\exp[i\boldsymbol{\rho}'\cdot(\mathbf{R}-\mathbf{r}_k)]\hat{\varphi}(\boldsymbol{\rho}')\,d\boldsymbol{\rho}'\,d\boldsymbol{\rho}\,d^N\mathbf{r}\,d\mathbf{R}$$
$$=\hat{n}\iiint\frac{1}{2\pi}{}^3(\rho_\mu\rho_\nu)\,|\hat{\varphi}(\boldsymbol{\rho})|^2\left(\frac{\pi m}{2K_BT}\right)^{1/2}\frac{1}{|\boldsymbol{\rho}|}d\boldsymbol{\rho}$$
$$+\hat{n}^2\iiint\delta(\boldsymbol{\rho})(\rho_\mu\rho_\nu)\,|\hat{\varphi}(\boldsymbol{\rho})|^2\left(\frac{\pi m}{2K_BT}\right)^{1/2}\frac{1}{|\boldsymbol{\rho}|}d\boldsymbol{\rho}$$
$$=\hat{n}\iiint\frac{1}{2\pi}{}^3(\rho_\mu\rho_\nu)\,|\hat{\varphi}(\boldsymbol{\rho})|^2\left(\frac{\pi m}{2K_BT}\right)^{1/2}\frac{1}{|\boldsymbol{\rho}|}d\boldsymbol{\rho}. \tag{118}$$

The expression given by the second equality has a term proportional to $\hat{n}$ which comes from the double sum $\sum_j\sum_k$ when $j=k$, and a term proportional to $\hat{n}^2$ which comes when $j\neq k$. The $\mathbf{R}$ integration produced a factor $\delta(\boldsymbol{\rho}+\boldsymbol{\rho}')$ which converted $\hat{\varphi}(\boldsymbol{\rho})\hat{\varphi}(\boldsymbol{\rho}')$ into $\hat{\varphi}(\boldsymbol{\rho})\hat{\varphi}(-\boldsymbol{\rho})=\hat{\varphi}(\boldsymbol{\rho})\hat{\varphi}^*(\boldsymbol{\rho})\equiv|\hat{\varphi}(\boldsymbol{\rho})|^2$ as is seen using (114). The $|\boldsymbol{\rho}|$ denominator in each term causes no trouble because of compensating numerator factors in $\rho_\mu\rho_\nu$, which also explain how the $\hat{n}^2$ term vanishes. If $|\hat{\varphi}(\boldsymbol{\rho})|^2$ vanishes for large $|\boldsymbol{\rho}|$ sufficiently rapidly, then the remaining integral is finite. In general, when W_*^{eq} is present and no approximation is used the value of $\lambda^0_{\mu\nu}$ does not have such simple $\hat{n}$ or T dependence.

These techniques need to be extended for the analysis of higher order terms in (94) using (98).

CRITICAL BEHAVIOR

That something can be said about critical behavior within the context of this paper follows from the observation that a Brownian particle can be used as a probe which manifests the state of a fluid. If the fluid passes through a phase transition, the Brownian particle should exhibit noticeably altered behavior. In particular, for high enough temperatures the fluid will be a liquid or gas, and the Brownian motion should be normal with a finite value for $\lambda_{\mu\nu}$. However, if we freeze the fluid, we should expect that $\lambda_{\mu\nu}\to\infty$. Using (6) and (8), we see that this implies a vanishing diffusion constant D.[34]

If a complete analysis of (94) using (98) were possible, we would end up with the infinite series: $\lambda_{\mu\nu}=\lambda^0_{\mu\nu}+\lambda^1_{\mu\nu}+\lambda^2_{\mu\nu}+\cdots$ in which the superscripts indicate the number of times a potential appears in the corresponding term. We suspect that intrinsic diffusion will result in each term being finite, as was seen for $\lambda^0_{\mu\nu}$. When $T=0$, however, we have a special case in which each term is in fact infinite. For $\lambda^0_{\mu\sigma}$, this is seen in (115) because the exponential can no longer damp out the integrand when $T=0$. Consequently, $\lambda_{\mu\nu}$ diverges for $T=0$; a fact consistent with intuitive expectations. For $T\neq 0$ each term in $\lambda_{\mu\nu}=\lambda^0_{\mu\nu}+\lambda^1_{\mu\nu}+\lambda^2_{\mu\nu}+\cdots$ will be finite, or at least this is plausible following our analysis of $\lambda^0_{\mu\nu}$, and if the temperature is sufficiently large, we expect the series to be summable. This expectation is based upon qualitative analysis of (94) using (98) which suggests that the $\lambda_{\mu\nu}$ series goes like a series in powers of the ratio of potential energy and thermal energy. For large temperatures this ratio gets small and the series converges. The possibility exists that each term in $\lambda_{\mu\nu}=\lambda^0_{\mu\nu}+\lambda^1_{\mu\nu}+\lambda^2_{\mu\nu}+\cdots$ is finite while the sum diverges, if the temperature is small, but nonzero. This suggests that there exists a temperature T_c such that for $T>T_c$ the series sums while for $T\leq T_c$ the series diverges, even though it is comprised of finite terms. If T_c exists, we identify it as the critical temperature for the fluid. It is desirable to try to achieve a rigorous basis for the analysis of (94) using (98) so that these conjectures may be tested.

CONCLUDING REMARKS

The theory of multiplicative stochastic processes has provided a method for nonperturbative approximate solution of complicated dynamical equations. It has been shown how to interpret an exact dynamical quantity as a stochastic quantity. By such an interpretation an averaged equation is written which serves as a nonperturbative approximation for the original dynamics. In this manner we have derived the Fokker–Planck equation for Brownian motion, starting from the exact dynamics for a heavy particle moving in a molecular fluid.

The discovery of "intrinsic diffusion" has permitted an analysis of the ergodicity of the force autocorrelation function which was derived. Because Brownian motion leads to ordinary diffusion, "intrinsic diffusion" and ordinary diffusion are connected by our analysis. "Intrinsic diffusion" will also arise in other contexts.[33]

The analysis of Brownian motion in this paper leads to a calculation of λ, the dissipative constant, in terms of the interparticle potentials. The calculation involves the time integral of a force autocorrelation function which connects λ with the dynamics of the system. A conjectured relationship between λ and critical phenomena is presented, which if eventually justified rigorously, would provide a dynamical basis for the understanding of some aspects of critical behavior.

In our other work on the quantum mechanical density matrix,[3,5] the problem of calculating the matrices, $Q_{\alpha\beta\mu\nu}$, which are the analogs of λ, has been raised. We shall pursue this problem in another paper.

ACKNOWLEDGMENT

This work was partially supported by NSF grant GP-39309.

APPENDIX: DERIVATION OF (50) FROM (48) AND (49).

The correlation formula in (48) follows directly from the solution to Eqs. (1) and (2). In addition, the Gaussian property of $\widetilde{F}(t)$ in (1) is inherited by $\widetilde{p}(t)$ in (48).[11] This implies

$$\langle\langle \widetilde{p}(t_1)\cdots\widetilde{p}(t_{2m-1})\rangle\rangle = 0 \quad \text{for } m=1,2,\cdots \tag{A1}$$

and

$$\begin{aligned}\langle\langle \widetilde{p}(t_1)\cdots\widetilde{p}(t_{2m})\rangle\rangle &= \frac{1}{2^m m!}\sum_{p\in S_{2m}}\prod_{j=1}^{m}\langle\langle \widetilde{p}(t_{p(2j-1)})\widetilde{p}(t_{p(2j)})\rangle\rangle\\ &= \frac{(MK_BT)^m}{2^m m!}\sum_{p\in S_{2m}}\prod_{j=1}^{m}\\ &\quad\times\exp\left[-\frac{\alpha}{M}\left|t_{p(2j)}-t_{\ (2j-1)}\right|\right]\\ &\qquad \text{for } m=1,2,\cdots\end{aligned} \tag{A2}$$

The formal solution to (49) is

$$f(x,t)=\exp\left(-\int_0^t \frac{\widetilde{p}(s)}{M}\,ds\frac{\partial}{\partial x}\right)f(x,0), \tag{A3}$$

as is verified by substitution and differentiation. Expanding the exponential and averaging gives

$$\langle\langle f(x,t)\rangle\rangle = \left[\left\langle\sum_{n=0}^{\infty}\left(-\frac{1}{M}\right)^n\frac{1}{n!}\left(\int_0^t\widetilde{p}(s)\,ds\right)^n\frac{\partial^n}{\partial x^n}\right\rangle\right]f(x,0). \tag{A4}$$

Using (A1) and (A2) in (A4) gives

$$\begin{aligned}\langle\langle f(x,t)\rangle\rangle &= \sum_{m=0}^{\infty}\left(-\frac{1}{M}\right)^{2m}\frac{1}{(2m)!}\frac{(MK_BT)^m}{2^m m!}\sum_{p\in S_{2m}}\prod_{j=1}^{m}\int_0^t\int_0^t\\ &\quad\times\exp\left(-\frac{\alpha}{M}\left|t_{p(2j)}-t_{p(2j-1)}\right|\right)dt_{p(2j)}\,dt_{p(2j-1)}\\ &\quad\times\frac{\partial^{2m}}{\partial x^{2m}}f(x,0)\\ &= \sum_{m=0}^{\infty}\left(\frac{K_BT}{M}\right)^{2m}\frac{1}{2^m m!}\left(\int_0^t\int_0^t\exp\left[-\frac{\alpha}{m}\left|s-s'\right|ds\,ds'\right]\right)^m\\ &\quad\times\left(\frac{\partial^2}{\partial x^2}\right)^m f(x,0)\\ &= \sum_{m=0}^{\infty}\frac{1}{2^m m!}\frac{1}{M^{2m}}\left(\int_0^t\int_0^t\langle\langle\widetilde{p}(s)\widetilde{p}(s')\rangle\rangle\,ds\,ds'\right)^m\\ &\quad\times\left(\frac{\partial^2}{\partial x^2}\right)^m f(x,0).\end{aligned} \tag{A5}$$

Therefore,

$$\frac{\partial}{\partial t}\langle\langle f(x,t)\rangle\rangle = \int_0^t\frac{1}{M^2}\langle\langle\widetilde{p}(t)\widetilde{p}(s)\rangle\rangle\,ds\frac{\partial^2}{\partial x^2}\langle\langle f(xt)\rangle\rangle \tag{A6}$$

which is (50), and which follows from (A5) by differentiation, followed by rearrangement of terms.

Equation (A6) can be integrated to give (51) of the text. The solution to (51) is given by

$$\begin{aligned}\langle\langle f(xt)\rangle\rangle &= \left\{4\pi\frac{K_BT}{\alpha}\left[t-\frac{M}{\alpha}+\frac{M}{\alpha}\exp\left(-\frac{\alpha}{M}t\right)\right]^{-1/2}\right\}\\ &\quad\times\exp\left(-\frac{\alpha}{4K_BT}\frac{(x-x_0)^2}{\{t-(M/\alpha)+(M/\alpha)\exp[-(\alpha/M)t]\}}\right)\end{aligned} \tag{A7}$$

if $\langle\langle f(x,0)\rangle\rangle = \delta(x-x_0)$. This expression can be checked by substitution. It leads to the Orstein—Fürth formula[35]

$$\langle\langle (x(t)-x_0)^2\rangle\rangle = \frac{2MK_BT}{\alpha^2}\left[\frac{\alpha}{M}t-1+\exp\left(-\frac{\alpha}{M}t\right)\right]. \tag{A8}$$

Note also that for the short times, $t<M/\alpha$, that (A7) reduces to

$$\left(4\pi\frac{K_BT}{2M}\,t^2\right)^{-1/2}\exp\left(-\frac{(x-x_0)^2}{4(K_BT/2M)t^2}\right) \tag{A9}$$

which is the Green's function for intrinsic diffusion in the text. However, for long time, $t>M/\alpha$, (A7) becomes the one-dimensional analog of (109) in the text which is valid for ordinary diffusion. The intrinsic diffusion Green's function has the form (A9) for all times!

[1] R.F. Fox, J. Math. Phys. **13**, 1196 (1972).
[2] R.F. Fox, J. Math. Phys. **13**, 1726 (1972).
[3] R.F. Fox, J. Math. Phys. **14**, 20 (1973).
[4] R.F. Fox, J. Math. Phys. **14**, 1187 (1973).
[5] R.F. Fox, J. Math. Phys. **15**, 217 (1974).
[6] R. Kubo and N. Hashitsume, Supp. Prog. Theor. Phys. **46**, 210 (1970).
[7] R.F. Fox and G.E. Uhlenbeck, Phys. Fluids **13**, 1893 (1970).
[8] R.F. Fox and G.E. Uhlenbeck, Phys. Fluids **13**, 2881 (1970).
[9] P. Langevin, Comp. Rend. **146**, 530 (1908).
[10] G.E. Uhlenbeck and L.S. Ornstein, Phys. Rev. **36**, 823 (1930).
[11] M.C. Wang and G.E. Uhlenbeck, Rev. Mod. Phys. **17**, 323 (1945).
[12] L. Onsager and S. Machlup, Phys. Rev. **91**, 1505 (1953).
[13] L.D. Landau, and E.M. Lifshitz, *Fluid Mechanics* (Pergamon, New York, 1959), Chap. 17.
[14] J. Foch, Phys. Fluids **14**, 893 (1971).
[15] J. Foch, Phys. Fluids **15**, 224 (1972).
[16] C. Cohen, Ph.D. Thesis, Princeton University (1971).

[17] H. Haken, *Handbuch Der Physik* (Springer-Verlag, Berlin, 1970), Vol. 25, Pt. 2c.
[18] A. Einstein, *Theory of Brownian Movement* (Dover, New York, 1956).
[19] R. F. Fox, Ph.D. Thesis, The Rockefeller University (1969).
[20] L. D. Landau, and E. M. Lifshitz, *Electrodynamics of Continuous Media* (Pergamon, New York, 1960), Chap. 13.
[21] K. Huang, *Statistical Mechanics* (Wiley, New York, 1963), pp. 76—77.
[22] R. F. Fox, J. Math. Phys. **15**, 000 (1974) (Sep).
[23] J. G. Kirkwood, J. Chem. Phys. **14**, 180 (1946).
[24] J. Ross, J. Chem. Phys. **24**, 375 (1956).
[25] J. L. Lebowitz, and E. Rubin, Phys. Rev. **131**, 2381 (1963).
[26] R. Kubo, unpublished lecture notes (1963).
[27] R. Zwanzig, J. Chem. Phys. **40**, 2527 (1964).
[28] T. S. Chow and J. J. Hermans, J. Chem. Phys. **56**, 3150 (1972).
[29] E. H. Hauge and A. Martin-Löf, J. Stat. Phys. **7**, 259 (1973).
[30] H. Mori, Phys. Lett. **9**, 136 (1964).
[31] H. Mori, Prog. Theor. Phys. **33**, 423 (1965).
[32] A. I. Khinchin, *Mathematical Foundations of Statistical Mechanics* (Dover, New York, 1949), Chap. 3.
[33] R. F. Fox (unpublished).
[34] H. E. Stanley, *Phase Transitions and Critical Phenomena* (Clarendon, Oxford, 1971), Chap. 13 and 15.
[35] See Ref. 10, Sec. III, pp. 830—832.

Weyl tensor decomposition in stationary vacuum space-times

E. N. Glass*

Joseph Henry Laboratories, Princeton University, Princeton, New Jersey 08540

(Received 8 May 1974)

The electric and magnetic parts of the Weyl tensor are represented as symmetrized derivatives of gradient vector fields in stationary vacuum space-times. It is shown that a necessary and sufficient condition for a stationary vacuum space-time to be static is that the Weyl tensor be electric type. It is further shown that the only stationary vacuum space-time with vanishing electric type Weyl tensor is flat space.

In a recent work,[1] Hallidy proved the asymptotic result that a necessary condition for asymptotically flat, stationary, vacuum space—times to be static is that the Weyl tensor be electric type to order r^{-7}, where r is an affine parameter along null geodesics. This result is a special case of a theorem which we present in this note.

Theorem: A necessary and sufficient[2] condition for a stationary vacuum space—time to be static is that the Weyl tensor be electric type.

Proof: Consider a vacuum space—time with timelike Killing vector ξ^α. The unit vector field along ξ^α is defined by

$$u^\alpha := \phi^{-1}\xi^\alpha,$$

where $\phi^2 := \xi^\alpha \xi_\alpha$. Stationary space—times are distinguished from static ones by the twist of the timelike Killing vector[3]:

$$\Omega^\mu := \tfrac{1}{2}\eta^{\mu\nu\rho\sigma}\xi_\nu \xi_{\rho;\sigma},$$

where $\eta^{0123} = -(-g)^{-1/2}$. Ω_μ is curl free and can be expressed locally as a gradient field

$$\Omega_\mu = \nabla_\mu \Omega.$$

The twist bivector is given by

$$\Omega_{\mu\alpha} = \eta_{\mu\alpha\rho\sigma}u^\rho \Omega^\sigma,$$

and it will be useful to define a vector proportional to the acceleration of the Killing field

$$A_\mu := -\nabla_\mu \phi = \phi^{-1}\xi^\alpha \nabla_\alpha \xi_\mu .$$

The Killing vector and Weyl tensor are related by

$$\xi_{\mu;\rho\sigma} = C^\alpha{}_{\sigma\rho\mu}\,\xi_\alpha, \tag{1}$$

which follows from Killing's equation, the Ricci identity, and the equality of the Riemann and Weyl tensors in vacuum. Electric and magnetic parts of the Weyl tensor are defined by[4]

$$\begin{aligned} E_{\mu\rho} &:= C_{\mu\nu\rho\sigma}u^\nu u^\sigma, \\ B_{\mu\rho} &:= \overset{*}{C}_{\mu\nu\rho\sigma}u^\nu u^\sigma, \end{aligned} \tag{2}$$

where

$$\overset{*}{C}_{\mu\nu\rho\sigma} := \tfrac{1}{2}\eta_{\mu\nu\alpha\beta}C^{\alpha\beta}{}_{\rho\sigma}.$$

Both electric and magnetic tensors are symmetric, trace-free, and orthogonal to u^α.

The Weyl tensor decomposition follows directly from Eq. (1) and its dual:

$$\begin{aligned} \phi^2 E_{\mu\nu} &= \phi \perp A_{(\mu;\nu)} + \phi^{-2}\,\Omega_{\mu\alpha}\Omega_\nu{}^\alpha, \\ \phi^2 B_{\mu\nu} &= \perp \Omega_{(\mu;\nu)} + \phi^{-1}\gamma_{\mu\nu}\Omega^\alpha A_\alpha, \end{aligned} \tag{3}$$

where $\perp$ symbolizes the projection operator $\gamma^\mu{}_\nu := g^\mu{}_\nu - u^\mu u_\nu$, and projects all free indices. The "necessary" part of the theorem follows when the twist of ξ^α is set equal to zero in Eq. (3).

Note that the trace-free property of $E_{\mu\nu}$ and $B_{\mu\nu}$ provides the relationship[5] between the twist potential Ω and the Killing vector norm ϕ:

$$\begin{aligned} D^2\Omega &= 3\phi^{-1}(D^\alpha\Omega)(D_\alpha\phi), \\ D^2\phi &= -2\phi^{-3}(D^\alpha\Omega)(D_\alpha\Omega), \end{aligned} \tag{4}$$

where $D_\alpha := \gamma_\alpha{}^\beta\nabla_\beta$, and $D^2 := \gamma^{\alpha\beta}D_\alpha D_\beta$.

To prove sufficiency, examine the divergence of Eq. (2):

$$\begin{aligned} \nabla^\mu E_{\mu\rho} &= -\phi^{-1}E_{\nu\rho}A^\nu + 3\phi^{-2}B_{\nu\rho}\Omega^\nu, \\ \nabla^\mu B_{\mu\rho} &= -\phi^{-1}B_{\nu\rho}A^\nu - 3\phi^{-2}E_{\nu\rho}\Omega^\nu, \end{aligned} \tag{5}$$

where the divergence free property of the Weyl tensor in vacuum and

$$\phi\,\xi_{\mu;\nu} = 2A_{[\mu}\xi_{\nu]} + \Omega_{\mu\nu}$$

have been used. When the magnetic part of the Weyl tensor is equated to zero in Eq. (5) the resulting condition

$$E_{\nu\rho}\Omega^\nu = 0$$

implies $\Omega^\nu = 0$ when $\det(E_{\nu\rho}) \neq 0$, which is the case for Petrov types I, II, and D.[6] If asymptotic flatness is imposed as a boundary condition,[7] then stationary vacuum solutions can only be type I or D.

Lemma: Any stationary vacuum space—time[2] with vanishing electric type Weyl tensor must be flat.

Proof: Setting $E_{\mu\rho}$ equal to zero in Eq. (5) results in $B_{\nu\rho}\Omega^\nu = 0$. It follows that $\det(B_{\nu\rho}) = 0$ (if it were unequal to zero then the solution $\Omega^\nu = 0$ would yield a contradiction by the theorem proved above). For the space—times under consideration $\det(B_{\nu\rho}) = 0$ iff$B_{\nu\rho} = 0$.

We thank B. Mashhoon and R. Geroch for valuable comments.

*Present address: Physics Department, University of Windsor, Windsor, Ontario N9B 3P4, Canada.

[1] W. Hallidy, J. Math. Phys. 15, 413 (1974).

[2] Sufficient for Petrov type I, II, and D spaces without boundary conditions. Sufficient for all asymptotically flat, stationary, vacuum space—times.

[3] Covariant derivatives are denoted by semicolons and ∇_μ. Parentheses around indices denote symmetrization, brackets around indices denote antisymmetrization, and the convention for the Riemann tensor is fixed by $R^\mu{}_{\nu\rho\sigma}A_\mu \equiv A_{\nu;\rho\sigma} - A_{\nu;\sigma\rho}$.

[4] The definitions of the electric and magnetic parts of the Weyl tensor in Ref. 1 and in this work agree for stationary space—times. See E.T. Newman and R. Penrose, Proc. Roy. Soc. A **305**, 175 (1968).

[5] Equation (4) agrees with R. Geroch, J. Math. Phys. **12**, 918 (1971), Eq. (A18), where his $(\lambda, \omega) = (\phi^2, 2\Omega)$ here.

[6] A.Z. Petrov, *Einstein Spaces* (Pergamon, Oxford, 1969), p. 110.

[7] R.W. Lind, "Shear-free, Twisting Einstein—Maxwell Metrics in the Newman—Penrose Formalism" (preprint).

The maximal solvable subgroups of *SO(p,q)* groups

J. Patera*, P. Winternitz*, and H. Zassenhaus†

*Centre de Recherches Mathématiques, Université de Montréal, Montréal, P. Q., Canada

†Department of Mathematics, Ohio State University, Columbus, Ohio

(Received 16 May 1974)

A recursive procedure is developed that makes it possible to determine all conjugacy classes under both $SO(p,q)$ and $O(p,q)$ of the maximal solvable subalgebras of the Lie algebras $LO(p,q)$ [and the continuous maximal solvable subgroups of $SO(p,q)$]. The cases of greatest physical interest with $p \geq q \geq 0$ and $p+q \leq 6$ are considered in detail (they include the Lorentz group, de Sitter groups, and the conformal group of space-time). Formulas (in terms of Fibonacci numbers) are given for the number of $O(p,q)$ [and $SO(p,q)$] equivalence classes of maximal solvable subalgebras of $LO(p,q)$.

I. INTRODUCTION

The classification of all subgroups of a given group, in particular all Lie subgroups of a given Lie group, is of considerable physical interest for many reasons. Thus, a classification of subgroups of an invariance group of a physical system leads to a classification of possible interactions, breaking the original symmetry, but still preserving some remnants of it. Further, each nonequivalent chain of subgroups of a given group G can be used to provide a different basis for the representation theory of the group G and hence leads, e.g., to different explicit formulas for harmonic analysis on the group. Since one of the important applications of group theory in physics is to provide expansion formulas for physical quantities (scattering amplitudes, form-factors, etc.), it is clearly important to consider different possible bases and thus different expansions systematically.

In a previous article[1] we have presented a general method for finding all conjugacy classes of maximal solvable subgroups of arbitrary semisimple Lie groups and have applied this method to the pseudounitary groups $SU(p,q)$. All semisimple subgroups of complex semisimple Lie groups have been found by Dynkin.[2] More recently the same problem was solved for the real semisimple Lie groups by Cornwell.[3] Since the Levi theorem[4] can be used to decompose an arbitrary Lie algebra into the direct sum of a semisimple algebra and a maximal solvable ideal, it is clear that a classification of all solvable subalgebras of a given algebra represents a crucial step towards classifying all subalgebras. The classification of maximal solvable subalgebras is a necessary and decisive step towards this goal. All these questions were discussed in somewhat greater detail in our previous article.[1]

Finally, let us mention that some of the $O(p,q)$ groups are of great interest in physics. The obvious examples are the $O(2,1)$ group, figuring in Regge pole theory[5] and many other applications, the Lorentz group $O(3,1)$, which is of obvious crucial importance in relativistic physics, the de Sitter groups $O(3,2)$ and $O(4,1)$, playing an important role in relativistic cosmology and elsewhere,[6,7] and finally the conformal group $O(4,2)$, playing an ever increasing role in particle physics, especially at very high energies.[7,8]

Let us note that while all subgroups of $O(2,1)$ and $O(3,1)$ are known,[9] the same cannot be said of any of the other $O(p,q)$ groups. Physical applications of all classes of subgroups of the homogeneous Lorentz group $O(3,1)$ and also the homogeneous Galilei group are discussed, e.g., in a review[10] and lectures.[11]

In Sec. II of this article we present a recursive method, by means of which it is possible to obtain all conjugacy classes of maximal solvable subgroups of a general $SO(p,q)$ group, and to determine their number, their dimension, the dimension of their maximal compact subgroup, etc. The cases of $p \geqslant q \geqslant 0$, $p+q \leqslant 6$ are considered in detail in Sec. III, together with some general results for $q \leqslant 3$, $q \leqslant p < \infty$. Some conclusions and the future outlook are presented in Sec. IV.

II. RECURSIVE PRESCRIPTION FOR MAXIMAL SOLVABLE SUBGROUPS OF SO (p,q)

Any solvable subgroup of a linear group G can be embedded into a maximal solvable subgroup[12] of G, which emphasizes the importance of finding all maximal solvable subgroups of G.

We want to classify the possible structures of the maximal solvable subalgebras S of the Lie algebras

$$LO(p,q)=\{X \mid X\in R^{n\times n} \text{ and } X^T(I_p\oplus -I_q)+(I_p\oplus -I_q)X=0\} \tag{1}$$

associated with the real forms $O(p,q)$ of the orthogonal group $O(p+q,C)$:

$$O(p,q)=\{U \mid U\in R^{n\times n} \text{ and } U^T(I_p\oplus -I_q)U=I_p\oplus -I_q\}, \tag{2}$$

where p and q are two integers satisfying

$$p\geqslant q\geqslant 0, \quad n=p+q, \tag{3}$$

and the superscript T indicates a transposed matrix. We shall consider equivalence classes of such subgroups S with respect to conjugacy under $O(p,q)$ and also under $SO(p,q)$ [transformations U satisfying $\det U = 1$, in addition to (2)].

We shall proceed recursively, i.e., represent the maximal solvable subgroups of $O(p,q)$ in terms of those of $O(p-1,q-1)$ or $O(p-2,q-2)$.

We state our main result in the form of a theorem, which we then prove.

Theorem 1: Nonconjugate maximal solvable subalgebras of $LO(p,q)$ can be obtained in three different ways:

(i) If p and q are not both odd then $LO(p,q)$ has a Cartan subalgebra (which is Abelian and hence solvable) consisting only of elements that generate $O(p,q)$ transformations, contained in the maximal compact subgroup $O(p)\times O(q)$. They form a maximal solvable subalgebra of dimension $[p/2]+[q/2]$ ($[x]$ denotes the entire part of x). A general element of this Cartan algebra can be written as

$$S=\sum_{i=1}^{[p/2]} X_i(E_{2i-1,2i}-E_{2i,2i-1}) + \sum_{j=1}^{[q/2]} Y_j(E_{p+2j-1,p+2j}-E_{p+2j,p+2j-1}), \tag{4}$$

where X_i and Y_j are real numbers, E_{ik} are $n\times n$ matrices with all elements equal to zero, except for an entry 1 on the intersection of the ith row and kth column.

(ii) One (for $q=1$) or two (for $q>1$) further types of maximal solvable subalgebras of $LO(p,q)$ are obtained as follows. We perform a similarity transformation $Z^{-1}LO(p,q)Z$ to a new realization of $LO(p,q)$, where Z is a nonsingular matrix, satisfying

$$Z^{-1}(I_p\oplus -I_q)Z=\begin{pmatrix} 0 & 0 & D_\alpha \\ 0 & I_{p-\alpha}\oplus -I_{q-\alpha} & 0 \\ D_\alpha & 0 & 0 \end{pmatrix} = D^\alpha_{pq},\quad \alpha=1 \text{ or } 2,\quad \alpha\leqslant q, \tag{5}$$

with

$$D_1=1,\quad D_2=\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{6}$$

The maximal solvable subalgebras can be represented as

$$S^\alpha_{pq}=\begin{pmatrix} X^\alpha_{11} & X^\alpha_{12} & X^\alpha_{13} \\ 0 & X^\alpha_{22} & X^\alpha_{23} \\ 0 & 0 & X^\alpha_{33} \end{pmatrix}, \tag{7}$$

where

$$(S^\alpha)^T D^\alpha_{pq} + D^\alpha_{pq} S^\alpha = 0, \tag{8}$$

so that we have

(a) $\alpha=1$: $X_{11}=a,\quad X_{12}=(b_1,\ldots,b_{p+q-2}),\quad X_{13}=0,$
($a, b_1,\ldots,b_{p+q-2}$ are arbitrary real numbers).

$$X_{23}=-(I_{p-1}\oplus -I_{q-1})X_{12}^T,\quad X_{33}=-a, \tag{9}$$

and X_{22} is a representation of a maximal solvable subalgebra of $LO(p-1,q-1)$ such that

$$X_{22}^T(I_{p-1}\oplus -I_{q-1})+(I_{p-1}\oplus -I_{q-1})X_{22}=0.$$

(b) $\alpha=2$: $X_{11}=\begin{pmatrix} a & b \\ -b & a \end{pmatrix},\quad X_{33}=-X_{11},\quad X_{13}=\begin{pmatrix} e & 0 \\ 0 & -e \end{pmatrix},$

$$X_{12}=\begin{pmatrix} c_1, c_2,\ldots, c_{p+q-4} \\ d_1, d_2,\ldots, d_{p+q-4} \end{pmatrix},\quad X_{23}=-(I_{p-2}\oplus -I_{q-2})X_{12}^T D_2,$$

$$X_{33}=-X_{11}^T \tag{10}$$

($a, b, c_1,\ldots,c_{p+q-4}, d_1,\ldots,d_{p+q-4}$ are arbitrary real numbers), and X_{22} is a representation of a maximal solvable subalgebra of $LO(p-2,q-2)$ such that

$$X_{22}^T(I_{p-2}\oplus -I_{q-2})+(I_{p-2}\oplus -I_{q-2})X_{22}=0.$$

All maximal solvable subalgebras S_{pq} are thus expressed in terms of $S_{p-1,q-1}$ and $S_{p-2,q-2}$ or are given by (4). Each algebra obtained in this manner determines an entire conjugacy class under $O(p,q)$ and none of the algebras, given by expressions (4)–(10) are conjugate to each other. This also gives a list of all conjugacy classes under $SO(p,q)$ with the sole exception when $p=q=$even. In this case the one class under $O(p,q)$ obtained by taking $\alpha=2$ at each step, splits into two nonequivalent classes under $SO(p,q)$. A representation S_{pq} of one of these two classes is obtained from (b). It is transformed into a representative $\tilde{\tilde{S}}_{pq}$ of the second conjugacy class by

$$\tilde{\tilde{S}}_{pq}=U^{-1}\tilde{S}_{pq}U,$$

where

$$U=\begin{pmatrix} &&&&&&&&&&1\\ &&&&&&&&&\cdot&\\ &&&&&&&&\cdot&&\\ &&&&&&&\cdot&&&\\ &&&&&&1&&&&\\ &&&&1&0&&&&&\\ &&&&0&1&&&&&\\ &&&\cdot&&&&&&&\\ &&\cdot&&&&&&&&\\ &\cdot&&&&&&&&&\\ 1&&&&&&&&&& \end{pmatrix}.$$

Proof: Let us make the following observations[1]:

1. If $n\leqslant 2$ then L itself is solvable so that $S=L$.

2. If the restriction of the action of L on the natural representation space $R^{n\times 1}$ of L over the real field R to S is irreducible, then by Lie's theorem we have $n\leqslant 2$ so that according to observation 1 we find that $S=L$.

3. As a consequence of Engel's theorem the nilpotent matrices of S form an ideal $N(S,L)$ in S.

4. If there is a decomposition

$$R^{n\times 1}=\sum_{\alpha=1}^{r} R_\alpha \tag{11}$$

of $R^{n\times 1}$ into the direct sum of real irreducible with respect to S linear subspaces R_α (R—S-subspaces) that are orthogonal to each other with respect to the nondegenerate symmetric matrix

$$D_{p,q}=I_p\oplus -I_q$$

such that $x^TDy=0$ if $x\in R_\alpha$, $y\in R_\beta$, $1\leqslant\alpha<\beta\leqslant r$, then we have $N(S,L)=0$ so that S is Abelian and S is contained in a Cartan subalgebra of L. Because of the maximal property of S we know that S itself must be a Cartan subalgebra. In other words, S is a compact Cartan subalgebra of L. All compact Cartan subalgebras of the Lie algebra $LO(p,q)$ are conjugate under the group $SO(p,q)$. They exist precisely if p and q are not both odd numbers.

Conversely, every compact Cartan subalgebra of L is a maximal solvable subalgebra; and up to conjugacy under $SO(p,q)$ we find that S is given by (4).

5. If no orthogonal decomposition of $R^{n\times 1}$ of the form (11) is possible, then there is an irreducible linear R—S-subspace m which is D-isotropic: $x^TDx=0$ for all x of m. Hence the D-perpendicular subspace $m^\perp=\{y\,|\,y\in R^{n\times 1}$ and $x^TDy=0\}$ is R—S-invariant and it contains m so that $R^{n\times 1}\supset m^\perp\supseteq m\supset 0$; the R—S-irreducibility of m implies by Lie's theorem that the dimension α of m over R is either 1 or 2.

There exists a nonsingular matrix Z_α of degree n over R with row vector u_i $(i=1,2,\ldots,n)$ such that

$$m=\sum_{i=1}^{\alpha} a_i u_i;\quad m^{\perp}=\sum_{i=\alpha+1}^{n} b_i u_i;\quad \alpha=1 \text{ or } 2;\ a_i, b_i \in R, \tag{12}$$

satisfying (5) and transforming a maximal solvable subalgebra S into the form

$$Z_\alpha^{-1} X Z_\alpha=\begin{pmatrix} \Delta_1(X) & X_{12} & X_{13} \\ 0 & \Delta_2(X) & X_{23} \\ 0 & 0 & \Delta_3(X) \end{pmatrix},\quad X\in S, \tag{13}$$

with representations $\Delta_i(x)$ of S over R of degrees α, $n-\alpha$, and α, respectively for $i=1,2$, and 3, such that $\Delta_2(x)$ belongs to $LO(p-\alpha, q-\alpha)$.

The solvability of S implies that the Lie algebras $\Delta_1(S)$ are solvable. The maximal property of S implies that $\Delta_1(S)$ and $\Delta_3(S)$ are maximal solvable subalgebras of $R^{\alpha\times\alpha}$ and that $\Delta_2(S)$ is a maximal solvable subalgebra of $LO(p-\alpha, q-\alpha)$.

The irreducibility of Δ_1 implies that $\Delta_1=R\equiv S_1$ (i.e., a real number) if $\alpha=1$, and that for $\alpha=2$ after suitable transformation by a nonsingular matrix we have

$$\Delta_1(S)=R\begin{pmatrix}1&0\\0&1\end{pmatrix}+R\begin{pmatrix}0&1\\-1&0\end{pmatrix}\equiv S_2.$$

Further, the matrix (13), in order to belong to Z_α^{-1} $LO(p,q)Z_\alpha$ must satisfy (8) which leads us to the conditions (9) and (10). Conversely, if $\alpha=1$ or 2 and if $\alpha\leqslant q$ and if $S_{p-\alpha,q-\alpha}$ is a maximal solvable subalgebra of $LO(p-\alpha, q-\alpha)$, then the matrices (7) with X_{11}^{α} of S_α, X_{22}^{α} of $S_{p-\alpha,q-\alpha}$, $X_{33}^{\alpha}=-(X_{11}^{\alpha})^T$ and X_{12}^{α}, X_{13}^{α} and X_{23}^{α} as in (9) or (10) form a maximal solvable subalgebra of the Lie algebra $Z_\alpha^{-1}LO(p,q)Z_\alpha$, where

$$Z_1=\begin{pmatrix} 1 & 0 & 1 \\ 0 & I_{n-2} & 0 \\ -1 & 0 & 1 \end{pmatrix} \text{ and } Z_2=\begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & & 1 & 0 \\ & 0 & I_{n-4} & 0 & \\ 0 & -1 & & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{14}$$

In order to verify the maximality of S it is sufficient to refer to the general criterion given in the Theorem of our previous paper.[1]

Finally, let us show that all the different subalgebras constructed according to the Theorem 1 are inequivalent, i.e., that each different sequence of numbers $\alpha_1,\ldots,\alpha_r$ leads to a different class of subalgebras. To do this it is sufficient to show that an R—S-invariant subspace m cannot be transformed by an $O(p,q)$ transformation into any different R—S-invariant subspace $m'\neq m$.

If m is a minimal nonzero R—S-invariant linear subspace of the representation space $R^{n\times 1}$ of S over R, then, of course, it is invariant under the nilpotent subalgebra S_0 formed by all matrices

$$\begin{pmatrix} 0_\alpha & X_{12} & 0_\alpha \\ & 0_{n-2\alpha} & \tilde{X}_{12} \\ & & 0_\alpha \end{pmatrix} \tag{15}$$

with $X_{12}\in R^{\alpha\times(n-2\alpha)}$ and $\tilde{X}_{12}=-I_{p-\alpha,q-\alpha}X_{12}^{T}D_\alpha$. Because of the irreducibility property of m as a representation space of the solvable Lie algebra S over R, it follows from Lie's theorem that the nilpotent subalgebra S_0 annihilates m. But the only n-columns annihilated by S_0 are the linear combinations of the first α unit n-columns over R. Hence m is uniquely determined.

These remarks lead to an inductive construction of all maximal solvable subalgebras of $LO(p,q)$ up to conjugacy under $O(p,q)$. There are as many classes of conjugacy under $O(p,q)$ as there are finite sequences $\alpha_1,\ldots,\alpha_r$ of nonnegative integers α_i for which

(1) $\alpha_i=1$ or 2, $0\leqslant i\leqslant r$;

(2) $\sum_{i=1}^{r}\alpha_i\leqslant q$;

(3) $(n+1)\left(p+\sum_{i=1}^{r}\alpha_i\right)$ is even.

Conjugacy under the subgroup $SO(p,q)$ of index 2 in $O(p,q)$ will be tantamount to conjugacy under $O(p,q)$ if a matrix of determiant -1 can be found in $O(p,q)$ which transforms S into itself ("normalizes" S). This is because that matrix can be multiplied into any other transforming matix if necessary to produce a transforming matrix of determinant 1. However, if all matrices of $O(p,q)$ normalizing S are of determinant 1, then the conjugacy class under $O(p,q)$ splits into two conjugacy classes under $SO(p,q)$. This will happen precisely if $p=q$ is even and if we deal with the case in which $r=p/2$, $\alpha_1=\alpha_2=\ldots=\alpha_r=2$.

III. PROPERTIES OF THE MAXIMAL SOLVABLE SUBALGEBRAS AND EXAMPLES

A. Number of maximal solvable subalgebras and their dimensions

The recursive procedure for constructing maximal solvable subalgebras of $LO(p,q)$ immediately provides us with recursion relations for the number of nonconjugate solvable subalgebras N_{pq}, their dimensions d_{pq}^{α} and the dimensions c_{pq}^{α} of their maximal compact subalgebras. These recursion relations can actually be solved, but we do this only for the number of algebras N_{pq}, since the expressions for d_{pq}^{α} and c_{pq}^{α} are not very illuminating.

Theorem 2: The number N_{pq} of maximal solvable subalgebras of $LO(p,q)$ that are not conjugate under $O(p,q)$ is equal to

$$N_{pq}=-1+3F_q+2F_{q-1}\quad \text{for } n=p+q \text{ odd} \tag{16a}$$

$$=\tfrac{1}{2}[(-1)^q-1]+2F_q+F_{q-1}\quad \text{for } n=p+q \text{ even}, \tag{16b}$$

where

$$F_q=\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^q-\left(\frac{1-\sqrt{5}}{2}\right)^q\right] \tag{17}$$

are the Fibonacci numbers.[13] (Notice that aside from the even or odd condition on $p+q$, N_{pq} depends on q only).

The same formulas hold when equivalence under $SO(p,q)$ is considered, with the exception of $LO(p,p)$; $p=$ even, when

$$\tilde{N}_{pp}=\tfrac{1}{2}[(-1)^p+1]+2F_p+F_{p-1}. \tag{18}$$

Proof: It follows directly from Theorem 1, i.e., formulas (9) and (10), that N_{pq} satisfies the recursion relation

$$N_{pq}=N_{p-1,q-1}+N_{p-2,q-2}+1 \quad \text{for } n=p+q \text{ odd} \qquad (19a)$$

$$=N_{p-1,q-1}+N_{p-2,q-}+\epsilon_q \quad \text{for } n=p+q \text{ even}, \qquad (19b)$$

where

$$\epsilon_q=\begin{cases}0 & \text{for } q \text{ odd}\\ 1 & \text{for } q \text{ even.}\end{cases}$$

Let us first consider n odd. Putting $M_q=N_{pq}+1$, we find

$$M_q=M_{q-1}+M_{q-2} \qquad (20)$$

which is the recursion relation for the Fibonacci numbers.[13] Hence we can write the solution of (19a) as

$$N_{pq}=-1+\lambda_1F_q+\lambda_2F_{q-1}. \qquad (21)$$

To find λ_1 and λ_2 we must calculate N_{pq} for two values of q independently. For $q=0$ (the compact case) we obviously have $N_{p0}=1$, for $q=1$ and p even, e.g., $LO(2,1)$, we have $N_{p1}=2$. Since $F_{-1}=F_1=F_2=1$ and $F_0=0$, we find $\lambda_1=3$ and $\lambda_2=2$, thus proving formula (16a).

For n even we apply (19b) twice to obtain

$$N_{pq}=2N_{p-2,q-2}+N_{p-3,q-3}+\epsilon_q+\epsilon_{q-1}=2N_{p-2,q-2}+N_{p-3,q-3}+1. \qquad (22)$$

Putting $M_q=N_{pq}+\frac{1}{2}$, we find

$$M_q=2M_{q-2}+M_{q-3} \qquad (23)$$

We search for basic solutions of (23) in the form

$$M_q=\alpha^q. \qquad (24)$$

Substituting (24) into (23) we find three solutions for α:

$$\alpha_1=-1, \quad \alpha_{2,3}=(1\pm\sqrt{5})/2.$$

The general solution of (22) can hence be written as

$$N_{pq}=-\tfrac{1}{2}+\lambda_1(-1)^q+\lambda_2F_q+\lambda_3F_{q-1}. \qquad (25)$$

To find λ_1, λ_2, and λ_3 we need three values of N_{pq}. For $q=0$, p even, we again have $N_{p0}=1$; for $q=1$, $p=$odd [e.g., $LO(1,1)$ or $LO(3,1)$], we have $N_{p1}=1$; and for $q=2$, p even (e.g., $LO(2,2)$], we have $N_{p2}=3$. Putting successively $q=0,1$, and 2 in (25), we obtain three equations for λ_i, giving $\lambda_1=\frac{1}{2}$, $\lambda_2=2$, and $\lambda_3=1$, thus proving (16b).

It was shown in our last comments of the previous section that if $p=q=$ even, then one of the $O(p,q)$ classes of maximal solvable subalgebras splits into two $SO(p,q)$ classes. This gives formula (18) and completes our proof of Theorem 2.

Let us just give the recursion relations for the dimension d^{α}_{pq} of S_{pq} and the number of compact elements c^{α}_{pq}, which follow directly from (9) and (10).

Indeed, if $\alpha=1$, i.e., S_{pq} is obtained from $S_{p-1,q-1}$ as indicated by Eq. (9), then we have

$$d^1_{pq}=d_{p-1,q-1}+p+q-1,$$
$$c^1_{pq}=c_{p-1,q-1}. \qquad (26)$$

If $\alpha=2$, i.e., S_{pq} is obtained from $S_{p-2,q-2}$ as indicated by Eq. (10), we have

$$d^2_{pq}=d_{p-2,q-2}+2(p+q)-5,$$
$$c^2_{pq}=c_{p-2,q-2}+1. \qquad (27)$$

For the compact case (4) we have

$$d^c_{pq}=c^c_{pq}=\left[\frac{p+q}{2}\right], \quad p\cdot q=\text{even}. \qquad (28)$$

The relations (26) and (27) can obviously be solved but the solutions depend on the number of steps of type $\alpha=1$ and $\alpha=2$ and we shall not go into this here. Instead, let us give some examples of physical interest.

B. Examples

As mentioned in the Introduction, the $O(p,q)$ groups of greatest physical interest are those with $p+q\leqslant 6$, $p\geqslant q\geqslant 0$. Some properties of the corresponding maxi-

TABLE I. Some properties of the maximal solvable subalgebras S_{pq} of $LO(p,q)$ for $p\geq q\geq 1$, $2\leq p+q\leq 6$.

Algebra	Number N_{pq} of conjugacy clases of subalgebras S_{pq} under $O(p,q)$	Dimension d_{pq} of S_{pq}	Number c_{pq} of compact elements of S_{pq}
$LO(1,1)$	1	1	0
$LO(2,1)$	2	1	1
		2	0
$LO(3,1)$	1	4	1
$LO(2,2)$	3	2	2
		3	1
		4	0
$LO(3,2)$	4	2	2
		5	1
		5	1
		6	0
$LO(4.1)$	2	2	2
		5	1
$LO(4,2)$	3	3	3
		8	2
		9	1
$LO(3,3)$	4	7	2
		8	1
		8	1
		9	0
$LO(5,1)$	1	7	2

TABLE II. Some properties of the maximal solvable subalgebras S_{pq} of $LO(p,q)$ for $1\le q\le 3$, $q\le p<\infty$.

Algebra		Number N_{pq} of conjugacy classes of subalgebras S_{pq} under $O(p,q)$	Dimension d_{pq} of S_{pq}	Number c_{pq} of compact elements of S_{pq}
$LO(p,1)$	p odd	1	$(3p-1)/2$	$(p-1)/2$
	p even	2	$p/2$ $(3p-2)/2$	$p/2$ $(p-2)/2$
$LO(p,2)$	p odd	4	$(p+1)/2$ $(3p+1)/2$ $(5p-5)/2$ $(5p-3)/2$	$(p+1)/2$ $(p-1)/2$ $(p-1)/2$ $(p-3)/2$
	p even	3	$(p+2)/2$ $(5p-4)/2$ $(5p-2)/2$	$(p+2)/2$ $p/2$ $(p-2)/2$
$LO(p,3)$	p odd	4	$(3p+5)/2$ $(7p-5)/2$ $(7p-5)/2$ $(7p-3)/2$	$(p+1)/2$ $(p-1)/2$ $(p-1)/2$ $(p-3)/2$
	p even	7	$(p+2)/2$ $(3p+4)/2$ $5p/2$ $(5p-6)/2$ $(5p+2)/2$ $(7p-6)/2$ $(7p-4)/2$	$(p+2)/2$ $p/2$ $p/2$ $(p-2)/2$ $(p-2)/2$ $(p-2)/2$ $(p-4)/2$

mal solvable subalgebras S_{pq} are summarized in Table I. The same properties of $O(p,q)$ groups for $1\leqslant q\leqslant 3$, $q\leqslant p<\infty$ are given in Table II. In Table III we list the number N_{pq} of equivalence classes under $O(p,q)$ of maximal solvable subalgebras of $LO(p,q)$ as a function of q. Naturally, there is some overlap between the three tables. The numbers N_{pq} can be checked using formulas (16) and (17).

It is of interest to note that the complex extension of the $LO(p,q)$ algebras, namely $LO(n,C)$ with $n=p+q$ has only one maximal solvable subalgebra up to conjugacy under $O(n,C)$, namely the Borel subalgebra.[14] This can be written in the form

$$B(n,C)=\begin{pmatrix}
\alpha_{11}, & \alpha_{12}, & \alpha_{13},\ldots,\alpha_{1n-2}, & \alpha_{1n-1}, & 0\\
0, & \alpha_{22}, & \alpha_{23},\ldots,\alpha_{2n-2}, & 0 & -\alpha_{1n-1}\\
0, & 0, & \alpha_{33},\ldots,0 & -\alpha_{2n-2}, & -\alpha_{1n-2}\\
\cdots & \cdots & \cdots & \cdots & \cdots\\
 & & -\alpha_{33}, & -\alpha_{23}, & -\alpha_{13}\\
 & 0 & 0, & -\alpha_{22}, & -\alpha_{12}\\
 & & 0, & 0, & -\alpha_{11}
\end{pmatrix} \tag{29}$$

where α_{ik} are arbitrary complex numbers. The complex dimension of $B(n,C)$ is obviously

$$\dim B(n,C)=\begin{cases} n^2/4 & \text{for } n \text{ even} \\ (n^2-1)/4 & \text{for } n \text{ odd.}\end{cases} \tag{30}$$

For $LO(p,p)$, $LO(p,p-1)$, and $LO(p,p-2)$ [and for no other $LO(p,q)$ algebras] the largest of the maximal solvable subalgebras have the real dimension given by (30) and their complex extensions are conjugate to the Borel subalgebra (29).

For practical applications it may be of use to have the maximal solvable subalgebras of $LO(p,q)$ explicitly in matrix form. The results can be obtained directly using the recursive procedure of Theorem 1. Let us, however, spell out a few cases of special interest explicitly. We shall give the $(p+q)$-dimensional matrix representations of some of the algebras S_{pq} and also the matrix D_{pq}, so as to show which form of $LO(p,q)$ we find convenient to use. Remember that the $LO(p,q)$ matrices satisfy

$$X^TD+DX=0 \tag{31}$$

and that a nonsingular matrix Z can always be found, transforming one realization of $LO(p,q)$ into another, so that

$$Z^{-1}XZ=\tilde{X}. \tag{32}$$

The symbol S^c denotes a compact (and hence Abelian) subalgebra. Consider individual algebras:

$O(1,1)$: One-dimensional Lorentz transformations.

$$S_{11}^1=\begin{pmatrix} a & 0\\ 0 & -a\end{pmatrix}\equiv LO(1,1),\quad D=\begin{pmatrix} 0 & 1\\ 1 & 0\end{pmatrix}.$$

$O(2,1)$. The Lorentz group in two spacelike and one timelike dimensions.

$$S_{21}^c=\begin{pmatrix} 0 & a & 0\\ -a & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\quad D=\begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & -1\end{pmatrix},$$

$$S_{21}^1=\begin{pmatrix} a & b & 0\\ 0 & 0 & -b\\ 0 & 0 & -a\end{pmatrix},\quad D^1=\begin{pmatrix} 0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix}.$$

TABLE III. Number N_{pq} of conjugacy classes under $O(p,q)$ of maximal solvable subalgebras of $LO(p,q)$.

q	0	1	2	3	4	5	6	7	8	9	10
N_{pq}, p odd	1	1	4	4	12	12	33	33	88	88	232
N_{pq}, p even	1	2	3	7	8	20	21	54	55	143	144

$O(3,1)$. The homogeneous Lorentz group.

$$S_{31}^1=\begin{pmatrix} a & b & c & 0\\ 0 & 0 & d & -b\\ 0 & -d & 0 & -c\\ 0 & 0 & 0 & -a\end{pmatrix},\quad D^1=\begin{pmatrix} 0&0&0&1\\ 0&1&0&0\\ 0&0&1&0\\ 1&0&0&0\end{pmatrix}.$$

This is the four-dimensional solvable algebra found previously (e.g., in Ref. 9). The element d generates a compact subgroup, the other elements noncompact ones. Denoting generators of space rotations about the ith axis L_i and Lorentz "boosts" along the ith axis K_i we can identify the generators as follows:

$$d\sim L_1,\quad a\sim K_1,\quad b\sim L_3-K_2,\quad c\sim L_2+K_3.$$

$O(2,2)\sim O(2,1)\times O(2,1)$:

$$S_{22}^c\begin{pmatrix} 0 & a & 0 & 0\\ -a & 0 & 0 & 0\\ 0 & 0 & 0 & b\\ 0 & 0 & -b & 0\end{pmatrix},\quad D=\begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&-1&0\\ 0&0&0&-1\end{pmatrix},$$

$$S_{22}^1=\begin{pmatrix} a & b & c & 0\\ 0 & 0 & d & -b\\ 0 & d & 0 & c\\ 0 & 0 & 0 & -a\end{pmatrix},\quad D^1=\begin{pmatrix} 0&0&0&1\\ 0&1&0&0\\ 0&0&-1&0\\ 1&0&0&0\end{pmatrix}$$

(S_{22}^1 has no compact elements)

$$S_{22}^2=\begin{pmatrix} a & b & e & 0\\ -b & a & 0 & -e\\ 0 & 0 & -a & -b\\ 0 & 0 & b & -a\end{pmatrix},\quad D^2=\begin{pmatrix} 0&0&0&1\\ 0&0&1&0\\ 0&1&0&0\\ 1&0&0&0\end{pmatrix}$$

(b corresponds to a compact subgroup).

Since $O(2,2)$ satisfies the criterion $p=q=$even the $O(2,2)$ class of subalgebras, represented by S_2 splits into two $SO(2,2)$ classes. Indeed, the $O(2,2)$ matrix O with det $O=-1$

$$O=\begin{pmatrix} 0&0&0&1\\ 0&1&0&0\\ 0&0&1&0\\ 1&0&0&0\end{pmatrix}$$

takes S_2 into

$$\tilde{S}_{22}^2=O^TS_2O=\begin{pmatrix} -a & 0 & b & 0\\ -e & a & 0 & -b\\ -b & 0 & -a & 0\\ 0 & b & e & a\end{pmatrix}$$

which cannot be transformed back into S_2 by an $SO(2,2)$ transformation.

Note also that $O(2,2)$ is locally isomorphic to $O(2,1)\times O(2,1)$ and that the $O(2,2)$ and $O(2,1)$ maximal solvable subalgebras satisfy

$$S_{22}^c\sim S_{21}^c\oplus S_{21}^c,\quad S_{22}^2\sim S_{21}^1\oplus S_{21}^c,$$
$$S_{22}^1\sim S_{21}^1\oplus S_{21}^1,\quad \tilde{S}_{22}^2\sim S_{21}^c\oplus S_{21}^1.$$

$O(4,1)$: One of the de Sitter groups.

$$S_{41}^c=\begin{pmatrix} 0 & a & 0 & 0 & 0\\ -a & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & b & 0\\ 0 & 0 & -b & 0 & 0\\ 0 & 0 & 0 & 0 & 0\end{pmatrix},D=\begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&-1\end{pmatrix},$$

$$S_{41}^1=\begin{pmatrix} a & b & c & d & 0\\ 0 & 0 & 0 & e & -b\\ 0 & 0 & 0 & 0 & -c\\ 0 & -e & 0 & 0 & -d\\ 0 & 0 & 0 & 0 & -a\end{pmatrix},\quad D^1=\begin{pmatrix} 0&0&0&0&1\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 1&0&0&0&0\end{pmatrix}.$$

(The element e generates a compact subgroup.)

$O(3,2)$: The other de Sitter group.

$$S_{32}^c=\begin{pmatrix} 0 & a & 0 & 0 & 0\\ -a & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & b\\ 0 & 0 & 0 & -b & 0\end{pmatrix},\quad D=\begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&-1&0\\ 0&0&0&0&-1\end{pmatrix}$$

$$S_{32}^1=\begin{pmatrix} a & b & c & d & 0\\ 0 & 0 & e & 0 & -b\\ 0 & -e & 0 & 0 & -c\\ 0 & 0 & 0 & 0 & d\\ 0 & 0 & 0 & 0 & -a\end{pmatrix},\quad D^1=\begin{pmatrix} 0&0&0&0&1\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&-1&0\\ 1&0&0&0&0\end{pmatrix}.$$

(The element e generates a compact subgroup.)

$$S_{32}^2=\begin{pmatrix} a & b & c & d & 0\\ 0 & e & f & 0 & -d\\ 0 & 0 & 0 & -f & -c\\ 0 & 0 & 0 & -e & -b\\ 0 & 0 & 0 & 0 & -a\end{pmatrix},\quad D^2=\begin{pmatrix} 0&0&0&0&1\\ 0&0&0&1&0\\ 0&0&1&0&0\\ 0&1&0&0&0\\ 1&0&0&0&0\end{pmatrix}.$$

(No compact subgroups.)

$$S_{33}^3=\begin{pmatrix} a & b & c & e & 0\\ -b & a & d & 0 & -e\\ 0 & 0 & 0 & -d & -c\\ 0 & 0 & 0 & -a & -b\\ 0 & 0 & 0 & b & -a\end{pmatrix},\quad D^3=D^2.$$

(Here b corresponds to a compact subgroup.)

$O(4,2)$. The conformal group of space–time.

$$S_{42}^{c}=\begin{pmatrix} 0 & a & 0 & 0 & 0 & 0\\ -a & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & b & 0 & 0\\ 0 & 0 & -b & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & c\\ 0 & 0 & 0 & 0 & -c & 0 \end{pmatrix},$$

$$D=\begin{pmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ & & 1 & & & \\ & & & -1 & & \\ & 0 & & & -1 & \\ & & & & & -1 \end{pmatrix},$$

$$S_{42}^{1}=\begin{pmatrix} a & b & c & d & e & 0\\ 0 & f & g & h & 0 & -e\\ 0 & 0 & 0 & j & -g & -c\\ 0 & 0 & -j & 0 & -h & -d\\ 0 & 0 & 0 & 0 & -f & -b\\ 0 & 0 & 0 & 0 & 0 & -a \end{pmatrix},$$

$$D^{1}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(No compact elements.)

$$S_{42}^{2}=\begin{pmatrix} a & b & c & d & g & 0\\ -b & a & e & f & 0 & -g\\ 0 & 0 & 0 & h & -e & -c\\ 0 & 0 & -h & 0 & -f & -d\\ 0 & 0 & 0 & 0 & -a & -b\\ 0 & 0 & 0 & 0 & b & -a \end{pmatrix},\quad D^{2}=D^{1}$$

(The element b corresponds to the compact subgroup.) The groups $O(3,3)$, $O(5,1)$, etc. can be considered quite analogously, but we shall stop here.

IV. CONCLUSIONS

In this article we have solved the problem of finding all maximal solvable subalgebras of $LO(p,q)$ up to conjugacy under $O(p,q)$ and $SO(p,q)$. The result is summarized in the recursive procedure of Theorem 1 in Sec. II. A surprisingly large number N_{pq} of conjugate maximal solvable subalgebras exists for general $LO(p,q)$-algebras. Indeed, formulas (16) and (17) show that N_{pq} grows exponentially with q (for $p \geq q \gg 1$). This is to be compared with the fact that the complex extension of $LO(p,q)$, namely $LO(n,C)$ with $n=p+q$ has just one maximal solvable subalgebra [the Borel subalgebra (29)].

We see that the situation is much more complex than for the pseudounitary groups $SU(p,q)$ $(p \geq q \geq 0)$ where precisely $q+1$ maximal solvable subalgebras exist. Thus, e.g., for $q=10$, p odd the algebra $LO(p,q)$ has 232 nonconjugate maximal solvable subalgebras, whereas $LSU(p,q)$ has only 11.

The methods of this article and the previous one[1] can readily be applied to find the maximal solvable subalgebras of an arbitrary semisimple algebra (over an arbitrary field). We plan to return to other real forms of the Cartan semisimple algebras (the classical algebras and possibly also the exceptional ones) in a future publication.

As mentioned in the Introduction, a classification of the maximal solvable subalgebras of a given Lie algebra is a crucial step towards classifying all subalgebras. Work, making use of the results of this paper and the previous one,[1] on the classification of all continuous subgroups of the conformal group $SU(2,2)\sim O(4,2)$ is in progress.

[1] J. Patera, P. Winternitz, and H. Zassenhaus, J. Math. Phys. 15, 1378 (1974).

[2] E. B. Dynkin. Mat. Sbornik. 30, 349 (1952); Trudy Mosk. Mat. Obschch. 1, 39 (1952) [Am. Math. Soc. Transl. Ser. 2, Vol. 6, 111 and 245 (1957)].

[3] J. F. Cornwell, Reports on Math. Phys. 2, 239, 289 (1971) and 3, 91 (1972); J.M. Ekins and J. F. Cornwell, "Semisimple real subalgebras of non-compact semisimple real Lie algebras IV" (to be published in Reports on Math. Phys.); "Semisimple subgroups of linear semisimple Lie groups" (to be published in J. Math. Phys.); and "Semisimple real subalgebras of non-compact semisimple real Lie algebras V", Preprint 1973.

[4] N. Jacobson, *Lie Algebras* (Wiley, New York, 1962).

[5] P. D. B. Collins and E. J. Squires, *Regge Poles in Particle Physics* (Springer, Berlin, 1968).

[6] T. O. Philips and E. P. Wigner, in *Group Theory and its Applications*, edited by E. M. Loebl (Academic, New York, 1968), Vol. 1.

[7] *Lectures in Theoretical Physics*, *De Sitter and Conformal Groups and Their Applications*, edited by A. O. Barut and W. E. Brittin (Colorado Assoc. U. P., Boulder, Colorado, 1971), Vol. XIII.

[8] H. Kastrup, Ann. Phys. 7, 388 (1962). T. Fulton, F. Rohrlich, and L. Witten. Rev. Mod. Phys. 34, 462 (1962); L. Castell, Nucl. Phys. B 4, 343 (1967).

[9] P. Winternitz and I. Friš, Yad. Fiz. 1, 889 (1965) [Sov. J. Nucl. Phys. 1, 636 (1965)].

[10] E. G. Kalnins, J. Patera, R. T. Sharp, and P. Winternitz, "Elementary Particle Reactions and the Lorentz and Galilei Groups" in *Group Theory and its Applications*, edited by E. M. Loebl (Academic, New York), Vol. 3 (to be published).

[11] P. Winternitz, *Expansions of Relativistic Scattering Amplitudes and Harmonic Analysis on the Lorentz Group*, Parts 1, 2, and 3, University of Arizona 1970-71 Lecture Series, edited by D. E. Myers.

[12] H. Zassenhaus, Hamb. Abh. 12, 289 (1938).

[13] D. E. Knuth, *Fundamental Algorithms. The Art of Computer Programming.* (Addison-Wesley, Reading, Mass., 1968), Vol. 1.

[14] A. Borel, Ann. Math. 64, 20 (1956).

Invariant imbedding and Fredholm integral equations with degenerate kernels

S. Ueno

Department of Information Science, Kanazawa Institute of Technology, Ogigaoka Nonoichi, Ishikawa, Japan

(Received 4 February 1974)

In a manner similar to that given in preceding papers by Bellman and Ueno, with the aid of the Bellman–Krein formula for the resolvent, we show how to solve Fredholm integral equations of the second kind with degenerate kernels. The standard procedure for solution is to convert it into an equivalent matrix equation, but in this paper it is transformed into a Cauchy problem which can be solved effectively by high speed digital computers.

1. INTRODUCTION

In mathematical physics, neutron transport, radiative transfer, rarefied gas dynamics, and biomathematics we deal frequently with Fredholm integral equations of the second kind. There exists a class of Fredholm equations which are solved by reducing to a system of algebraic equations. The kernel is called degenerate if it is a finite sum of terms, each term being product of two factors, one of which depends on t, and the other only on y. The standard procedure in solving Fredholm equations with degenerate kernels is to transform them into a system of linear algebraic equations, containing a number of integrals (cf. Courant and Hilbert[1]).

In a series of preceding papers (cf. Bellman and Ueno,[2-6] hereinafter referred to as Papers I–V, respectively) we showed how to solve Fredholm (or Volterra) equations with the aid of invariant imbedding. In those papers it is shown that the use of a Bellman–Krein-like formula for the resolvent permits us to convert the two-point boundary value problem into an initial value problem. In recent years several authors have applied invariant imbedding to the solution of Fredholm equations with degenerate kernels (cf. Kagiwada, Kalaba, Schumitzky, and Ueno[7]; Kagiwada, Kalaba, and Ueno[8]; Kalaba and Vereeke[9]; Kagiwada and Kalaba[10]; Kalaba and Zagustin[11]; Leong and Sen[12]; Bellman and Ueno[5]). In this short paper, we present an alternative approach to the solution of Fredholm equations with degenerate kernels. Making use of the Bellman–Krein formula for the resolvent, we show how to convert the two-point boundary value problem. The Cauchy system obtained makes tractable the numerical computation by high-speed digital computers. Another method for this problem is presented in a preceding paper (cf. Ueno[13]).

2. FREDHOLM INTEGRAL EQUATIONS WITH DEGENERATE KERNELS

Consider the Fredholm integral equation

$$u(t)=g(t)+\int_0^x k(t,y)\,u(y)\,dy, \tag{1}$$

where $0\leq t\leq x$, $g(t)$ is a given forcing function, and the kernel $k(t,y)$ is a finite sum of continuous functions of the form

$$k(t,y)=\sum_{n=1}^{N}\sum_{m=1}^{M} r(n,m)\,a(n;t)\,b(m;y). \tag{2}$$

It is assumed that each of the sets of functions $\{a(n;t)\}$, $\{b(m;y)\}$ are linearly independent square integrable functions in the basic interval $(0,x)$. If $N=M$ and

$$r(n,m)=d(n,m), \tag{3}$$

where $d(n,m)$ is a Kronecker delta, equal to unity for $n=m$ and zero otherwise, we have a Pincherle–Goursat kernel given by

$$k(t,y)=\sum_{n=1}^{N} a(n;t)\,b(n;y). \tag{4}$$

Denote by $u(t,x)$ the solution of the equation

$$u(t,x)=g(t)+\int_0^x k(t,y)\,u(y,x)\,dy, \tag{5}$$

in order to exhibit explicitly the dependence of u function on the integration interval x. Let $K(t,y;x)$ be the resolvent kernel for Eq. (5). Thus, the K function is uniquely determined by the pair of equations

$$K(t,y;x)=k(t,y)+\int_0^x K(t,z;x)\,k(z,y)\,dz \tag{6}$$

and

$$K(t,y;x)=k(t,y)+\int_0^x k(t,z)\,K(z,y;x)\,dz, \tag{7}$$

where $k(t,y)$ is given by Eq. (2). Then, the solution of Eq. (5) is provided by

$$u(t,x)=g(t)+\int_0^x K(t,y;x)\,g(y)\,dy. \tag{8}$$

3. INVARIANT IMBEDDING OF AUXILIARY EQUATIONS

In a manner similar to that given by several authors (cf. Kalaba and Vereeke[9]; Kagiwada and Kalaba[10]; Kalaba and Zagustin[11]; Leong and Sen[12]; Bellman and Ueno[5]; Ueno[13]), we introduce an auxiliary equation given by

$$J(n;t,x)=a(n;t)+\int_0^x k(t,y)\,J(n;y,x)\,dy, \tag{9}$$

where $n=1,2,3,\ldots,N$, and $0\leq t\leq x$. On recalling Eq. (7) for $y=x$, we get

$$\begin{aligned}K(t,x;x)&=k(t,x)+\int_0^x k(t,z)\,K(z,x;x)\,dz\\&=\sum_{n=1}^{N}\sum_{m=1}^{M} r(n,m)\,a(n;t)\,b(m;x)\\&\quad+\int_0^x k(t,z)\,K(z,x;x)\,dz.\end{aligned} \tag{10}$$

Superposition of Eqs. (9) and (10) results in

$$K(t,x;x)=\sum_{n=1}^{N}\sum_{m=1}^{M} J(n;t,x)\,r(n,m)\,b(m;x). \tag{11}$$

Furthermore, Eq. (9) is expressed in terms of the resolvent

$$J(n;t,x)=a(n;t)+\int_0^x K(t,z;x)\,a(n;z)\,dz. \tag{12}$$

On differentiating it with respect to x, we have

$$J_x(n;t,x)=K(t,x;x)\,a(n;x)+\int_0^x K_x(t,z;x)\,a(n;z)\,dz, \tag{13}$$

where the subscript x denotes partial differentiation with respect to x.

With the aid of the Bellman–Krein formula for the resolvent given by (cf. Paper IV)

$$K_x(t,y;x)=K(t,x;x)\,K(x,y;x), \tag{14}$$

where $0\le t,y\le x$, Eq. (13) becomes

$$\begin{aligned}J_x(n;t,x)&=K(t,x;x)\,a(n;x)+\int_0^x K(t,x;x)\,K(x,z;x)\,a(n;z)\,dz\\&=K(t,x;x)\,J(n;x,x)\\&=J(n;x,x)\sum_{i=1}^N\sum_{j=1}^M J(i;t,x)\,r(i,j)\,b(j;x).\end{aligned} \tag{15}$$

On differentiating Eq. (5) with respect to x, we obtain

$$u_x(t,x)=k(t,x)\,u(x,x)+\int_0^x k(t,y)\,u_x(y,x)\,dy. \tag{16}$$

The superposition principle permits us to express $u_x(t,x)$ as

$$u_x(t,x)=K(t,x;x)\,u(x,x), \tag{17}$$

where $K(t,x;x)$ is given by Eq. (11) and

$$\begin{aligned}u(x,x)&=g(x)+\int_0^x k(x,y)\,u(y,x)\,dy\\&=g(x)+\sum_{n=1}^N\sum_{m=1}^M r(n,m)\,a(n;x)\int_0^x b(m;y)\,u(y,x)\,dy.\end{aligned}$$

Putting (18)

$$c(m;x)=\int_0^x b(m;y)\,u(y,x)\,dy, \tag{19}$$

$u(x,x)$ is rewritten in the form

$$u(x,x)=g(x)+\sum_{n=1}^N\sum_{m=1}^M r(n,m)\,a(n;x)\,c(m;x). \tag{20}$$

Differentiation of Eq. (19) with respect to x provides

$$c_x(m;x)=b(m;x)\,u(x,x)+\int_0^x b(m;y)\,u_x(y,x)\,dy \tag{21}$$

On recalling Eqs. (11), (17), and (20), we have

$$\begin{aligned}c_x(m;x)&=b(m;x)\,u(x,x)+\int_0^x K(y,x;x)\,u(x,x)\,b(m;y)\,dy\\&=u(x,x)\Big(b(m;x)+\sum_{i=1}^N\sum_{j=1}^M r(i,j)\,b(j;x)\\&\qquad\times\int_0^x J(i;y,x)\,b(m;y)\,dy\Big)\\&=\Big(g(x)+\sum_{i=1}^N\sum_{j=1}^M r(i,j)\,a(i;x)\,c(j;x)\Big)\\&\quad\times\Big(b(m;x)+\sum_{i=1}^N\sum_{j=1}^M r(i,j)\,b(j;x)\\&\qquad\times\int_0^x J(i;y,x)\,b(m;y)\,dy\Big).\end{aligned} \tag{22}$$

Then, a Cauchy system for $J(n;t,x)$ and $c(m;x)$ is given by Eq. (15) and (22), together with the initial conditions

$$J(n;t,t)=[J(n;x,x)]_{x=t}, \tag{23}$$

$$c(m;0)=0. \tag{24}$$

In what follows, we show how to determine an auxiliary function $J(n;x,x)$. Recalling Eqs. (2) (9), we have

$$\begin{aligned}J(n;x,x)&=a(n;x)+\sum_{i=1}^N\sum_{j=1}^M r(i,j)\,a(i;x)\\&\quad\times\int_0^x b(j;y)\,J(n;y,x)\,dy.\end{aligned} \tag{25}$$

Putting

$$R(j,n;x)=\int_0^x b(j;y)\,J(n;y,x)\,dy, \tag{26}$$

where $j=1,2,3,\ldots,M$ and $n=1,2,3,\ldots,N$, Eq. (25) becomes

$$J(n;x,x)=a(n;x)+\sum_{i=1}^N\sum_{j=1}^M r(i,j)\,a(i;x)\,R(j,n;x). \tag{27}$$

On allowing for Eqs. (15) and (27), differentiation of Eq. (26) with respect to x results in

$$\begin{aligned}R_x(j,n;x)&=b(j;x)\,J(n;x,x)+\int_0^x b(j;y)\,J_x(n;y,x)\,dy\\&=\Big(a(n;x)+\sum_{u=1}^N\sum_{v=1}^M r(u,v)\,a(u;x)\,R(v,n;x)\Big)\\&\quad\times\Big(b(j;x)+\sum_{p=1}^N\sum_{q=1}^M r(p,q)\,b(q;x)\,R(j,p;x)\Big).\end{aligned} \tag{28}$$

Equation (28) is the required initial value differential equation for R function with an initial condition

$$R(j,n;0)=0. \tag{29}$$

When once the R function has been determined by solving Eq. (28), $J(n;x,x)$ is computed by Eq. (27).

4. STATEMENT OF THE CAUCHY SYSTEM

It has been shown that, provided that x_1 is sufficiently small and $0\le t\le x<x_1$, the $R(i,j;x)$-function satisfies a Cauchy system

$$\begin{aligned}R_x(i,j;x)&=\Big(a(j;x)+\sum_{u=1}^N\sum_{v=1}^M r(u,v)\,a(u,x)\,R(v,j;x)\Big)\\&\quad\times\Big(b(i;x)+\sum_{p=1}^N\sum_{q=1}^M r(p,q)\,b(q;x)\,R(i,p;x)\Big),\end{aligned} \tag{30}$$

where $i=1,2,3,\ldots,M$, and $j=1,2,3,\ldots,N$, together with an initial condition

$$R(i,j;0)=0. \tag{31}$$

An initial value differential equation for the J function takes the form

$$J_x(n;t,x)=J(n;x,x)\sum_{i=1}^N\sum_{j=1}^M J(i;t,x)\,r(i,j)\,b(j;x), \tag{32}$$

together with an initial condition at $t=x$

$$J(n;t,t)=[J(n;x,x)]_{x=t},\quad n=1,2,3,\ldots,N. \tag{33}$$

Then, the J function at $t=x$ can be expressed in terms of the R function

$$J(n;x,x)=a(n;x)+\sum_{i=1}^N\sum_{j=1}^M r(i,j)\,a(i;x)\,R(j,n;x). \tag{34}$$

Once J and R functions have been determined by the above procedure, we can compute $c(m;x)$ and $u(t,x)$ by solving a set of differential equations. A Cauchy system for c and u functions takes the form

$$c_x(m;x)=\Big(g(x)+\sum_{i=1}^N\sum_{j=1}^M r(i,j)\,a(i;x)\,c(j;x)\Big)$$

$$\times\left(b(m;x)+\sum_{p=1}^{N}\sum_{q=1}^{M}r(p,q)\,b(q;x)\,R(m,p;x)\right) \tag{35}$$

and

$$u_x(t,x)=\sum_{i=1}^{N}\sum_{j=1}^{M}J(i;t,x)\,r(i,j)\,b(j;x)$$
$$\times\Big[g(x)+\sum_{p=1}^{N}\sum_{q=1}^{M}r(p,q)\,a(p;x)\,c(q;x)\Big], \tag{36}$$

together with initial conditions

$$c(m;0)=0, \tag{37}$$

$$u(t,t)=g(t)+\sum_{i=1}^{N}\sum_{j=1}^{M}r(i,j)\,a(i;t)\,c(j;t), \tag{38}$$

where $0\leqslant t\leqslant x$, and $m=1, 2, 3, \ldots, M$.

[1] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1953).

[2] R. Bellman and S. Ueno, J. Math. Phys. 14, 1489 (1973).

[3] R. Bellman and S. Ueno, Astrophys. Sp. Sci. 16, 241 (1972).

[4] R. Bellman and S. Ueno, University of Southern California, TR. No. 71-44, December 1971; J. Math. Phys. 15, 17 (1974).

[5] R. Bellman and S. Ueno, University of Southern California, TR. No. 72-19, April 1972; J. Math. Anal. Appl. 44, 264 (1973).

[6] R. Bellman and S. Ueno, University of Southern California, TR. No. 72-22, May 1972 (to be published in Utilitas Mathematica).

[7] H.H. Kagiwada, R.E. Kalaba, A. Schumitzky, and S. Ueno, The RAND Corporation, Memorandum, RM-5516-PR, 1967.

[8] H.H. Kagiwada, R.E. Kalaba, and S. Ueno, The RAND Corporation, Memorandum, RM-5599-PR, 1968.

[9] R.E. Kalaba and B.J. Vereeke, The RAND Corporation, Memorandum, RM-5694-PR, 1968.

[10] H. H. Kagiwada and R. Kalaba, Assoc. Compt. Machinery 17, 412 (1970).

[11] R.E. Kalaba and E. Zagustin, J. Franklin Inst. 293, 277 (1971).

[12] T.K. Leong and K.K. Sen, Publ. Astr. Soc., Japan 23, 99 (1971).

[13] S. Ueno, Department of Electrical Engineering, University of Southern California TR. No. RB73-30, 1973.

Kinetic theory and the Lorentz gas

A. Weyland

Institute for Theoretical Physics, University of Nijmegen, The Netherlands

(Received 20 June 1974)

The high density properties of the velocity autocorrelation function and the diffusion coefficient are discussed in a one- and three-dimensional Lorentz gas on the basis of kinetic theory.

In kinetic theory, the use of field theoretical methods in the discussion about the small density (Ref. 1) and the long time behavior (Ref. 2) of the velocity correlation function in hard sphere fluids is well established.

Most recently a detailed analysis of the use of field theory in kinetic theory for hard sphere fluids has been given for all times and densities in an effort to derive a generalized Boltzmann equation (Ref. 3). In the case of the Lorentz gas with point scatterers the application of field theory is nearly straightforward. The basis was given in Ref. 4 and worked out in more detail in Ref. 5.

The interesting thing here is that explicit solutions can be given for rather complicated equations so that the model serves as an ideal testing ground. As a result of the diagram analysis the following Dyson equation was found:

$$(z - i\bar{k}\cdot\bar{v} - nM_{z,\bar{k}}(0))\,G_{z,\bar{k}} = 1 \tag{1}$$

Here $G_{z,\bar{k}}$ is the Laplace and Fourier transform of the one-particle propagator with respect to time and position of the moving particle. The velocity of the moving particle is $\bar{v}$, its radius σ, while the density of scatterers is n. $M_{z,\bar{k}}\ (\bar{\kappa})$ is the generalized collision operator, Fourier-transformed also to the position of the point scatterer (κ). In the case of the Boltzmann equation, the collision operator is the binary collision operator $T(0)$. In a one- and three-dimensional system the solution of the Boltzmann equation is simple. (The one-dimensional case is more interesting than one may think at first glance, as will be explained later on). That is, for $d=1$ and 3 ($d=$dimensionality) one finds (Ref. 5)

$$G^{(B1)}_{z,\bar{k}} = (z_1 - i\bar{k}\cdot\bar{v})^{-1}\{1-(nv)^2/[z_1+(kv)^2]\}^{-1} \times[1+nvP_1(z_1-i\bar{k}\cdot\bar{v})^{-1}], \tag{2}$$

$$G^{(B3)}_{z,\bar{k}} = (z_3 - i\bar{k}\cdot\bar{v})^{-1}\{1+\pi nv\sigma^2[1-\pi n\sigma^2/k\tan^{-1}(kv/z_3)]^{-1} \times P_3(z_3-i\bar{k}\cdot\bar{v})^{-1}\}, \tag{3}$$

where $z_1 = z+nv$, $z_3 = z+\pi nv\sigma^2$, and the projection operator P acts on $\bar{v}$ only. That is, $P_1\bar{v} = -\bar{v}$, $P_3\bar{v} = (4\pi)^{-1}\int d\Omega_v\,\bar{v} = 0$. More generally, $M_{z,\bar{k}}\,(\bar{\kappa})$ satisfies an equation of the form

$$M_{z,\bar{k}}(\bar{\kappa}) = T(\bar{\kappa}) + (2\pi)^{-d}\int d\bar{\kappa}'\,T(\bar{\kappa}-\bar{\kappa}')\,G_{z,\bar{k}+\bar{\kappa}'}M_{z,\bar{k}}(\bar{\kappa}'). \tag{4}$$

The Laplace transform of the velocity autocorrelation function, Φ_z, then follows from

$$\Phi_z = (z+\gamma_z)^{-1}, \quad \gamma_z = -n\hat{v}\cdot M_{z0}(0)\hat{v}/v^2, \tag{5}$$

while the diffusion coefficient is defined as $D = v^2/(d\gamma_0)$. In the Boltzmann limit γ_z is $2\,nv$, $8/3\;nv\sigma$, or $\pi nv\,\sigma^2$ for $d=1$, 2, or 3, respectively.

The question raised here is whether an equation like (4) is powerful enough to describe the velocity autocorrelation function and the diffusion coefficient well enough for all times and densities. Especially the high density limit could be a weak point because, equation (4) does not contain the typical correlations characteristic for a moving particle enclosed in a cage of surrounding scatterers (cage effect). As a consequence of this, the diffusion coefficient should become zero if the density of scatterers is higher than a "critical" one. Following computer experiments (Ref. 6), this happens for a "reduced" density n^* ($n^* = n\sigma^d$) of about 0.4 if $d=2$. For the velocity autocorrelation function this should have some consequences for the long time tail.

The basis for solving (4) will be formed by the so-called ring approximation (Ref. 7). Here one replaces $M_{z,\bar{k}}(\bar{\kappa}')$ in (4) by $T(\bar{\kappa}')$ and uses the Boltzmann propagator.

The results are

$$\gamma^{(R1)}_z = 2nv[1+2nv/(z_1+\sqrt{zz_1})] \quad (d=1), \tag{6}$$

$$\gamma^{(R3)}_z = \pi nv\sigma^2 - \tfrac{1}{8}\pi^2n^2v\sigma^2\int_0^\infty \kappa^2d\kappa f^2(\kappa) \times\{1-\pi n\sigma^2/\kappa\tan^{-1}(\kappa v/z_3)\}^{-1} \quad (d=3), \tag{7}$$

where

$$f(\kappa) = iv\sigma^2\sum_{l=0}(-i)^l(2l+1)[\Gamma(2-l/2)\,\Gamma(5/2)+l/2)]^{-1}j_l'(\kappa\sigma) \times\int_{-1}^{+1}dxP_1(x)\,(z_3-ikvx)^{-1} \tag{8}$$

Here $j'\ (\kappa\sigma)$ is the derivative of a spherical Bessel function and $P_1(x)$ a Legendre polynomial of the first kind. The integral in (8) can easily be performed using the recursion relation for the Legendre polynomials. From this one finds that the velocity autocorrelation function becomes negative within a few mean free times and remains negative with a tail of the form: $t^{-(d/2+1)}$ in accordance with the hydrodynamic approach given in Ref. 8).

For $d=1$, the diffusion coefficient drops down to half its Boltzmann value for all densities, whereas for $d=3$ it is at most two thirds of the corresponding Boltzmann value (infinite density limit). Compared to computer experiments ($d=3$, Ref. 6) and the exact solution ($d=1$) the approximation still contains too little memory effects.

Here one may note that in spite of the fact that the one-dimensional Lorentz gas is very special (only rattling between two neighboring scatterers is possible), the structure and the solutions of the "kinetic equations" are very similar to that of the three-dimensional Lorentz gas. This is so because the binary collision

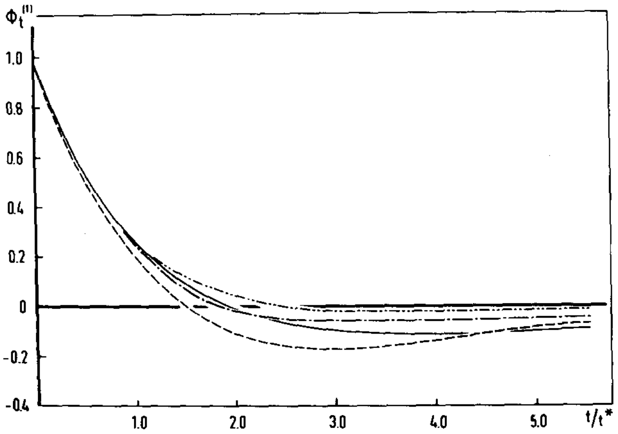


FIG. 1. The velocity autocorrelation function as a function of t/t^*, $t^*=\sigma/2n^*\bar{v}$, $d=1$. ..-.. Ring. -.-.- Superring. - - Exact solution equation (1) and (4) — Exact.

operator contains a real and a virtual part giving the moving particle the possibility to pass a scatterer, even in case $d=1$. For instance, both Boltzmann propagators have the same type of hydrodynamic mode leading to long time tails of the form $t^{-(d/2+1)}$. Another interesting aspect of comparing the one-dimensional with the three-dimensional case is the effect of the anisotropy in the scattering. For $d=1$ the scattering is anisotropic ($P_1\bar{v}=-\bar{v}$) while for $d=3$ the scattering is isotropic ($P_3\bar{v}=0$). This is an important reason why the effect of the ring approximation is larger in case $d=1$ than for $d=3$.

This pays off in the next approximation, the so-called superring approximation, in which one replaces the propagator in (4) by the Boltzmann propagator. For $d=1$ one finds

$$\gamma_z^{(S1)}=2nv(z_1+\sqrt{zz_1})/(z+\sqrt{zz_1}). \qquad (9)$$

This is valid for a point (moving) particle. Otherwise a small oscillation (which can be calculated exactly) is superimposed on the corresponding velocity autocorrelation function.

The diffusion coefficient now drops down to its exact value, namely zero, for all densities. The long time tail of the velocity autocorrelation function remains of the form $t^{-3/2}$, but its coefficient is four times larger than in the ring approximation. This is due to the fact that (9) is the sum of a geometric series, a series that diverges for $z=0$. This shows how Eq. (4) stops diffusion if the anisotropy in the scattering is large enough. For $d=3$, besides the obvious effect of the dimensionality, the scattering is isotropic leading to a diffusion coefficient which drops down to only half of its Boltzmann value if the density of scatterers is infinite. In calculating the velocity autocorrelation function in this approximaton (see Fig. 2) only the first term in (8) was kept. This is not unreasonable as one can see if one compares the corresponding long time tails and high density limits.

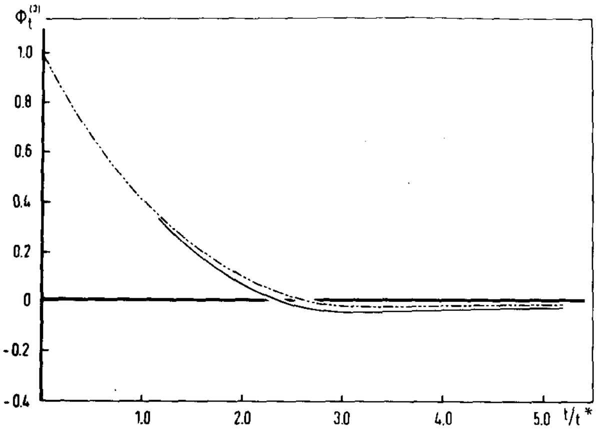


FIG. 2. The velocity autocorrelation function as a function of t/t^*, $t^*=\sigma/\pi n^*\bar{v}$, $d=3$, $n^*=0.3$. .-..-. Ring. — Superring.

For $d=2$, the mathematics is much more complicated (due to the solution of the Boltzmann equation). However, because here the scattering is anisotropic too, one should expect a result somewhere between $d=1$ and $d=3$.

A doubtful argument for this is that knowing the small density expansions of $\gamma_0^{(R2)}$ (Ref. 4) and using it as a term in geometric series in the superring approximation, one finds that diffusion stops at $n^*_c\approx 0.37$! That the argument is doubtful is also due to the fact that it does not work for $d=3$. The exact solution of (1) and (4) for $d=1$ yields a velocity autocorrelation function of the form

$$\Phi_z=(2nv)^{-1}2z_2[2z_2^2+1+(1+4z_2^2)^{1/2}]^{-1}, \qquad (10)$$

where $z_2=z/2nv$.

Its one-particle propagator differs from the Boltzmann propagator only in that the density n has to be multiplied by a function of z, m_z, which has to be determined selfconsistently.

The interesting thing here is that one finds an essential singularity in γ_z for $z=0$, leading to an exponential decay of the velocity autocorrelation function, while the diffusion coefficient remains zero. For $d=3$ the same tric can be used but it will never lead to a critical density above which diffusion stops completely. Whether this is a serious defect of Eq. (4) is not clear because no computer experiments are yet available indicating the value of a critical density in the three-dimensional Lorentz gas.

ACKNOWLEDGMENTS

The author is indebted to Professor J. M. J. van Leeuwen for contributions and stimulating discussions. He also thanks Dr. Ernst, Dr. van Beyeren, and Dr. C. Bruin for helpful discussions.

[1] J.R. Dorfman and E.G.D. Cohen, J. Math. Phys. 8, 282 (1967).
[2] J.R. Dorfman and E.G.D. Cohen, Phys. Rev. Lett. 25, 1257 (1970).
[3] H. van Beyeren, thesis, Physics Department, University of Nijmegen, 1974.
[4] J.M.J. van Leeuwen and A. Weyland, Physica 36, 456 (1967); 38, 35 (1968).
[5] A. Weyland, "The velocity autocorrelation function in the Lorentz gas," internal report 1968.
[6] C. Bruin, preprint, T.H. Delft.
[7] K. Kawasaki and I. Oppenheim, Phys. Rev. 136, A1519 (1964).
[8] M.H. Ernst and A. Weyland, Phys. Lett. A 34, 39 (1971).

Diffraction characteristics of a slit formed by two staggered parallel planes

S. C. Kashyap

Radio and Electrical Engineering Division, National Research Council of Canada, Ottawa, Ontario, Canada

(Received 19 April 1974)

The diffraction of a plane electromagnetic wave by a slit formed by two staggered parallel planes is investigated using an asymptotic Wiener–Hopf technique. By following a standard procedure the problem is formulated in terms of two coupled Wiener–Hopf equations. For large edge–edge separation, the decoupling of the equations is accomplished by evaluating certain integrals by the saddle point method of integration. The results thus obtained can be conveniently identified as rays emanating from the two edges. It is shown that various changes in transmission coefficient and diffraction pattern of a slit can be obtained by changing the angle of stagger of the planes. Plots of transmission coefficients and diffraction patterns are presented for various slit widths and angles of stagger to show these characteristics.

1. INTRODUCTION

The diffraction of a plane electromagnetic wave by a slit in a conducting screen has received great attention for a long time. An exact solution exists in terms of Mathieu functions.[1] Asymptotic diffraction theories also exist, although the interaction term in this case is available only for large slit widths.[2,3] To the author's knowledge, however, no attempt has ever been made to find the properties of a slit formed by two staggered parallel planes. It is the purpose of this paper to find the diffraction characteristics of such a slit for an E-polarized incident plane wave. The approach is based on a boundary value method which leads to two coupled Wiener—Hopf equations. The equations are solved by evaluating certain integrals by the saddle point method of integration. This results in a solution, which for large edge—edge separation is quite similar to that of Karp and Russek,[3] for the case of a wide slit in a screen, except for a few modifications. Finally, the behavior of the slit is illustrated with the aid of diffraction patterns and plots of transmissions coefficients for various values of slit width and angle of stagger of the planes.

2. FORMULATION OF PROBLEM

Consider the slit formed by two staggered parallel planes as shown in Fig. 1. Let an E-polarized plane wave

$$\phi_i = \exp(-ikx\cos\theta_0 - iky\sin\theta_0)\exp(-i\omega t) \tag{1}$$

be incident on the plates. The time dependence $\exp(-i\omega t)$ will hence forth be omitted. Let ϕ_t be the total field and let ϕ be the scattered field such that

$$\phi_t = \phi + \phi_i \tag{2}$$

everywhere. The scattered field satisfies the wave equation

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + k^2\phi = 0, \quad k = k_1 + ik_2, \tag{3}$$

where k is temporarily assumed to have a small positive imaginary part. Let Φ be the Fourier transform of the scattered field ϕ in the x direction; i.e.,

$$\Phi(\alpha, y) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty}\phi(x,y)\exp(i\alpha x)\,dx \tag{4}$$

with $\alpha = \sigma + i\tau$.

The solution of the transformed wave equation takes the form

$$\begin{aligned}\Phi(\alpha, y) &= Ae^{-\gamma y} & y \geqslant 0 && (5a)\\ &= Be^{-\gamma y} + Ce^{\gamma y} & -b \leqslant y \leqslant 0 && (5b)\\ &= De^{\gamma y} & y \leqslant -b && (5c)\end{aligned}$$

where $\gamma^2 = \alpha^2 - k^2$, with the assumption that $\gamma = |\sigma|$ as $\alpha = \sigma \rightarrow \pm\infty$, and $\gamma = -ik$ for $\alpha = 0$. Continuity of electric field at $y=0$ and $y=-b$ yields

$$A = B + C, \tag{6a}$$

$$D = B\exp(2\gamma b) + C. \tag{6b}$$

Let us introduce

$$\Phi_+(\alpha, y) = \frac{1}{(2\pi)^{1/2}}\int_0^{\infty}\phi\exp(i\alpha x)\,dx, \tag{7a}$$

$$\Phi_-(\alpha, y) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{0}\phi\exp(i\alpha x)\,dx, \tag{7b}$$

$$\tilde{\Phi}_+(\alpha, y) = \frac{1}{(2\pi)^{1/2}}\int_l^{\infty}\phi\exp[i\alpha(x-l)]\,dx, \tag{7c}$$

$$\tilde{\Phi}_-(\alpha, y) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{l}\phi\exp[i\alpha(x-l)\,dx, \tag{7d}$$

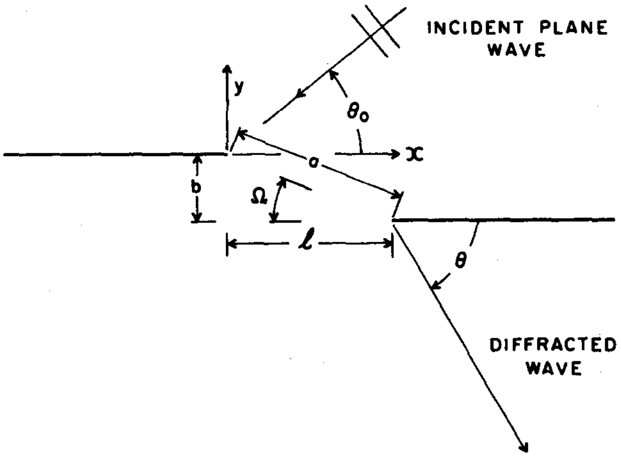


FIG. 1. Configuration of a slit formed by two staggered planes.

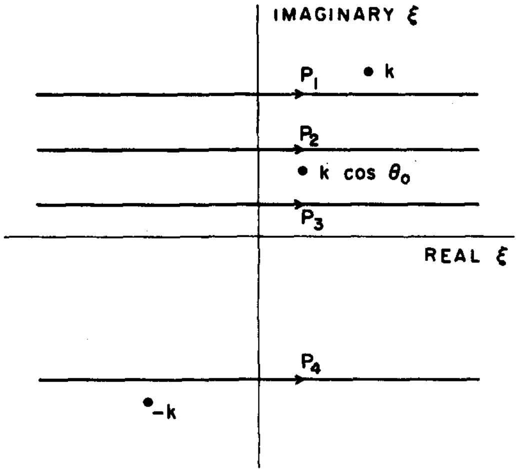


FIG. 2. The integration contours in the complex ξ plane.

where the "+" functions are regular in the upper $\tau > -k_2$ half plane and "−" functions are regular in the lower $\tau < +k_2$ half plane except for the presence of a pole at $\alpha = k\cos\theta_0$ in certain cases (see Fig. 2).

Elimination of A, B, C, D, from (5) and some manipulation yield

$$\Phi_+(\alpha,0)+\frac{i}{(2\pi)^{1/2}(\alpha-k\cos\theta_0)}=-\frac{\exp(i\alpha l)\exp(-\gamma b)\tilde{J}_+(\alpha,-b)}{\gamma}-\frac{J_-(\alpha,0)}{\gamma}, \tag{8}$$

$$\tilde{\Phi}_-(\alpha,-b)-\frac{i\exp[-ika\cos(\theta_0+\Omega)]}{(2\pi)^{1/2}(\alpha-k\cos\theta_0)}=-\frac{\tilde{J}_+(\alpha,-b)}{\gamma}-\frac{\exp(-i\alpha l)\exp(-\gamma b)J_-(\alpha,0)}{\gamma}, \tag{9}$$

where

$$J_-(\alpha,0)=\tfrac{1}{2}[\Phi'_-(\alpha,0^+)-\Phi'_-(\alpha,0^-)]=-\gamma C, \tag{10a}$$

$$\tilde{J}_+(\alpha,-b)=\tfrac{1}{2}[\Phi'_+(\alpha,-b^+)-\tilde{\Phi}'_+(\alpha,-b^-)]$$
$$=-\gamma B\exp(-i\alpha l)\exp(\gamma b). \tag{10b}$$

The prime in the above equations indicates that the transform of $\partial\phi/\partial y$ has been taken. Equations (8) and (9) are the desired Wiener-Hopf equations containing unknowns $\Phi_+(\alpha,0)$, $\tilde{\Phi}_-(\alpha,-b)$, $J_-(\alpha,0)$, and $\tilde{J}_+(\alpha,-b)$. In these equations $\Phi_+(\alpha,0)$ is regular in the $\tau > -k_2$ half plane and $\Phi_-(\alpha,-b)$ is regular in the lower $\tau < +k_2$ half plane. Similarly, $\tilde{J}_+(\alpha,-b)$ and $J_-(\alpha,0)$ are regular for $\tau > -k_2$ and $\tau < +k_2$, respectively, except for a pole at $\alpha = k\cos\theta_0$ in each case (see Fig. 2).

Let us first consider Eq. (9). With the aid of Cauchy's theorem, we can write

$$\frac{\exp(-i\alpha l)\exp(-\gamma b)J_-(\alpha,0)}{(\alpha+k)^{1/2}}=\frac{1}{2\pi i}\int_{P_2}\frac{\exp(-i\xi l)\exp(-\gamma b)J_-(\xi,0)}{(\xi+k)^{1/2}(\xi-\alpha)}d\xi-\frac{1}{2\pi i}\int_{P_1}\frac{\exp(-i\xi l)\exp(-\gamma b)J_-(\xi,0)}{(\xi+k)^{1/2}(\xi-\alpha)}d\xi \tag{11a}$$

$$=T_+(\alpha)+T_-(\alpha) \tag{11b}$$

where the paths P_1 and P_2 are shown in Fig. 2. Thus Eq. (9) can be rewritten as

$$\underset{\text{“−”}}{(\alpha-k)^{1/2}\tilde{\Phi}_-(\alpha,-b)}-\underset{\text{“−”}}{\frac{i\exp[-ika\cos(\theta_0+\Omega)]}{(2\pi)^{1/2}(\alpha-k\cos\theta_0)}\times[(\alpha-k)^{1/2}-(k\cos\theta_0-k)^{1/2}]}+\underset{\text{“−”}}{T_-(\alpha)}$$
$$=-\underset{\text{“+”}}{\frac{\tilde{J}_+(\alpha,-b)}{(\alpha+k)^{1/2}}}-\underset{\text{“+”}}{T_+(\alpha)}+\underset{\text{“+”}}{\frac{i(k\cos\theta_0-k)^{1/2}\exp[-ika\cos(\theta_0+\Omega)]}{(2\pi)^{1/2}(\alpha-k\cos\theta_0)}} \tag{12}$$

where Ω is the angle of stagger of the planes as shown in Fig. 1. With reference to Fig. 2, the "−" sign in Eq. (12) indicates that the terms are regular for $\tau < k_2$ and the "+" sign indicates regularity in $\tau > k_2\cos\theta_0$. On applying the Wiener–Hopf technique[4] and equating both sides of Eq. (12) to zero, we obtain

$$\tilde{J}_+(\alpha,-b)=-T_+(\alpha)(\alpha+k)^{1/2}+\left(\frac{k}{\pi}\right)^{1/2}\times\frac{(\alpha+k)^{1/2}\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{(\alpha-k\cos\theta_0)} \tag{13}$$

Substituting for $\tilde{J}_+(\alpha,-b)$ in (8) and rearranging, we have

$$\underset{\text{“+”}}{\Phi_+(\alpha,0)(\alpha+k)^{1/2}}+\underset{\text{“+”}}{\frac{i}{(2\pi)^{1/2}(\alpha-k\cos\theta_0)}[(\alpha+k)^{1/2}-(k\cos\theta_0+k)^{1/2}]}$$
$$=-\frac{i(k+k\cos\theta_0)^{1/2}}{(2\pi)^{1/2}(\alpha-k\cos\theta_0)}+\frac{\exp(i\alpha l)\exp(-\gamma b)T_+(\alpha)(\alpha+k)^{1/2}}{(\alpha-k)^{1/2}}$$
$$\underset{\text{“−”}}{-\frac{J_-(\alpha,0)}{(\alpha-k)^{1/2}}}\quad\underset{\text{“∓”}}{-\left(\frac{k}{\pi}\right)^{1/2}\times\frac{\sin(\theta_0/2)(\alpha+k)^{1/2}\exp(i\alpha l)\exp(-\gamma b)\exp[-ika\cos(\theta_0+\Omega)]}{(\alpha-k\cos\theta_0)(\alpha-k)^{1/2}}} \tag{14}$$

The "−" sign in the above equation indicates regularity for $\tau < k_2\cos\theta_0$, whereas the "+" sign indicates regularity for $\tau > -k_2$. The terms marked "∓" are regular in neither half plane. They can be split into regular functions using Cauchy's formula. Thus

$$\frac{\exp(i\alpha l)\exp(-\gamma b)(\alpha+k)^{1/2}T_+(\alpha)}{(\alpha-k)^{1/2}}$$

$$=\frac{1}{2\pi i}\int_{P_4}\frac{\exp(i\xi l)\exp(-\gamma b)(\xi+k)^{1/2}T_+(\xi)}{(\xi-k)^{1/2}(\xi-\alpha)}\,d\xi$$

$$-\frac{1}{2\pi i}\int_{P_3}\frac{\exp(i\xi l)\exp(-\gamma b)(\xi+k)^{1/2}T_+(\xi)}{(\xi-k)^{1/2}(\xi-\alpha)}\,d\xi$$

$$=V_+(\alpha)+V_-(\alpha) \qquad (15)$$

and

$$\left(\frac{k}{\pi}\right)^{1/2}$$

$$\times\frac{\sin(\theta_0/2)(\alpha+k)^{1/2}\exp(i\alpha l)\exp(-\gamma b)\exp[-ika\cos(\theta_0+\Omega)]}{(\alpha-k)^{1/2}(\alpha-k\cos\theta_0)}$$

$$=\left(\frac{k}{\pi}\right)^{1/2}\frac{\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{2\pi i}$$

$$\times\int_{P_4}\frac{\exp(i\xi l)\exp(-\gamma b)(\xi+k)^{1/2}}{(\xi-k)^{1/2}(\xi-k\cos\theta_0)}\,d\xi$$

$$-\left(\frac{k}{\pi}\right)^{1/2}\frac{\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{2\pi i}$$

$$\times\int_{P_3}\frac{\exp(i\xi l)\exp(-\gamma b)(\xi+k)^{1/2}}{(\xi-k)^{1/2}(\xi-k\cos\theta_0)}\,d\xi$$

$$=U_+(\alpha)+U_-(\alpha) \qquad (16)$$

where the paths P_3 and P_4 are shown in Fig. 2. Then (14) can be rewritten as

$$\Phi_+(\alpha,0)(\alpha+k)^{1/2}+\frac{i}{(2\pi)^{1/2}}[(\alpha+k)^{1/2}-(2k)^{1/2}\cos(\theta_0/2)]$$
$$-V_+(\alpha)+U_+(\alpha)$$
$$=-i\left(\frac{k}{\pi}\right)^{1/2}\frac{\cos(\theta_0/2)}{(\alpha-k\cos\theta_0)}-\frac{J_-(\alpha,0)}{(\alpha-k)^{1/2}}+V_-(\alpha)-U_-(\alpha). \qquad (17)$$

On applying the Wiener–Hopf technique and equating both sides of Eq. (17) to zero, we obtain

$$J_-(\alpha,0)=-i\left(\frac{k}{\pi}\right)^{1/2}\frac{\cos(\theta_0/2)(\alpha-k)^{1/2}}{(\alpha-k\cos\theta_0)}+V_-(\alpha)(\alpha-k)^{1/2}$$
$$-U_-(\alpha)(\alpha-k)^{1/2}. \qquad (18)$$

3. FAR FIELD OF THE SLIT

Substitution of the expressions obtained for $\tilde{J}_+(\alpha,-b)$ (Eq. 13) and $J_-(\alpha,0)$ in (10) and use of (6b) and (5c) yields the following expression for the transform of the field in the region $y<-b$:

$$\Phi(\alpha,y)=-\left(\frac{k}{\pi}\right)^{1/2}$$

$$\times\frac{\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]\exp(i\alpha l)\exp(\gamma b)\exp(\gamma y)}{(\alpha-k\cos\theta_0)(\alpha-k)^{1/2}}$$

$$+i\left(\frac{k}{\pi}\right)^{1/2}\frac{\cos(\theta_0/2)\exp(\gamma y)}{(\alpha-k\cos\theta_0)(\alpha+k)^{1/2}}$$

$$+\frac{T_+(\alpha)}{(\alpha-k)^{1/2}}\exp(i\alpha l)\exp(\gamma y)+\frac{U_-(\alpha)}{(\alpha+k)^{1/2}}\exp(\gamma y)$$

$$-\frac{V_-(\alpha)}{(\alpha+k)^{1/2}}\exp(\gamma y). \qquad (19)$$

Thus the final solution depends on the evaluation of $U_-(\alpha)$, $T_+(\alpha)$, and $V_-(\alpha)$. Let us first consider $U_-(\alpha)$. Replacing α by $-k\cos\omega$ in (16) for the sake of convenience, we obtain

$$U_-(-k\cos\omega)=-\frac{\sqrt{k}\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{2\pi i\sqrt{\pi}}$$
$$\times\int_{P_3}\frac{(\xi+k)^{1/2}\exp(i\xi l)\exp(-\gamma b)}{(\xi-k)^{1/2}(\xi-k\cos\theta_0)(\xi+k\cos\omega)}\,d\xi. \qquad (20)$$

Isolation of poles at $\xi=k\cos\theta_0$ and $\xi=-k\cos\omega$ results in

$$U_-(-k\cos\omega)=$$
$$-\frac{\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{2\pi i\sqrt{k\pi}(\cos\theta_0+\cos\omega)}$$
$$\times\left(\int_{P_3}\frac{(\xi+k)^{1/2}-(k+k\cos\theta_0)^{1/2}}{(\xi-k\cos\theta_0)(\xi-k)^{1/2}}\exp(i\xi l)\exp(-\gamma b)\,d\xi\right.$$
$$-\int_{P_3}\frac{(\xi+k)^{1/2}-(k-k\cos\omega)^{1/2}}{(\xi+k\cos\omega)(\xi-k)^{1/2}}\exp(i\xi l)\exp(-\gamma b)\,d\xi$$
$$+(k+k\cos\theta_0)^{1/2}\int_{P_3}\frac{\exp(i\xi l)\exp(-\gamma b)}{(\xi-k\cos\theta_0)(\xi-k)^{1/2}}\,d\xi$$
$$\left.-(k-k\cos\omega)^{1/2}\int_{P_3}\frac{\exp(i\xi l)\exp(-\gamma b)}{(\xi+k\cos\omega)(\xi-k)^{1/2}}\,d\xi\right). \qquad (21)$$

The integrands in the first two terms in (21) are smooth functions in the neighborhood of $\xi=k\cos\theta_0$ and $\xi=-k\cos\omega$ and can be evaluated asymptotically in a standard manner. The third and the fourth integrals may be identified with Sommerfeld half plane solutions. Evaluation of these integrals yields

$$U_-(-k\cos\omega)$$
$$=\frac{2\sin(\theta_0/2)\cos(\Omega/2)\exp[-ika\cos(\theta_0+\Omega)]}{\sqrt{k\pi}(\cos\theta_0+\cos\omega)}$$
$$\times\frac{\exp[i(ka-\pi/4)]}{(2\pi ka)^{1/2}}$$
$$\times\left(\frac{\cos(\Omega/2)-\cos(\theta_0/2)}{\cos\Omega-\cos\theta_0}-\frac{\cos(\Omega/2)-\sin(\omega/2)}{\cos\Omega+\cos\omega}\right)$$
$$+\frac{i\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{2\pi\sqrt{k\pi}(\cos\omega+\cos\theta_0)}$$
$$\times\left(\frac{\cos(\theta_0/2)}{\sin(\theta_0/2)}Q(a,\Omega/\pi-\theta_0)-\frac{\sin(\omega/2)}{\cos(\omega/2)}Q(a,\omega/\Omega)\right), \qquad (22)$$

in which

$$Q(\rho,\theta/\theta_0)$$
$$=2\pi^{1/2}\exp(i\pi/4)\left\{\exp[-ik\rho\cos(\theta-\theta_0)]F\left[\sqrt{2k\rho}\cos\left(\frac{\theta-\theta_0}{2}\right)\right]\right.$$
$$\left.+\exp[-ik\rho\cos(\theta+\theta_0)]F\left[\sqrt{2k\rho}\cos\left(\frac{\theta+\theta_0}{2}\right)\right]\right\} \qquad (23)$$

and $F(x)$ is the Fresnel integral

$$F(x)=\int_x^\infty\exp(iu^2)\,du. \qquad (24)$$

Asymptotic evaluation of the Fresnel integrals yields

$$U_-(\alpha)=\frac{i\sqrt{k}\cos(\theta_0/2)\exp(i2kb\sin\theta_0)}{\sqrt{\pi}(k\cos\theta_0-\alpha)}C(\theta_0<\Omega)$$

$$+\frac{\sqrt{k}\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{\sqrt{\pi}(\alpha-k)^{1/2}}$$

$$\times\frac{\exp(i\alpha l)\exp(-\gamma b)C(k\cos\Omega<\alpha)}{(\alpha-k\cos\theta_0)}$$

$$+\frac{\sqrt{k}\,2\cos^2(\Omega/2)\sin(\theta_0/2)\exp[i(ka-\pi/4)]}{\sqrt{\pi}(\cos\Omega-\cos\theta_0)(k\cos\Omega-\alpha)(2\pi ka)^{1/2}} \qquad (25)$$

where

$$C(\theta_1<\theta_2)=1 \quad \text{for } \theta_1<\theta_2$$

$$0 \quad \text{for } \theta_1\geqslant\theta_2. \qquad (26)$$

A similar asymptotic evaluation for $T_+(\alpha)$ and $V_-(\alpha)$ gives

$$T_+(\alpha)$$

$$=-\frac{\sqrt{k}\sin(\theta_0/2)\exp[-ika\cos(\theta_0+\Omega)]}{\sqrt{\pi}(k\cos\theta_0-\alpha)}C(\pi-\Omega<\theta_0)$$

$$-\frac{i\sqrt{k}\cos(\theta_0/2)(\alpha-k)^{1/2}\exp(-i\alpha l)\exp(-\gamma b)}{\sqrt{\pi}(\alpha+k)^{1/2}(\alpha-k\cos\theta_0)}$$

$$\times C(\alpha<-k\cos\Omega)$$

$$+\frac{i\sqrt{k}\,2\cos^2(\Omega/2)\cos(\theta_0/2)\exp[i(ka-\pi/4)]}{\sqrt{\pi}(\cos\Omega+\cos\theta_0)(k\cos\Omega+\alpha)(2\pi ka)^{1/2}}$$

$$+\frac{(\alpha-k)^{1/2}[V_-(\alpha)-U_-(\alpha)]\exp(-i\alpha l)\exp(-\gamma b)}{(\alpha+k)^{1/2}}$$

$$\times C(\alpha<-k\cos\Omega)$$

$$+\frac{2k\cos^2(\Omega/2)[V_-(-k\cos\Omega)-U_-(-k\cos\Omega)]\exp[i(ka-\pi/4)]}{(k\cos\Omega+\alpha)(2\pi ka)^{1/2}}, \qquad (27)$$

$$V_-(\alpha)=\frac{T_+(\alpha)(\alpha+k)^{1/2}\exp(i\alpha l)\exp(-\gamma b)}{(\alpha-k)^{1/2}}C(k\cos\Omega<\alpha)$$

$$+\frac{2k\cos^2(\Omega/2)\,T_+(k\cos\Omega)\exp[i(ka-\pi/4)]}{(k\cos\Omega-\alpha)(2\pi ka)^{1/2}}. \qquad (28)$$

Note that $T_+(k\cos\Omega)$ and $V_-(-k\cos\Omega)$ are not considered as unknowns. They may be obtained by setting $\alpha=k\cos\Omega$ and $\alpha=-k\cos\Omega$ in the solution of $T_+(\alpha)$ and $V_-(\alpha)$, respectively, once the latter are known.

Substituting for $T_+(\alpha)$, $U_-(\alpha)$, and $V_-(\alpha)$ in (19) and taking the inverse transform, the transmitted far field may be written as

$$\phi=\phi_{\text{sep}}+\phi_{\text{int}}, \qquad (29)$$

where ϕ_{sep} is the field diffracted by the structure when the two planes are isolated from each other. Evaluating the integrals involved in the inverse transformation of (19) by the method of steepest descent, ϕ_{sep} is recognized to be

$$\phi_{\text{sep}}=\frac{\exp[i(kr-\pi/4)]}{(2\pi kr)^{1/2}}\left(-\frac{2i\cos(\theta_0/2)\cos(\theta/2)}{\cos\theta_0+\cos\theta}\right.$$

$$\times[1-\exp(2ikb\sin\theta_0)C(\theta_0<\Omega)]C(\Omega<\theta)$$

$$+\frac{2i\sin(\theta_0/2)\sin(\theta/2)}{\cos\theta_0+\cos\theta}$$

$$\times\exp\{-ika[\cos(\theta_0+\Omega)+\cos(\theta-\Omega)]\}$$

$$\left.\times C(\theta_0<\pi-\Omega)[1-\exp(i2kb\sin\theta_0)C(\pi-\Omega<\theta)]\right). \qquad (30)$$

The first term in the large-parenthesis in the above expression corresponds to the ray diffracted from the upper edge whereas the second term gives the ray diffracted from the lower edge. The second term in each square bracket denotes the contribution from rays which are reflected from the lower or upper half plane as well diffracted by upper or lower edges, respectively. The expression differs from Keller's[2] for a slit in a screen in that it has extra terms corresponding to the reflection from the two half planes. Also it has extra shadow boundaries at $\theta_0=\pi-\Omega$ and $\theta=\Omega$ which may be easily removed by using Fresnel integral expressions as indicated in (20)—(25). However, the solution of (27) and (28) for $T_+(\alpha)$ and $V_-(\alpha)$ is much more complicated in this case. For $\Omega=0$, the asymptotic expression is identical to that of Keller[2].

Similarly, the expression for ϕ_{int}, which corresponds to interaction between the two half planes, reduces to

$$\phi_{\text{int}}=\frac{i\exp[i(kr+ka)]}{(2\pi kr)^{1/2}(2\pi ka)^{1/2}}\left(1+\frac{\exp(i2ka)\cos^4(\Omega/2)}{2i\pi ka\cos^2\Omega}\right)^{-1}$$

$$\times\left[\exp[-ika\cos(\theta-\Omega)]\frac{\cos(\theta_0/2)\cos^2(\Omega/2)\sin(\theta/2)}{(\cos\Omega+\cos\theta_0)(\cos\Omega-\cos\theta)}\right.$$

$$\times(1-\exp(i2kb\sin\theta_0)C(\theta_0<\Omega))$$

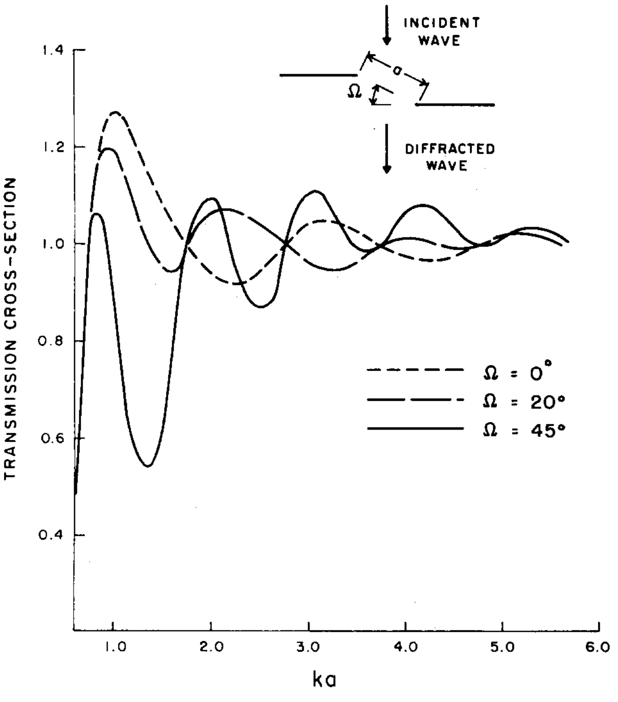


FIG. 3. Transmission cross section vs slit width (ka) for various stagger angles (Ω) of the planes.

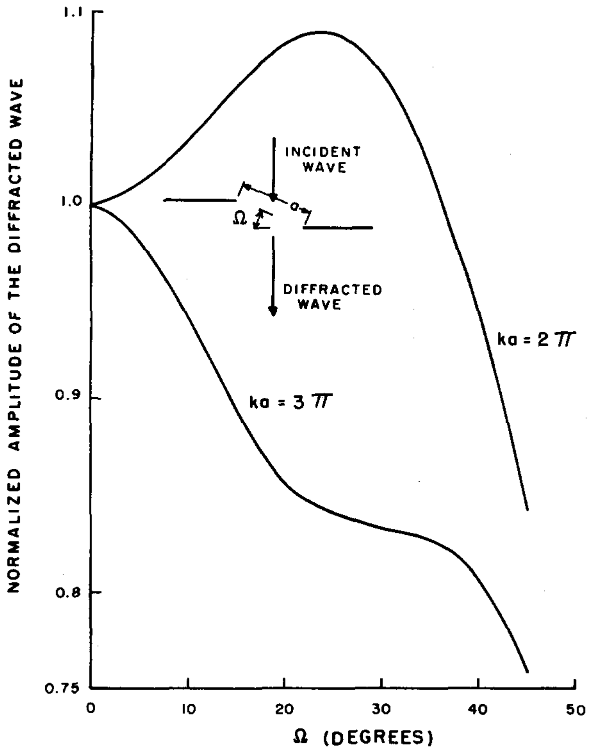


FIG. 4. Transmitted amplitude (normalized to $\Omega = 0$) vs stagger angle (Ω) for two slit widths.

$$\times(1-\exp(i2kb\sin\theta_0)\,C(\pi-\Omega<\theta))$$

$$+\exp[-ika\cos(\theta_0+\Omega)]\,\frac{\sin(\theta_0/2)\cos^2(\Omega/2)\cos(\theta/2)}{(\cos\Omega-\cos\theta_0)(\cos\Omega+\cos\theta)}$$

$$\times\, C(\theta_0<\pi-\Omega)\,C(\Omega<\theta)$$

$$+\frac{\exp[i(ka-3\pi/4)]}{(2\pi ka)^{1/2}}\,\frac{\cos^2(\Omega/2)}{\cos\Omega}\left(\frac{\cos(\theta_0/2)\cos^2(\Omega/2)\cos(\theta/2)}{(\cos\Omega+\cos\theta_0)(\cos\Omega+\cos\theta)}\right.$$

$$\times(1-\exp(i2kb\sin\theta_0)\,C(\theta_0<\Omega)\,C(\Omega<\theta))$$

$$+\exp\{-ika[\cos(\theta_0+\Omega)+\cos(\theta-\Omega)]\}$$

$$\times\frac{\sin(\theta_0/2)\cos^2(\Omega/2)\sin(\theta/2)}{(\cos\Omega-\cos\theta_0)(\cos\Omega-\cos\theta)}\,C(\theta_0<\pi-\Omega)$$

$$\left.\times(1-\exp(i2kb\sin\theta_0)\,C(\pi-\Omega<\theta))\right)\Bigg]. \tag{31}$$

This interaction field represents the field of four types of higher-order rays employed by Keller[2] in the slit solution except for some modifications. The first modification is that the factor $\exp[i(ka-\pi/4)/(2\pi ka)^{1/2}]$ related to the edge—edge ray field for an ordinary slit is multiplied in the present case by $\cos^2(\Omega/2)/\cos\Omega$, which reduces to unity for $\Omega=0$. The second modification is related to the factor $\cos(\theta_0/2)\cos(\theta/2)/(1+\cos\theta_0)(1+\cos\theta)$ which appears in the diffraction coefficient part of the edge—edge interaction term for an ordinary slit. It is replaced by $\cos(\theta_0/2)\cos(\theta/2)\cos^2(\Omega/2)/(\cos\Omega+\cos\theta_0)(\cos\Omega+\cos\theta)$. The last modification relates to the appearance of additional terms due to reflection from the lower and upper half planes. For the special case $\Omega=0$, ϕ_{int} reduces to the expression found by Karp and Russek[3] and Keller[2] for the slit in a conducting screen.

4. NUMERICAL RESULTS AND DISCUSSION

Computations of transmission cross section were performed for various slit widths and stagger angles. The transmission cross section was obtained from the imaginary part of $|F|/ka\cos\Omega$, where the far field scattered by the slit is given by $F(2\pi kr)^{-1/2}\exp[i(kr+\pi/4)]$. Figure 3 shows the results of these computations. Examination shows that the magnitude of the first maximum in the transmission cross section vs ka plot

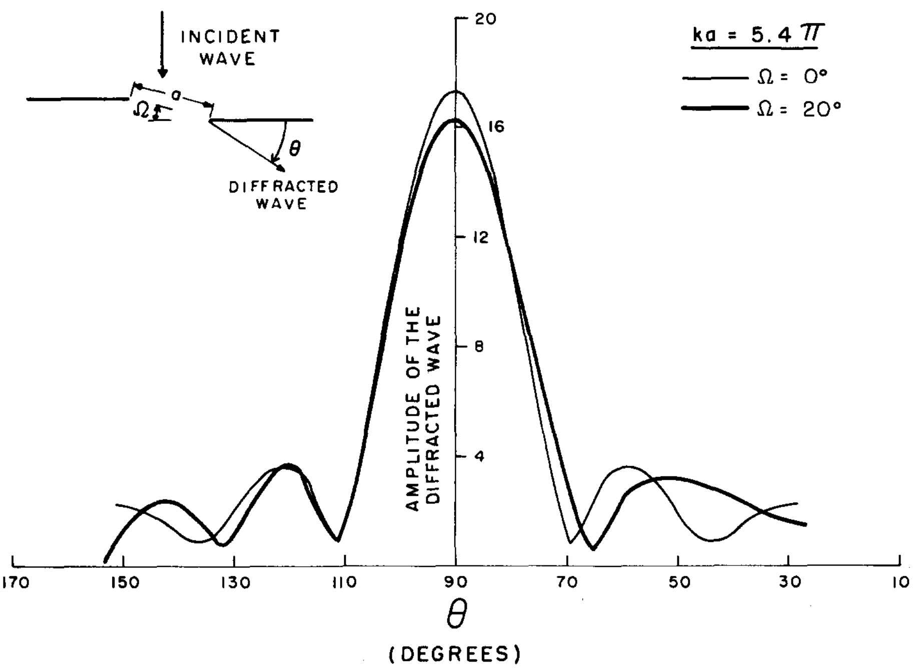


FIG. 5. Diffraction patterns of a slit formed by two staggered planes.

decreases with stagger angle. Also the separation between two maxima (or two minima) decreases with increasing stagger angle. Figure 4 gives similar data except that here we plot the magnitude of the diffracted field, normalized to its magnitude at $\Omega=0$. We note that it is possible to increase or decrease transmission by changing the stagger angle. Figure 5 shows diffraction patterns for a slit with Ω (stagger angle) $=0°$ and $20°$. It is evident that the patterns become asymmetric (for normal incidence) if the planes are staggered and that variations of bandwidth and main lobe amplitude may be obtained by changing the stagger angle.

Our asymptotic results are not valid near the shadow boundaries $\theta_0=\Omega$, $\pi-\Omega$; $\theta=\Omega$, $\pi-\Omega$ and for the case when $\Omega=\pi/2$. The diffracted field expressions for these cases may be obtained by expressing $T_+(\alpha)$ and $V_-(\alpha)$ in terms of Fresnel integrals as in the case of $U_-(\alpha)$ [Eqs. (20)–(25)]. However, the solution of (27) and (28) for $T_+(\alpha)$ and $V_-(\alpha)$ is much more complicated in this case.

5. CONCLUSIONS

The problem of diffraction of a plane wave by a slit formed by two staggered planes has been solved approximately by using the Wiener–Hopf technique and the saddle point method of integration. The results obtained are identified as rays emanating from the two edges and are similar to that of Karp and Russek[3] and Keller[2] for a slit in a plane screen except for a few modifications. The results indicate that various changes in main lobe amplitude, beamwidth and transmission cross section of a slit may be obtained by changing the angle of stagger of the planes.

ACKNOWLEDGMENTS

The author wishes to thank Dr. R. A. Hurd for many helpful discussions and Mrs. M. Steen for performing the computations.

[1] P. M. Morse and P. J. Rubenstein, Phys. Rev. 54, 895 (1938).

[2] J. B. Keller, J. Appl. Phys. 28, 426 (1957).

[3] S. N. Karp and A. Russek, J. Appl. Phys. 27, 886 (1956).

[4] B. Noble, *Methods Based On Wiener—Hopf Technique* (Macmillan, New York, 1958).

Asymptotic behavior of Markoffian kinetics

Koichiro Matsuno

Central Research Laboratories, Nippon Electric Company, Ltd., Kawasaki 211, Japan

(Received 24 April 1974)

The ergodic state with the least mean recurrence time in an irreducible Markoffian process having countable states, between any pair of which transition probability rates are definable, is the state with the least irreversible decay rate if neither the microscopic reversibility nor the doubly stochastic property holds.

1. THEOREM

Suppose a finite classical system whose dynamical evolution obeys an irreducible Markoffian process having countable states $\{i\}$ $(i=1,2,\cdots)$, between any pair of which transition probability rates are definable.[1] Hence the kinetics of the Markoffian process follows the Pauli master equation

$$\frac{\partial}{\partial t} P(i,t)=\sum_{j\neq i} W(i\leftarrow j)P(j,t)-\sum_{j\neq i} W(j\leftarrow i)P(i,t), \qquad (1.1)$$

where $P(i,t)$ is the probability density of the representative point R of the system found on coordinate i at time t and $W(i\leftarrow j)$ is the probability of transition from j to i per unit time. Here, the term *state* is identically referred to as *coordinate*. The theorem we shall prove is as follows:

Theorem A: Suppose the Markoffian kinetics following the Pauli master equation (1.1) with the set of transition probability rates $\{W(i\leftarrow j)\}$ each of which is independent of time. Then, the mean recurrence time $\tau(i\rightarrow i)$ of the coordinate i has its minimum value at coordinate $i=i_m$ giving the least irreversible decay rate

$$i_m\equiv\{i\,;\,\min\{W(i)\}\} \qquad (1.2)$$

with

$$W(i)\equiv\sum_{j\neq i} W(j\leftarrow i) \qquad (1.3)$$

if the auxiliary condition

$$\widetilde{W}(i_m)-W(i_m)>0 \qquad (1.4)$$

with

$$\widetilde{W}(i)\equiv\sum_{j\neq i} W(i\leftarrow j) \qquad (1.5)$$

is satisfied.

The auxiliary condition (1.4) excludes both the microscopic reversibility

$$W(i\leftarrow j)=W(j\leftarrow i) \text{ for any pair of } i \text{ and } j\,(\neq i) \qquad (1.6)$$

and the doubly stochastic property

$$W(i)=\widetilde{W}(i) \text{ for any } i. \qquad (1.7)$$

In order to prove Theorem A, we shall first define a particular measure and next prove a set of lemmas.

2. DEFINITION

As a prerequisite for the proof of Theorem A, we shall define a measure

$$\mu[i\,;T,\tau_0]\equiv[T]/T. \qquad (2.1)$$

The meanings of the symbols are as follows: We call as a unit-continuous event (UCE) of coordinate i the event that R keeps remaining on coordinate i from the moment R transits from any coordinate except i to i till the moment R again transits from i to anywhere else; $[T]$ is the total time duration of UCE's of coordinate i appearing during the time interval T. Here, the time duration of UCE of i is counted as zero if it is less than an appropriate time unit τ_0. $[T]$ is consecutively counted by the unit of τ_0 in a single sequence of the interval $T(\geqslant\tau_0)$ from an arbitrary initial time $t=t_{in}$, while, if two successive UCE's of any j with the intermission of the interval τ' occur with the time durations $\tau_1(>0)$ and $\tau_2(>0)$, respectively, and if $\tau'<\tau_0$, the successive events are regarded as if a single UCE of j with the time duration $(\tau_1+\tau'+\tau_2)$ had occurred. In general, the quantity $\mu[i\,;\,T,\tau_0]$ depends upon an initial time $t=t_{in}$ from which one begins to measure the interval T. If the initial time is specifically referred to, the quantity will be denoted as $\mu[i\,;T\mid t_{in},\tau_0]$.

Next, we shall introduce the probability measure $P_i(t_2,t_1)$ of the event that R found on i at time $t=t_1$ keeps remaining on the same coordinate until $t=t_2(>t_1)$ without suffering any transition. One readily finds

$$P_i(t_2,t_1)\equiv P_i(t_2-t_1)=\{P_i(1)\}^{(t_2-t_1)} \qquad (2.2)$$

with

$$P_i(1)\equiv\exp[-W(i)], \qquad (2.3)$$

since the event is Markoffian as seen from the master equation (1.1) and since each member of the set $\{W(i\leftarrow j)\}$ has been supposed to be independent of time.

Furthermore, we shall introduce another quantity $P_c(j\mid i\,;T)$ which is characteristic of the probability of a particular event continuing over the interval T as follows:

$$P_c(j\mid i\,;T)\equiv P_j(T)/P_i(T) \qquad (2.4)$$

for the event that R is found on j at an initial time $t=t_{in}$ followed by another one in which it keeps staying on j till a later time $t=t_{in}+T$, and

$$P_c(j\mid i\,;T)\equiv 0 \qquad (2.5)$$

for the event that R is found on any coordinate except j even a moment during the interval T between an initial time $t=t_{in}$ and a later time $t=t_{in}+T$. If the initial time t_{in} for fixing the quantity $P_c(j\mid i\,;T)$ is specifically referred to, the quantity will be denoted as $P_c(j\mid i\,;T\mid t_{in})$. Since the quantity $P_c(j\mid i\,;\,T\mid t_{in})$ is related to the probability of a particular event, it follows that

$$P_c(j\mid i\,;T\mid t_{in})=0,\ \text{ if } P_c(j\mid i\,;T\mid t_{in})=\epsilon \ \text{ with } \epsilon\rightarrow 0. \qquad (2.6)$$

One notes that $P_c(j\mid i\,;T\mid t_{in})=0$ if $\mu[j\,;T\mid t_{in},\tau_0]=0$, and that $\mu[j;T\,|\,t_{\text{in}},\tau_0]=1$ if $P_c(j\,|\,i;T\,|\,t_{\text{in}})\neq 0$ in the sense of (2.6).

3. LEMMAS

Lemma A:

$$0\leq \mu[\alpha\;;T\,|\,t_{\text{in}},\tau_0]\leq 1 \tag{3.1}$$

with

$$\alpha\subset\{i\},$$

where α means an arbitrary local subset of coordinates, i.e., states belonging to the total countable states $\{i\}$ and t_{in} is an arbitrary initial time.

Proof: If one replaces a single coordinate by a local subset of coordinates in the definition of (2.1), the present *lemma* immediately follows. QED

Lemma B:

$$\mu[\alpha\;;T\,|\,t_{\text{in}},\tau_0]+\mu[\beta\,;T\,|\,t_{\text{in}},\tau_0]\leq\mu[\alpha\cup\beta,\,T\,|\,t_{\text{in}},\tau_0] \tag{3.2}$$

with

$$\alpha,\beta\subset\{i\}\text{ satisfying }\alpha\cap\beta=\phi,$$

where t_{in} is arbitrary.

Proof: Since the representative point R cannot be found simultaneously in both the subsets α and β without any interception between the two, the present *lemma* results. QED

Lemma C:

$$\lim_{T\to\infty}\mu[j\,;T\,|\,t_{\text{in}},\tau_0]<1\quad\text{for } j\neq k_m \tag{3.3}$$

with

$$k_m\equiv\{i\,;\min\{\tau(i\to i)\}\}. \tag{3.4}$$

Proof: If it occurred that

$$\lim_{T\to\infty}\mu[j\,;T\,|\,t_{\text{in}},\tau_0]=1\quad\text{for } j\neq k_m,$$

the definition of (2.1) would give the inequality $\tau(j\to j)<\tau(k_m\to k_m)$. This apparently contradicts condition (3.4). QED

Lemma D:

$$\lim_{T\to\infty}\mu[k_m\,;T\,|\,t_{\text{in}},\tau_0]>0 \tag{3.5}$$

with

$$\tau_0>\tau(k_m\to k_m)$$

if the auxiliary condition

$$\tau(k_m\longleftrightarrow k_m)>0 \tag{3.6}$$

is satisfied, where $\tau(i\longleftrightarrow i)$ is the average holding time of R on coordinate i from the moment R transits to i from anywhere else till the moment it again transits from $\underline{i}$ to anywhere else.

Proof: The inequality (3.5) is straightforward from the definition of (2.1) under the condition (3.6). QED

Lemma E:

$$P_c(j\,|\,i\,;T)=0\quad\forall\{\mu[j\,;T\,|\,t_{\text{in}},\tau_0]=0\}, \tag{3.7}$$

where $\forall\{\mu=0\}$ stands for any event satisfying $\mu=0$ and t_{in} is arbitrary.

Proof: For any event satisfying the condition $\mu[j\,;T\,|\,t_{\text{in}},\tau_0]=0$, it never occurs that R keeps remaining on coordinate j any longer than τ_0 without suffering transitions. Hence the definition of (2.5) leads to *Lemma E*. QED

Lemma F:

$$P_c(j\,|\,i\,;T)\neq 0\quad\text{where }\exists\{\mu[j\,;T\,|\,t_{\text{in}},\tau_0]=1\}, \tag{3.8}$$

where $\exists\{\mu=1\}$ stands for a certain event satisfying $\mu=1$ and t_{in} is arbitrary.

Proof: For a certain event satisfying $\mu[j\,;T\,|\,t_{\text{in}},\tau_0]=1$, the possibility is not completely excluded that R may keep staying on j during the interval T if an initial time t_{in} is appropriately chosen. Such an event yields $P_c(j\,|\,i\,;T)\neq 0$ because of the definition of (2.4). QED

4. PROOF

Before entering the proof of Theorem A, we shall prove another auxiliary theorem.

Theorem B: Suppose $\mu[x\,;T\,|\,t_{\text{in}},\tau_0]=1$, where x is an arbitrary coordinate among $\{i\}$ and t_{in} is an arbitrary initial time. Then,

$$\mu[i\,;T\,|\,t_{\text{in}},\tau_0]=1\rightleftarrows P_c(j(\neq i)\,|\,i\,;T\,|\,t_{\text{in}})=0 \tag{4.1}$$

for any coordinate j except i.

Proof: If $\mu[i\,;T\,|\,t_{\text{in}},\tau_0]=1$, then $\mu[j(\neq i);\,T\,|\,t_{\text{in}},\tau_0]=0$ because of Lemmas A and B. Hence one obtains $P_c(j(\neq i)\,|\,i\,;T\,|\,t_{\text{in}})=0$ from Lemma E. Next, suppose $P_c(j(\neq i)\,|\,i\,;T\,|\,t_{\text{in}})=0$ under the constraint $\mu[x\,;T\,|\,t_{\text{in}},\tau_0]=1$ for an arbitrary t_{in}. On the other hand, $P_c(x\,|\,i\,;T\,|\,t'_{\text{in}})\neq 0$ follows for a certain event satisfying $\mu[x\,;T\,|\,t'_{\text{in}},\tau_0]=1$ because of Lemma F. If $x\neq i$, a contradiction would occur. Consequently, $\mu[i\,;T\,|\,t_{\text{in}},\tau_0]=1$ follows. QED

Proof of Theorem A: Since $P_c(j\,|\,i\,;T)$ is either $P_j(T)/P_i(T)$ or identically zero as shown in (2.4) and (2.5), Theorem B in the limit $T\to\infty$ will yield

$$\lim_{T\to\infty}\mu[i\,;T\,|\,t_{\text{in}},\tau_0]=\begin{cases}1 & \text{for } i=i_m\\ 0 & \text{for } i\neq i_m\end{cases} \tag{4.2}$$

if and only if the ansatz

$$\lim_{T\to\infty}\mu[x\,;T\,|\,t_{\text{in}},\tau_0]=1\quad\text{for an arbitrary } x \tag{4.3}$$

holds, where the limiting procedure presented in (2.6) is employed. On the other hand, Lemma C says that the ansatz

$$\lim_{T\to\infty}\mu[x\,;T\,|\,t_{\text{in}},\tau_0]=1$$

never holds unless $x=k_m$. Hence, the necessary condition, by which Lemma C does not contradict the statement presented in (4.2) and (4.3), is the equality

$$i_m=k_m. \tag{4.4}$$

A sufficient condition allowing for the identity of state i_m to state k_m is the inequality $\tau(i_m\longleftrightarrow i_m)>0$, which we shall prove, because of both Lemma D and the ansatz (4.3). In fact, the definition of (2.1) would not exclude the event $\mu[k_m,\,T\,|\,t'_{\text{in}},\,\tau_0]=1$ as a particular case of

Lemma D since $\tau(k_m \to k_m) < \tau(j \to j)$ for any $j(\neq k_m)$, where $T(\geqslant \tau_0)$ is arbitrary and t'_{in} is a particular time.

The average holding time $\tau(i \longleftarrow i)$ can be evaluated as follows: In the asymptotic limit, any coordinate must be balanced in the sense that the transition of R from coordinate i to anywhere else with the rate $W(i)$ should be compensated by its just reversed transition with the rate $\widetilde{W}(i)$. The condition of the asymptotic balance is expressed as

$$W(i) = \widetilde{W}(i) \exp[-W(i)\tau(i \longleftarrow i)] \tag{4.5}$$

or, equivalently, as

$$\tau(i \longleftarrow i) = \frac{1}{W(i)} \ln \frac{\widetilde{W}(i)}{W(i)} \tag{4.6}$$

with the aid of the probability function given in (2.2). Hence the inequality of (1.4) yields

$$\tau(i_m \longleftarrow i_m) > 0. \tag{4.7}$$

Since the positiveness of the average holding time $\tau(i_m \longleftarrow i_m)$ has been proved, it turns out that the state k_m with the least mean recurrence time is identical to the state i_m with the least irreversible decay rate under the auxiliary condition (1.4). This completes the proof of Theorem A. QED

In order to make sure that the state i_m is really ergodic, we shall estimate the magnitude of the mean recurrence time $\tau(i_m \to i_m)$. Let us consider the probability

$$P^{(\nu)}_{\text{cont}}(j \neq i;\, t_2, t_1)$$

of an event denoted as $\nu(=1,2,\cdots)$ that R is never found on i from the moment $t=t_1$ till $t=t_2(>t_1)$. Hence one obtains

$$\frac{P^{(\nu)}_{\text{cont}}(j \neq i;\, t_2, t_1)}{P_i(t_2 - t_1)} \leqslant \frac{\exp[-\{\min_{j\neq 1}\{W(j)\}\}(t_2 - t_1)]}{\exp[-W(i)(t_2 - t_1)]}. \tag{4.8}$$

As a particular case of this inequality, it follows that

$$\frac{P^{(\nu)}_{\text{cont}}(j \neq i_m;\, t_2, t_1)}{P_{i_m}(t_2 - t_1)} \leqslant \exp\left(-\frac{(t_2 - t_1)}{\tau_\gamma}\right) \tag{4.9}$$

with

$$\tau_\gamma \equiv \left[\min_{j \neq i_m} \{W(j) - W(i_m)\}\right]^{-1}. \tag{4.10}$$

If the time interval $(t_2 - t_1)$ is much greater than τ_γ, the right-hand side of (4.9) would almost vanish. The probability of an event that R could never be found on i_m all through the time interval, which is much greater than τ_γ, would thus become vanishingly small compared with the probability that R would keep staying on i_m during the same interval.

Suppose one can classify each event, which continues over the interval τ_0 along the time axis, into the following three categories. The first group is for the events in which R keeps staying on i_m all through the time interval τ_0. The second one is for the events in which R is never found on i_m even a moment during the interval τ_0. And the third one is for all the other events, i.e., the events in which R is found on i_m even a moment during the interval τ_0. If the inequality

$$\tau_0 \gg \tau_\gamma \tag{4.11}$$

is satisfied, only events of the first and third groups will appear with the probability of almost unity since the probability of an event belonging to the second group is found to be vanishingly small compared with the one belonging to the first group. Hence one finds that the mean recurrence time $\tau(i_m \to i_m)$ has the magnitude of

$$\tau(i_m \to i_m) \lesssim \tau_0,$$

where τ_0 is given in (4.11).

5. DISCUSSIONS AND CONCLUDING REMARKS

We shall present the physical significances and implications of the theorems proven in the followings.

Although Theorem A has been proved in the case that the transition probability rates $\{W(i \leftarrow j)\}$ are independent of time, one can readily prove a similar theorem even if $\{W(i \leftarrow j)\}$ are periodic functions of time. All one has to do is to choose a fundamental period as a time unit of the kinetics and to follow the similar arguments presented in Sec. 2—4.

Markoffian kinetics sometimes could result from a certain projection eliminating irrelevant variables from a dynamical system,[2,3] which has a fixed Poincaré cycle. The recurrence times of the reduced Markoffian process do not agree with the Poincaré cycle of the dynamical system which admits a pointwise detection, that is, a fine-grained observation. The present discrepancy is due to the fact that the projection always causes a contraction of information on the side of observer and that the Markoffian kinetics necessarily accompanies a set of coarse-grained states. Because of this projection, the dynamical system, which is not yet subject to the thermodynamic limit, would reduce to a Markoffian process with much smaller recurrence times than the original Poincaré cycle. Hence, the size of the least mean recurrence time depends entirely upon the coarse-graining which the observer employs.

If either the microscopic reversibility or the doubly stochastic property holds, the average holding time of any state would vanish following the expression given in (4.6). This in turn yields the observation that one cannot identify the state with the least mean recurrence time among those each of which has the vanishing holding time on average. In fact, it is known that a steady distribution with equal weight follows if either the microscopic reversibility or the doubly stochastic property is satisfied.[1]

There is an argument that a steady distribution could be established in a Markoffian process even if the microscopic reversibility does not hold.[4,5] The principle of detailed balance is an example giving such a steady distribution which yields the probability distribution being proportional to the inverse of the mean recurrence time of each Markoffian state.[1] The steady distribution, however, excludes fluctuations of each recurrence time.[6] Hence, one observes that the probabilistic kinetics based upon the steady distribution cannot deal with the macrokinetics associated to macroscopic fluctuations[6,7] of each recurrence time around its mean value.

The state with the least mean recurrence time appears most frequently among the countable states of an irreducible Markoffian process. An observer, who is interested only in a *once-and-for-all* event which occurs along the time axis and not in a sort of ensemble average of events, would regard the state with the least mean recurrence time just like an everlasting state if he cannot follow the fast transition dynamics with the characteristic time less than the memory holding time of his own and if the holding time is of the order of the least mean recurrence time.

A certain class of nonlinear statistical mechanics off equilibrium belongs to Markoffian kinetics.[8] Each Markoffian state represents a structure which could be realized in nonlinear statistical mechanics. Although it has been pointed out that if the structure is extensive[9,10] or, more generally, is invariant under a scale transformation,[11] the probabilistic kinetics could reduce to a set of a few typical kinetics, specifically, in the thermodynamic limit, the structures which one meets in nonlinear statistical mechanics off equilibrium are by no means restricted to those scale-invariant ones. If a structure with an intensive characteristic[8] appears, the one most important subject with which the observer is concerned will be to identify the information associated with the structure as a whole. We have shown that the state with the least mean recurrence time is the one with the least irreversible decay rate under a reasonable condition regardless of whether the corresponding structure is extensive or intensive.

In conclusion, a principal significance of the theorems proven is seen in the fact that the structure which the observer employing an appropriate coarse-grained time unit for time measurement may regard as being in an asymptotically stable state is the one with the least irreversible decay rate. The same statement has been argued to apply also to a nonlinear system even if it does not obey Markoffian kinetics.[6] Henceforth, given a nonlinear system arbitrarily off equilibrium, one realizes that the more stable structure which the observer employing the coarse-graining may identify in the course of time evolution is the structure with the less irreversible decay rate. Such an observer regards the least irreversible decay rate as a selection rule available to a nonlinear system off equilibrium not yet subjected to the thermodynamic limit. This exhibits a distinct contrast to the thermodynamic second law as the selection rule in the thermodynamic limit.

[1] For example, J. L. Doob, *Stochastic Processes* (Wiley, New York, 1953).

[2] L. van Hove, Physica **21**, 517 (1955).

[3] R.W. Zwanzig, in *Lectures in Theoretical Physics*, edited by W.E. Brittin and ohters (Interscience, New York, 1961).

[4] R. Graham, Springer Tracts in Mod. Phys. **66**, 1 (1973).

[5] K. Tomita and H. Tomita, Phys. Lett. **46** A, 265 (1973) and Prog. Theor. Phys. **51**, 1731 (1974).

[6] K. Matsuno, J. Stat. Phys. **11**, 87 (1974).

[7] K. Matsuno, Phys. Lett. **47** A, 99 (1974).

[8] M. Eigen, Naturwiss. **58**, 465 (1971).

[9] N.G. van Kampen, Can. J. Phys. **39**, 551 (1961).

[10] R. Kubo, K. Matsuo, and K. Kitahara, J. Stat. Phys. **9**, 51 (1973).

[11] H. Mori, Prog. Theor. Phys. (to be published).

Quantumlike formulation of stochastic problems

Emilio Santos

Departamento de Optica, Universidad de Valladolid, and GIFT, Valladolid, Spain

(Received 6 March 1974)

It is shown that a stochastic process can be viewed as a set of states (normalized positive linear functionals) over an Abelian C^*-algebra. Alternatively, the stochastic process can be associated with a set of representations of the algebra as a subalgebra of the (noncommutative) C^*-algebra of bounded operators in a Hilbert space. Then, an operator equation can be associated with every stochastic equation in some general conditions. The formalism is applied to Brownian motion. Then, we study the nonrelativistic motion of a single particle in stochastic electrodynamics, a theory which has been proposed as a possible alternative to quantum mechanics. The equations of motion, which are derived, coincide with the basic ones of quantum mechanics. The differences between this theory and quantum mechanics are summarized.

INTRODUCTION

The motive for the present paper is the attempt to give an answer to the following question: Is a classical theory of the microworld possible? For many years, a negative answer was given to this question on the basis of von Neumann theorem against hidden variables in quantum mechanics. It is now clear that the theorem does not answer the question, which is therefore open.[1] What seems almost sure is that if a classical theory of the microworld is possible, it must be a stochastic one. Here, classical theory means that every material system can be described by a set of variables, every one having a well-defined value at any time. Stochastic theory means that the laws of motion cannot be exactly stated, so that stochastic hypotheses must be made about it, with the result that the equations of motion become stochastic.

Many attempts have been made to derive quantum mechanics from classical stochastic hypotheses, but without real success until now[2]. The difficulty might be that the mathematical techniques developed to deal with stochastic systems are not suitable for the specific stochastic system which is —maybe—behind quantum mechanics. The problem can be approached in just the opposite direction, i.e., developing a quantum-like formalism to deal with general stochastic systems. This idea has been previously considered by Collins and Hall,[3] but the formalism has not been developed to the point of allowing the solution of actual stochastic problems.

In the present paper, a general quantum-like formalism is developed making use of the theory of Banach algebras. Then, it is applied to the classical stochastic problem best known in physics: Brownian motion. After that, the formalism is used to study stochastic electrodynamics, a theory which seems the best candidate to be a classical alternative to quantum mechanics.[4] Aside from the interest that the formalism has by way of answering the fundamental question posed above, it may be useful in other branches of probability theory.

I. ABELIAN C*-ALGEBRA OF A RANDOM VARIABLE

The usual way to deal with a random variable is through its probability distribution. An alternative approach is the use of the moments of the distribution. In very general cases both approaches are equivalent, though the second one is the most interesting in physics because it is closer to the experiments. In fact, most times the measurement process does not give the full probability distribution of a variable, but only one or a few of its moments (for instance, the mean and the standard deviation).

Let X be a real random variable (i.e., one whose range is a subset of the set of real numbers). The moments of the probability distribution are the expectation values of the integer powers of the variable, and they will be written

$$\langle X^n \rangle \equiv E[X^n]. \tag{1.1}$$

In stochastic problems it may be necessary to use functions of X more general than the integer powers, so that we will consider the set, $\mathcal{A}$, of all polynomials of X with complex coefficients. It is straightforward to endow this set with the structure of an involutive algebra (or * -algebra), which is Abelian (commutative). The expectation value defines a linear functional over the algebra, which is fully determined by the probability distribution. In very general cases, the functional also determines the probability distribution. Then, it follows that giving the probability distribution of a random variable is equivalent to giving a linear functional over its associated *-algebra.

As a simple example, let us consider a discrete random variable, X, whose range is the set of values x_k ($k=1,2,\ldots,n$). It is rather trivial to verify that a representation of the *-algebra of (complex) polynomials of X can be obtained by means of diagonal matrices in such a way that the matr'x associated with the polynomial $f(X)\in\mathcal{A}$ is

$$F \equiv \begin{pmatrix} f(x_1) & 0 & \cdots & 0 \\ 0 & f(x_2) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & f(x_n) \end{pmatrix} \tag{1.2}$$

Linear operations involving polynomials in X correspond to similar operations involving the associated matrices and the same is true for the product of two polynomials. Now, for a random variable, X, whose range is a finite set of n real numbers, every function of X can be identified with a polynomial of degree less than n. Therefore, the following result is obtained: The set of (complex) functions of a discrete random variable whose range is a set of n real numbers, is isomorphic to some subset of the set of linear operators of a (complex) vector space of n dimensions.

It is plausible to attempt an extension of this result to more general random variables. To do this, we make use of the following well-known theorem: A probability measure on a compact configuration space $\mathcal{D}$ is a state (a positive normalized linear functional) over the Abelian C^*-algebra, $\mathcal{A}$, of complex continuous functions on $\mathcal{D}$.[5] This generalizes the statement previously made about the algebra of functions of a discrete random variable. We do not consider here the possible generalizations of this result but, in the following, we will assume without proof that a suitable Abelian C^*-algebra of functions can be associated with any random variable of physical interest, in such a way that the probability measure is a state over the algebra.

Once the C^*-algebra, $\mathcal{A}$, of a random variable, X, is defined, and a state over $\mathcal{A}$ is determined by the probability distribution, the GNS (Gelfand—Naimark—Segal) construction allows one to find a representation of $\mathcal{A}$ in a Hilbert space $\mathcal{H}$ in such a way that

$$\langle f\rangle = \langle\psi|\pi(f)|\psi\rangle, \tag{1.3}$$

where $|\psi\rangle\in\mathcal{H}$ is a unit cyclic vector (we will use Dirac notation throughout) and $\pi(f)$ is the operator on $\mathcal{H}$ associated with $f\in\mathcal{A}$. Furthermore, the representation is unique up to unitary mappings. [The vector $|\psi\rangle$ is said cyclic if the set $\{\pi(f)|\psi\rangle, f\in\mathcal{A}\}$ is dense in $\mathcal{H}$.]

As a simple example, let us consider again the case of a discrete random variable. The probability distribution associates a nonnegative number p_k with every number x_k of the range of the random variable. The expectation value of a function, $f(X)\in\mathcal{A}$, of the random variable is given by

$$\langle f\rangle = \sum_{k=1}^{n} p_k f(x_k). \tag{1.4}$$

It is obvious that this defines a linear normalized positive functional over $\mathcal{A}$. The normalization corresponds to the condition that the sum of all probabilities, p_k, is unity. Equation (1.4) can be written in matrix form if we define the row matrix of the probabilities

$$\langle P| \equiv p_1 p_2 \cdots p_n, \tag{1.5}$$

and a column matrix whose elements are all unity:

$$|1\rangle \equiv \begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}. \tag{1.6}$$

These matrices are written in Dirac notation in order to emphasize that they can be assumed to represent vectors in the (complex) Hilbert space of n dimensions. Then, Eq. (1.4) is written, taking Eq. (1.2) into account,

$$\langle f\rangle = \langle P|F|1\rangle. \tag{1.7}$$

As F is diagonal, it commutes with any diagonal matrix S, so that we can also write

$$\langle f\rangle = \langle P|S^{-1}FS|1\rangle \equiv \langle\psi|F|\psi\rangle. \tag{1.8}$$

The matrix S has been so chosen that $\langle P|S^{-1}$ is the transposed conjugate (Hermitian conjugate) of $S|1\rangle$. To do this, it is enough to choose S as the matrix whose elements are

$$S_{jk} = \delta_{jk} P_j^{1/2} \exp(i\varphi_k), \tag{1.9}$$

where φ_k are arbitrary real numbers. (The inverse of the matrix S so defined exists whenever all the probabilities p_j are different from zero. If some of them were zero, we put it equal to ϵ and take the limit $\epsilon\to 0$ at the end of the calculation. It is not difficult to see that the matrices $\langle\psi|$ and $|\psi\rangle$ are well defined in this limit). Now, $|\psi\rangle = S|1\rangle$ can be considered a unit vector in the n-dimensional complex Hilbert space and Eq. (1.8) is a particular case of Eq. (1.3). It can be seen that this construction of the Hilbert space for a discrete random variable agrees with the one of GNS.

In this example it is easy to see that:

(1) Every real random variable $f\in\mathcal{A}$ can be associated with a Hermitian operator $\pi(f)$ on a Hilbert space $\mathcal{H}$.
(2) The set of probability distributions of the variables is determined by a unit vector $|\psi\rangle\in\mathcal{H}$.
(3) The range of the variable is the spectrum of its associated operator [remember that $\pi(f)$ can be represented by a diagonal matrix F whose eigenvalues are the numbers $f(x_k)$.]
(4) The probability that $f(X)$ takes on the value $f(x_k)$ is equal to the square modulus of the projection of $|\psi\rangle$ into the subspace associated with the eigenvalue.

The analogy of these results with the postulates of quantum mechanics is striking. It is enough to replace "real random variable" by "(Hermitian) observable," "set of probability distributions" by "state of the system," and "range of the variable" by "possible outcomes as a result of the measurement of the observable." It is to be noted that, here, each "set of probability distributions" corresponds to a "pure state" of quantum mechanics (i.e., it is associated with a vector—rather than a density matrix—in the Hilbert space).

It is straightforward to extend these results to systems of random variables. In fact, any system of random variables can be reduced to a single one. If X_j are random variables whose ranges are the spaces $\mathcal{D}_j$, the set $\{X_j\}$ is equivalent to a single random variable whose range is the Cartesian product of the spaces $\mathcal{D}_j$. All the previous formalism is valid with small changes. In particular, the Hilbert space of the GNS construction is the tensor product of the Hilbert spaces of the individual random variables. As an example, let us consider two random variables: X with range $\{x_1, x_2\}$ and Y with range $\{y_1, y_2\}$. Any function $f(X, Y)$ can be represented by the matrix

$$F \equiv \begin{pmatrix} f(x_1,y_1) & 0 & 0 & 0 \\ 0 & f(x_1,y_2) & 0 & 0 \\ 0 & 0 & f(x_2,y_1) & 0 \\ 0 & 0 & 0 & f(x_2,y_2) \end{pmatrix}. \tag{1.10}$$

It is obvious that the matrices associated with X and Y commute and also that every joint probability distribution can be represented by a (cyclic) vector in the

Hilbert space of four dimensions. This is a particular case of a general property: The set of continuous functions of a system of random variables defined in compact spaces is an Abelian C^*-algebra, and every joint probability distribution corresponds to a state over the algebra. The Abelian property of the algebra is related to the possibility of joint probability distributions, a relation similar to the one which exists in quantum mechanics between the commutativity of observables and the possibility of joint probability distributions of the corresponding variables (which is usually called compatibility of the measures). Actually, not all observables commute in quantum mechanics, while only Abelian C^*-algebras have been considered here till now. In the next section, it will be shown that noncommutative C^*-algebras appear in a natural fashion in the study of stochastic processes.

II. STOCHASTIC PROCESSES

In this section a Hilbert space approach to stochastic processes is presented, which provides a tool for the solution of a wide class of stochastic equations. The relation of this formalism with the standard representation of stochastic processes as curves in Hilbert space[6] will be considered elsewhere.

A stochastic process is a time-dependent family of random variables with the same range. (Only continuous processes will be considered here.) Then, according to the results of the previous section, the process can be viewed as a family of states over an Abelian C^*-algebra $\mathcal{A}$. It was shown that the states can be associated with vectors in a Hilbert space $\mathcal{H}$, so that the stochastic process is associated with a time-dependent "state" vector $|\psi(t)\rangle \in \mathcal{H}$. This is an active (or "Schrödinger") picture of the stochastic process. A more interesting, passive picture is defined in the following.

The expectation value of any function $f(X) \in \mathcal{A}$ of the stochastic process $X(t)$ at time t is

$$\langle f(X)\rangle_t = \langle \psi(t) | \hat{f} | \psi(t)\rangle, \tag{2.1}$$

where $\hat{f} \in \mathcal{M}$ is the representative of $f(X) \in \mathcal{A}$, and $\mathcal{M}$ is the set of bounded operators on $\mathcal{H}$. Note that a sufficient condition for the process $X(t)$ being stationary is that the (unit) vectors $|\psi(0)\rangle$ and $|\psi(t)\rangle$ coincide up to a phase factor. As long as the state vector is assumed normalized at any time, it is possible to find some unitary operator $U(t)$ acting on $\mathcal{H}$ such that

$$|\psi(t)\rangle = U(t)|\psi(0)\rangle, \quad \langle\psi(t)| = \langle\psi(0)|U(t)^\dagger. \tag{2.2}$$

Indeed, $U(t)$ may be chosen in many different forms. It must be pointed out that the unitary operator may depend on the initial state. [In sharp contrast, in quantum mechanics it is asusmed that the evolution operator $U(t)$ is the same for all inital states, which implies that linear relations are conserved by the evolution. It is *not* claimed that this is true for stochastic processes in general.] Now, from Eqs. (2.1) and (2.2) it follows that

$$\langle f(X)\rangle_t = \langle\psi(0)|\hat{f}(t)|\psi(0)\rangle, \quad \hat{f}(t) \equiv U(t)^\dagger \hat{f} U(t). \tag{2.3}$$

It is not difficult to show that the set

$$\mathcal{A}_t \equiv \{\hat{f}(t),\ f \in \mathcal{A}\} \subset \mathcal{M} \tag{2.4}$$

provides a representation of the C^*-algebra $\mathcal{A}$ in the Hilbert space $\mathcal{H}$. Therefore, the stochastic process can be viewed as a family of representations of its associated Abelian C^*-algebra in a Hilbert space. Furthermore, the vector $|\psi(0)\rangle \in \mathcal{H}$ defines a state over the (noncommutative) C^*-algebra $\mathcal{M}$. The probability distribution of the stochastic process at any time t is determined by the restriction of this state to the subalgebra $\mathcal{A}_t \subset \mathcal{M}$. This represents the passive (or "Heisenberg") picture of the process. It must be pointed out that, although each algebra $\mathcal{A}_t$ is Abelian, the elements of different algebras do not commute in general.

At this moment it is possible to define the derivative of the operator $x(t)$ as the limit

$$\dot{\hat{x}}(t) \equiv \lim_{t \to 0} \frac{\hat{x}(t+\Delta t) - \hat{x}(t)}{\Delta t} \in \mathcal{M}, \tag{2.5}$$

provided that this limit exists. The existence of the limit implies that a suitable choice has been made of the operators $U(t)$, which were not uniquely determined by Eq. (2.2). The important property of the operator $\dot{\hat{x}}(t)$ is that it can be associated with the derivative of the stochastic process $\dot{X}(t)$ in such a way that the following relation is fulfilled:

$$\frac{d}{dt}\langle f(X)\rangle_t = \langle\psi(0)|\frac{d}{dt} f(\hat{x}(t))|\psi(0)\rangle. \tag{2.6}$$

Here, the derivative of any power of $\hat{x}$ is given by

$$\frac{d}{dt}(\hat{x}(t)^n) = \sum_{j=0}^{n-1} \hat{x}(t)^j \dot{\hat{x}}(t) \hat{x}(t)^{n-j-1} \equiv nS(\dot{\hat{x}}\hat{x}^{n-1}), \tag{2.7}$$

where S (symmetrizer) means the average of all monomials obtained by rearranging the operators in all possible orderings.

After that, we have a general method for dealing with differential stochastic equations. A stochastic equation is some functional relation between several stochastic processes and its derivatives. By introducing auxiliary processes every equation can be written as a set of first order equations. We will assume that all these can be put in the following form:

$$\mathring{X}_j(t) = g_j(X_1(t),\ X_2(t),\ \ldots,\ X_n(t)), \tag{2.8}$$

where $X_1(t), \ldots, X_n(t)$ are stochastic processes, and the g_j $(j = 1, \ldots, n)$ are some (given) functions which map the Cartesian product of the ranges of $X_1 \cdots X_n$ onto the range of $\dot{X}_j$. Now, the set of processes can be considered as a single process, a result similar to the one obtained for random variables at the end of Sec. I. Therefore, the set of Eqs. (2.8) can be written as the single equation

$$\dot{X}(t) = g(X(t)), \tag{2.9}$$

where g is some function of the range of X onto the range of $\dot{X}$. This relation means that the processes $\dot{X}$ and $g(X)$ have the same range and the same probability distribution at any time. This implies that, for any function $f(X) \in \mathcal{A}$, the following relation holds:

$$\langle f(\dot{X})\rangle_t = \langle f(g(X))\rangle_t. \tag{2.10}$$

Taking Eq. (2.3) into account, it is seen that Eq. (2.10) is fulfilled if we assume

$$\dot{\hat{x}}(t) = g(\hat{x}(t)),$$

provided that we use the vector $|\psi(0)\rangle$ in order to calculate the expectation values of functions of $\dot{\hat{x}}$ or $g(\hat{x})$. In summary, the stochastic equation can be written as well as an equation involving time-dependent operators on a Hilbert space. The initial conditions on the stochastic equation are usually given as a set of relations involving the initial probability distributions of the processes. In the quantum-like formalism, this amounts to determining the operator $\hat{x}(0)\in\mathcal{M}$ and the state vector $|\psi(0)\rangle\in\mathcal{H}$. It seems that all these results could be generalized to stochastic equations more general than Eq. (2.8), and this will be assumed without proof in the following.

In many cases, stochastic equations relate some unknown stochastic processes with other processes which are known from the beginning. The most important of these are the stationary Gaussian processes, which are considered in the following. We define a stochastic process $Z(t)$ as stationary if its probability distribution is the same at all times and the new process

$$Y(\tau)\equiv Z(t+\tau)-Z(t) \tag{2.11}$$

depends on τ but not on t. In the quantum-like formalism, a Hermitian time-dependent operator $\hat{z}(t)$ must be associated with the process $Z(t)$ and another one, $\hat{y}(\tau)$, with $Y(\tau)$. The operators fulfil an equation similar to Eq. (2.11). A Gaussian distribution is obtained for $Y(\tau)$ if two adjoint operators $\hat{y}_+(\tau)$, $\hat{y}_-(\tau)$ are defined such that

$$\hat{y}_+(\tau)+\hat{y}_-(\tau)=\hat{y}(\tau), \tag{2.12}$$

$$\langle\psi|\hat{y}_-(\tau)=\hat{y}_+(\tau)|\psi\rangle=0,\quad [\hat{y}_+(\tau),\ \hat{y}_-(\tau)]=f(\tau). \tag{2.13}$$

In fact, it is well known from the quantum theory of the harmonic oscillator that Eqs. (2.12) and (2.13) give rise to the following probability distribution for the random variable $Y(\tau)$,[7]

$$\rho(Y)=[2\pi f(\tau)]^{-1/2}\exp[-Y^2/2\,f(\tau)]. \tag{2.14}$$

The function $f(\tau)$ can be easily related to the power spectrum of the process $Z(t)$. In fact, the square mean of $Z(t+\tau)-Z(t)$ is, by Eq. (2.14),

$$f(\tau)=\langle[Z(t+\tau)-Z(t)]^2\rangle=2\langle Z(t)^2\rangle-2\langle Z(t)Z(t+\tau)\rangle. \tag{2.15}$$

Here, the (ensemble) average $\langle Z(t)\,Z(t+\tau)\rangle$ is the correlation function of the stochastic process, which is independent of t for a stationary process. [It has been assumed that $Z(t)$ has zero mean, which can be made without loss of generality.] The correlation function is related to the power spectrum $G_z(\omega)$ through the Wiener–Khintchine theorem

$$\langle Z(t)\,Z(t+\tau)\rangle=\int_0^\infty G_z(\omega)\cos\tau\omega\,d\omega. \tag{2.16}$$

Hence, the function $f(\tau)$ is related to the power spectrum of $Z(t)$ as follows:

$$f(\tau)=2\int_0^\infty G_z(\omega)(1-\cos\tau\omega)\,d\omega. \tag{2.17}$$

It is useful to define a new operator function $\hat{u}(\omega)$ which is the Fourier transform of $\hat{z}(t)$, i.e.,

$$\hat{u}(\omega)\equiv(2\pi)^{-1}\int_{-\infty}^{\infty}\hat{z}(t)\exp(-i\omega t)\,d\omega,\quad \hat{u}(-\omega)=\hat{u}(\omega)^\dagger, \tag{2.18}$$

where the last equality is a consequence of $\hat{z}(t)$ being Hermitian. Now, instead of Eqs. (2.13), we postulate the following ones, which will be proved to imply the former:

$$\langle\psi|\hat{u}(-\omega)=\hat{u}(\omega)|\psi\rangle=0\ \text{ if }\omega>0,$$
$$[\hat{u}(\omega),\hat{u}(\omega')]=G_z(\omega)\ \operatorname{sgn}\omega\,\delta(\omega+\omega'). \tag{2.19}$$

The function sgn ω is included to assure that the right-hand side of Eq. (2.19) chages sign with the interchange $\omega\rightarrow\omega'$, as the left-hand one does. After that, two (non-Hermitian) operators can be defined by

$$\hat{z}_+(t)=\int_0^\infty\hat{u}(\omega)\exp(i\omega t)\,d\omega,\quad \hat{z}_-(t)=\int_{-\infty}^{0}\hat{u}(\omega)\exp(i\omega t)\,dw, \tag{2.20}$$

such that the following relations hold:

$$\begin{aligned}&\langle\psi|\hat{z}_-(t)=\hat{z}_+(t)|\psi\rangle=0,\\&[\hat{z}_+(t),\hat{z}_+(t')]=[\hat{z}_-(t),\hat{z}_-(t')]=0,\\&[\hat{z}_+(t),\hat{z}_-(t')]=\int_0^\infty G_z(\omega)\exp[i\omega(t-t')]\,d\omega,\\&\hat{z}_+(t)+\hat{z}_-(t)=\hat{z}(t).\end{aligned} \tag{2.21}$$

These equations, which can be easily derived from Eqs. (2.18) to (2.20), fully characterize the stationary Gaussian stochastic process (with zero mean) $Z(t)$ in terms of its associated operator $\hat{z}(t)$. It is trivial to show that these equations imply Eqs. (2.13) if we identify

$$\hat{y}_\pm(\tau)=\hat{z}_\pm(t+\tau)-\hat{z}_\pm(t). \tag{2.22}$$

Actually, Eqs. (2.18) and (2.19) or Eqs. (2.21) contain more information than Eqs. (2.11), (2.12), (2.13), and (2.17) because they imply that the Fourier components of the process $Z(t)$ are independent, besides being Gaussian. These two properties will be considered to define the stationary Gaussian processes from now on.

III. BROWNIAN MOTION

The theory of Brownian motion started in 1905 with the work of Einstein, who studied the motion of a Brownian particle in ordinary space in absence of forces. The forces were included in 1908 by Smoluchowski. This theory is imperfect in the sense that it introduces infinite velocities, a difficulty which was eliminated by Uhlenbeck and Ornstein in 1930 by developing the theory with the space of coordinates and velocities. Our purpose here is to illustrate the use of the quantum-like formalism in a typical example involving continuous stochastic processes. Therefore, we will study the simplest nontrivial case, i.e., the Einstein–Smoluchowski theory in one dimension.[8]

We begin by considering Brownian motion in the absence of forces. The physical assumption is that Brownian motion is a limiting case of the problem of random steps. This assumption leads to the result that if the particle is at point z at time t, the probability density at time t' will be

$$\rho(z',t';z,t)=(4D\pi|t-t'|)^{-1/2}\exp[-(z-z')^2/4D|t-t'|], \tag{3.1}$$

where the constant D is called diffusion coefficient. In order to apply the formalism of the preceding section, we compare Eq. (3.1) with Eq. (2.14). It follows that

$$f(\tau)=2D|\tau|,$$

so that the power spectrum of the stochastic process $Z(t)$ is, taking Eq. (2.17) into account,

$$G_z(\omega)=2D/\pi\omega^2. \tag{3.2}$$

The stationary Gaussian stochastic process with this power spectrum is called Wiener process. In the quantum-like formalism its properties are stated by Eqs. (2.21), whence it follows, taking Eq. (3.2) into account,

$$[\dot{\hat{z}}_+(t), \dot{\hat{z}}_-(t')] = (2D/\pi)\int_0^\infty \exp[i\omega(t-t')]\,d\omega$$
$$= 2D\delta(t-t') + (2Di/\pi)\mathcal{P}(t-t')^{-1}, \qquad (3.3)$$

where $\mathcal{P}$ means principal part. As a matter of fact, only the real part of the commutator is important because it fully characterizes the probability distribution of the process $Y(\tau)$ at each time (although not the joint probability distribution at two times) and this is all which is needed in the following.

The study of Brownian motion in the presence of forces is based in the following physical assumption: An external force $f(x)$ gives rise to a drift velocity proportional to it, which must be added to the random velocity. This can be written in terms of operators as follows:

$$d\hat{x}(t)/dt = \dot{\hat{z}}(t) + (m\beta)^{-1}f(\hat{x}), \qquad (3.4)$$

where $m\beta$ is some constant (written this way to agree with standard notation) and $\dot{\hat{z}}(t)$ is the operator associated with the random velocity, which is characterized by Eqs. (2.21) and (3.3). Equation (3.4) is the operator counterpart of Langevin equation, and should be considered the "Heisenberg equation" of Brownian motion. It is a differential equation involving time-dependent operators, as is the equation of Heisenberg in quantum mechanics.

It is not easy to solve Eq. (3.4) in general with the quantum-like formalism. (Later on, we will derive from it a Schrödinger-type equation which is more practical.) Therefore, we will study a simple illustrative example in which the solution is possible: the Brownian particle in an oscillator well. In this case, the external force $f(x)$ is given by

$$f(x) = -kx. \qquad (3.5)$$

Then, Eq. (3.4) can be written

$$d\hat{x}(t)/dt = \dot{\hat{z}}(t) - \alpha\hat{x}(t), \quad \alpha \equiv k/m\beta. \qquad (3.6)$$

This equation is easily integrated to give

$$\hat{x}(t) = \exp(-\alpha t)\left(\int_{t_0}^{t}\dot{\hat{z}}(\tau)\exp(\alpha\tau)\,d\tau + \hat{x}(t_0)\exp(\alpha t_0)\right). \qquad (3.7)$$

It is possible to define two adjoint operators $\hat{x}_+(t)$, $\hat{x}_-(t)$ related to $\dot{\hat{z}}_+(t)$ and $\dot{\hat{z}}_-(t)$ by equations similar to Eq. (3.7). Let us assume that the Brownian particle is at point x_0 at time t_0. This initial condition is easily incorporated into the formalism by writing

$$\hat{x}_+(t_0) = \hat{x}_-(t_0) = x_0/2, \qquad (3.8)$$

which imply that the particle is at point x_0 at time t_0 with probability one [as Eqs. (2.12) and (2.13) imply Eq. (2.14)]. From Eqs. (3.3), (3.7), and (3.8) it follows

$$[\hat{x}_+(t) - (x_0/2)\exp(\alpha t_0 - \alpha t)]\,|\psi\rangle$$
$$= \langle\psi|\,[\hat{x}_-(t) - (x_0/2)\exp(\alpha t - \alpha t_0)] = 0, \qquad (3.9)$$
$$[\hat{x}_+(t),\ \hat{x}_-(t)] = (D/\alpha)[1 - \exp(2\alpha t_0 - 2\alpha t)], \quad t > t_0.$$

These equations mean that the probability distribution is Gaussian at all times, the center approaches exponentially to the origin (the position of lowest potential energy) and the dispersion increases up to D/α. This is the well-known solution of the problem of a Brownian particle in an oscillator potential.[8] This example demonstrates both the power of the quantum-like formalism and its limitations. One of these is that some initial conditions might be very difficult (or even impossible) to incorporate into the formalism. This would be the case if the initial probability distributions were not Gaussian.

Let us consider now the general solution of the stochastic equation whose operator couterpart is Eq. (3.4). This equation can be used to derive the evolution of the expectation values of the observable quantities. For example, the observable (random variable) $X(t)^n$ has the operator $\hat{x}(t)^n$ associated with it. The time variation of this operator is given by

$$d\hat{x}^n/dt = nS(\hat{x}^{n-1}\dot{\hat{x}}) = nS(\hat{x}^{n-1}\dot{\hat{z}}) + (m\beta)^{-1}n\hat{x}^{n-1}f(\hat{x}). \qquad (3.10)$$

The symmetrizer S is introduced to take into account the possible noncommutativity of $\hat{x}(t)$ and $\dot{\hat{x}}(t)$ or $\hat{x}(t)$ and $\dot{\hat{z}}(t)$ [see Eq. (2.7)]. On the other hand, $\hat{x}$ and $f(\hat{x})$ obviously commute. The evolution of the expectation values is given by

$$d\langle X^n\rangle/dt = \langle\psi|\,d\hat{x}^n/dt\,|\psi\rangle = n\langle\psi|\,S(\hat{x}^{n-1}\dot{\hat{z}})\,|\psi\rangle$$
$$+ (m\beta)^{-1}n\langle\psi|\,\hat{x}^{n-1}f(\hat{x})\,|\psi\rangle. \qquad (3.11)$$

The first term of the right-hand side can be changed to a more useful form by using the operators $\dot{\hat{z}}_+$ and $\dot{\hat{z}}_-$:

$$n\langle\psi|\,S(\hat{x}^{n-1}\dot{\hat{z}})\,|\psi\rangle = \sum_{j=1}^{n}\langle\psi|\,\hat{x}^{j-1}(\dot{\hat{z}}_+ + \dot{\hat{z}}_-)\hat{x}^{n-j}\,|\psi\rangle. \qquad (3.12)$$

Now, the operator $\dot{\hat{z}}_+$ must be carried to the right of each bracket and the operator $\dot{\hat{z}}_-$ to the left, in order to make use of the second Eq. (3.3). Hence, assuming that the commutator of $\hat{x}$ and $\dot{\hat{z}}_+$ is a number (which will be proved later), we have

$$\langle\psi|\,S(\hat{x}^{n-1}\dot{\hat{z}})\,|\psi\rangle = (n-1)\langle\psi|\,\hat{x}^{n-2}\,|\psi\rangle\,\mathrm{Re}[\dot{\hat{z}}_+(t),\ \hat{x}(t)]. \qquad (3.13)$$

It remains to calculate the real part of the commutator $[\dot{\hat{z}}_+(t),\ \hat{x}(t)]$. To do this, let us assume that it is some function $g(t, t')$. Then it follows

$$\mathrm{Re}[\dot{\hat{z}}_+(t),\ \dot{\hat{x}}(t')] = \frac{\partial g(t,t')}{\partial t'},$$

$$\mathrm{Re}[\dot{\hat{z}}_+(t), f(\hat{x}(t'))] = \frac{d\hat{f}}{d\hat{x}}\,g(t,t');$$

whence, taking Eqs. (3.3) and (3.4) into account,

$$2D\delta(t-t') = \mathrm{Re}[\dot{\hat{z}}_+(t), \dot{\hat{z}}(t')] = \frac{\partial g(t,t')}{\partial t'} + \frac{\partial\hat{f}}{\partial\hat{x}}\,g(t,t'). \qquad (3.14)$$

In order to integrate this equation, the integration constant is fixed by the following causality condition: It is assumed that the real part of the commutator of $\hat{x}(t')$ and $\dot{\hat{z}}_+(t)$ is zero for $t' < t$. This is related to the fact that, when we follow the evolution into the future, the position at a time is independent of the stochastic velocity at later times. (The opposite would be true if we were interested in following the evolution into the past.) So, the solution of Eq. (3.14) is

$$\mathrm{Re}[\dot{\hat{z}}_+(t), \hat{x}(t')] = 2D\theta(t'-t). \qquad (3.15)$$

This is valid up to times t' slightly later than t. If $t' \gg t$ the last term of Eq. (3.14) should be taken into account. At time $t' = t$, the step function $\theta(t'-t)$ must be defined to take the value 1/2. It is not necessary to calculate the imaginary (anti-Hermitian) part of the commutator

$[\dot{\hat{z}}_+(t),\ \hat{x}(t)]$ {which is just $\frac{1}{2}[\dot{\hat{z}}(t),\ \hat{x}(t)]$}. We need only to assume that it commutes with $\hat{x}(t)$ in order to obtain Eq. (3.13). This is a simplifying assumption which amounts to fixing the integration constant of the imaginary counterpart of Eq. (3.14). After this, Eqs. (3.11), (3.13), and (3.15) lead to the following equation for the evolution of the observable quantities:

$$\frac{d\langle X^n\rangle}{dt} = Dn(n-1)\langle X^{n-2}\rangle + (m\beta)^{-1}n\langle X^{n-1}f(X)\rangle.$$

Hence, for any polynomial $M(X)$, it follows

$$\frac{d\langle M\rangle}{dt} = D\langle d^2M/dX^2\rangle + (m\beta)^{-1}\langle f(X)dM/dX\rangle. \tag{3.16}$$

This equation holds in both Schrödinger and Heisenberg pictures. Assuming now that we work in Schrödinger picture and the coordinate representation, we have

$$\langle\psi(t)|M(\hat{x})|\psi(t)\rangle = \int dx\,dx'\langle\psi(t)|x\rangle\langle x|M(\hat{x})|x'\rangle\langle x'|\psi(t)\rangle$$

$$= \int dx|\psi(x,t)|^2M(x). \tag{3.17}$$

After introducing the probability density ρ as the square modulus of the "wavefunction" $\psi(x,t) \equiv \langle x|\psi(t)\rangle$, this equation can be written

$$\frac{d}{dt}\int \rho(x,t)M(x)\,dx = D\int d^2M/dx^2\rho(x,t)\,dx$$

$$+ (m\beta)^{-1}\int f(x)\,dM/dx\rho(x,t)\,dx.$$

As this is true for all M, it follows after a number of integrations by parts

$$\partial\rho/\partial t = D\partial^2\rho/\partial x^2 - (m\beta)^{-1}\partial(f\rho)/\partial x. \tag{3.18}$$

This is the equation of Smoluchowski, which can be considered the "Schrödinger equation" of Brownian motion. Certainly, our derivation is not shorter than the usual ones with conventional techniques, but it illustrates quite well that the quantum-like formalism is suitable to deal with stochastic problems as classical as Brownian motion. Eq. (3.18) looks truly classical because only the square modulus, $\rho(x,t)$, of the "wave function", $\psi(x,t)$, appears in it, but neither this fact nor the opposite are essential for the quantum-like formalism presented in this paper.

It is interesting to look at Eq. (3.18) from another point of view. We define the (Hermitian) operator

$$\hat{p}(t) \equiv -im[\dot{\hat{z}}_+(t) - \dot{\hat{z}}_-(t)|, \tag{3.19}$$

such that, from Eq. (3.15), we have

$$2im\,\mathrm{Re}[\hat{x}(t),\ \dot{\hat{z}}_+(t)] \equiv [\hat{x}(t),\ \hat{p}(t)] = i(2mD). \tag{3.20}$$

This is similar to the basic commutation relation of quantum mechanics provided that $2mD$ is identified with $\hbar$. (This relation between the Planck constant and the diffusion coefficient has been considered in all attempts to reduce quantum mechanics to a Brownian-like stochastic theory.[2]) From Eqs. (3.4) and (3.20) it follows that the time derivative of the first Eq. (2.21) can be written

$$\dot{\hat{z}}_+(t)|\psi\rangle = (\hat{v} - (m\beta)^{-1}f(\hat{x}) + i\hat{p}/m)|\psi\rangle = 0, \quad \hat{v} \equiv \dot{\hat{x}}(t). \tag{3.21}$$

[Hence it seems obvious that the momentum of the Brownian particle cannot be identified with the operator $\hat{p}$ but, maybe, with $\hat{p}' \equiv \beta^{-1}f(\hat{x}) - i\hat{p}$. Nevertheless, this is not quite correct because neither $\hat{p}'$ is Hermitian nor is Eq. (3.21) an operator equation.] In the Schrödinger picture and the coordinate representation Eq. (3.21) is written

$$\hat{v}\psi(x,t) = (m\beta)^{-1}f(x)\psi(x,t) - \frac{2D\partial\psi(x,t)}{\partial x}. \tag{3.22}$$

[Remember that Eq. (3.20) implies $\hat{p} \equiv i(2mD)\partial/\partial x$ in the coordinate representation.] Now, it is plausible to identify the probability density current $j(x,t)$ with the real part of $\psi^*(x,t)\hat{v}\psi(x,t)$, as in quantum mechanics. If this identity is made, Eq. (3.22) leads to

$$j(x,t) \equiv (m\beta)^{-1}f(x)\rho(x,t) - \frac{D\partial\rho(x,t)}{\partial x}, \quad \rho \equiv |\psi|^2.$$

The equation of continuity for the current so defined is just the equation of Smoluchowski, Eq. (3.18).

IV. STOCHASTIC ELECTRODYNAMICS

Stochastic electrodynamics is just classical electrodynamics with the hypothesis of a random background radiation in the whole space.[4] The line of reasoning which leads to the idea of background radiation is as follows. In space there are systems of charged particles—which will be called atoms—moving according to classical laws. Then, the atoms will be continuously radiating and some amount of radiation will be always present in space. If this is so, the radiation will act on the atoms and these will arrive at a state of dynamical equilibrium such that the rate of emission equals the rate of absorption. This may explain in a simple way the stability of atoms, without departing from classical theories. Once the existence of some amount of radiation in space is assumed, its spectral density is fixed by very general principles. In fact, the only spectrum which is Lorentz invariant is the one which associates a mean energy of $\hbar\omega/2$ with each normal mode of the radiation.[4] The constant $\hbar$ measures the intensity of the background radiation and it is identified with the Planck constant on experimental grounds.

It must be pointed out that if other fields of force exist in nature besides the electromagnetic one, some background radiation of each type must exist. Therefore, it would be better to speak of a general dynamical stochastic theory and not just about stochastic electrodynamics. The general theory will not be considered in the present paper, which deals with a simple example: the motion of a single charged particle in the presence of background radiation.

The nonrelativistic equation of motion of a charged particle interacting with the background radiation is

$$m\ddot{\mathbf{r}} = \mathbf{f}(\mathbf{r}) + e\mathbf{E} + m\tau\dddot{\mathbf{r}}, \quad \tau \equiv 2e^2/3mc^3, \tag{4.1}$$

where m is the mass of the particle, e its charge, and c the speed of light. The right-hand side of Eq. (4.1) represents the total force acting on the particle. The first term is the (given) external force which is assumed to derive from a potential $V(\mathbf{r})$. The second term is due to the electric field of the background radiation (the magnetic force is neglected in the nonrelativistic limit). The last term is the damping due to the reaction on the particle of the radiation emitted by it. The electric field of the background radiation can be considered a stochastic process, which we assume Gaussian. The power spectrum of one of its components, say E_x, is fixed by Lorentz invariance, as was mentioned above. It is written[4]

$$G_E(\omega)=2\hbar\omega^3/3\pi c^3. \tag{4.2}$$

From now on, we will work in one dimension for simplicity.

It is convenient to transform Eq. (4.1), because it has some unphysical solutions (for example, if E and f were zero, a solution would be

$$x=x_0\exp(t/\tau),$$

which is absurd). It is not difficult to show that an equation which has the same solutions of Eq. (4.1) except the unphysical ones is the following (written already in terms of the operators associated with the stochastic processes which describe the motion):

$$m\ddot{\hat{x}}=m\ddot{\hat{z}}+\int_0^\infty f(\hat{x}(t+\tau s))\exp(-s)\,ds, \tag{4.3}$$

$$\ddot{\hat{z}}(t)\equiv e\int_0^\infty \hat{E}(t+\tau s)\exp(-s)\,ds, \tag{4.4}$$

where $\hat{z}(t)$ is the operator associated to the stochastic displacement. It is easy to show from Eqs. (4.2) and (4.4) that the power spectrum of the process $Z(t)$ is

$$G_z(\omega)=m\hbar\tau/[\pi|\omega|(1+\tau^2\omega^2)]. \tag{4.5}$$

Hence, taking Eqs. (2.21) into account, it follows

$$[\hat{z}(t),\ \dot{\hat{z}}(t')]=(2i\hbar\tau/\pi m)\int_0^\infty(1+\tau^2\omega^2)^{-1}\cos[\omega(t-t')]\,d\omega$$
$$=(i\hbar/m)\exp[-|t-t'|/\tau]. \tag{4.6}$$

If there are no forces acting on the particle except those of the random background (i.e., $f=0$), then Eqs. (4.3) and (4.6) imply

$$[\hat{x}(t),\ \dot{\hat{x}}(t)]=i\hbar/m+g(t,t), \tag{4.7}$$

where $g(t,t')$ is any (maybe operator) function fulfilling

$$\frac{\partial^3 g(t,t')}{\partial t^2\partial t'}=0. \tag{4.8}$$

With the choice $g=0$, Eq. (4.7) becomes the familiar commutation relation of quantum mechanics provided that the momentum operator is defined by

$$\hat{p}(t)\equiv m\dot{\hat{x}}(t). \tag{4.9}$$

The choice $g=0$ produces a great simplification of Eq. (4.7) but it is by no means essential. For instance, we might be interested in the motion of a particle which at time t_0 has a known position x_0 and a known velocity v_0. In this case,

$$\hat{x}(t_0)=x_0,\quad \dot{\hat{x}}(t_0)=v_0,$$

so that

$$[\hat{x}(t_0),\ \dot{\hat{x}}(t_0)]=0,\quad g(t_0,t_0)=-i\hbar/m. \tag{4.10}$$

Therefore, the present theory does not exclude the possibility of a simultaneous knowledge of the position and the velocity of the particle (the question whether they can be actually measured is beyond the scope of the present paper). Nevertheless, the theory resulting from Eq. (4.10) may become quite complex and the choice $g=0$ may be unavoidable in practice. In this case, we are going to show that the basic equations of the theory agree with the ones of quantum mechanics. In order to calculate in general the commutator $[\hat{x}(t),\ \dot{\hat{x}}(t)]$, we define

$$G(t,t')\equiv[\hat{x}(t),\ \hat{x}(t')]. \tag{4.11}$$

Then, Eq. (4.3) leads to

$$\frac{m^2\partial^4 G}{\partial t^2\partial t'^2}-\frac{m\partial^2}{\partial t^2}\int_0^\infty ds\left(\frac{d\hat{f}}{d\hat{x}}\right)_{t'+\tau s}G(t,t'+\tau s)\exp(-s)$$
$$-\frac{m\partial^2}{\partial t'^2}\int_0^\infty ds\left(\frac{d\hat{f}}{d\hat{x}}\right)_{t+\tau s}G(t+\tau s,\ t')\exp(-s)$$
$$+\int_0^\infty ds\int_0^\infty ds'\left(\frac{d\hat{f}}{d\hat{x}}\right)_{t+\tau s}\left(\frac{d\hat{f}}{d\hat{x}}\right)_{t'+\tau s'}G(t+\tau s,\ t'+\tau s')$$
$$=[\ddot{\hat{z}}(t),\ \ddot{\hat{z}}(t')]. \tag{4.12}$$

The right-hand side is singular for $t=t'$ [see Eq. (4.6)]. On the left-hand side, the most singular term will be the one with the highest derivative, which is the first one. Then, in the neighbourhood of $t=t'$ we must identify

$$[\ddot{\hat{x}}(t),\ \ddot{\hat{x}}(t')]=\frac{\partial^4 G(t,t')}{\partial t^2\partial t'^2}\approx[\ddot{\hat{z}}(t),\ \ddot{\hat{z}}(t')], \tag{4.13}$$

whence Eq. (4.7) is obtained again, although in this case $g(t,t')$ fulfils an equation more involved than Eq. (4.8). After this, the solution of particular problems might be accomplished by techniques similar to the ones used for Brownian motion in the previous section. Some results are rather trivial. For example, from Eq. (4.3) and the first Eq. (2.21) it can be obtained the following generalization of the Ehrenfest equations, which includes the spontaneous emission:

$$md^2\langle\psi|\hat{x}(t)|\psi\rangle/dt^2=\int_0^\infty\langle\psi|f(\hat{x}(t+\tau s))|\psi\rangle\exp(-s)\,ds. \tag{4.14}$$

Nevertheless, the solution of more general problems may become very difficult.

A substantial simplification is obtained in the limit $\tau\to 0$. This is equivalent to $e\to 0$ or $\alpha\equiv e^2/\hbar c\to 0$ [see Eq. (4.1)], which is the same limit considered in going from quantum electrodynamics to ordinary quantum mechanics. Therefore, in this limit we must obtain an alternative to quantum electrodynamics. In the limit $e\to 0$, Eq. (4.3) is simply written

$$m\ddot{\hat{x}}=f(\hat{x}(t))\equiv -dV(\hat{x})/d\hat{x}, \tag{4.15}$$

and this plus Eq. (4.7) (with $g=0$) is all which is needed to solve any problem of the motion. These two equations agree completely with those of elementary quantum mechanics. In fact, it is not difficult to prove that they are equivalent to the familiar ones of the Heisenberg picture of quantum mechanics [taking the definition Eq. (4.9) into account]:

$$[\hat{x}(t),\ \hat{p}(t)]=i\hbar,\quad d\hat{M}/dt=(i/\hbar)[\hat{H},\hat{M}]+\frac{\partial\hat{M}}{\partial t},$$
$$\hat{H}\equiv\hat{p}(t)^2/2m+V(\hat{x}(t)),\quad \hat{M}\equiv M(\hat{x}(t),\ \hat{p}(t),\ t). \tag{4.16}$$

Although these equations are the basic ones for both quantum mechanics and stochastic electrodynamics (this one in the limit $e\rightarrow 0$) these theories are not fully equivalent. Aside from possible differences in domains which have not been considered here (for instance, many particle systems and relativistic motion), a prime difference is that Eqs. (4.16) have been derived only for charged particles, whilst quantum mechanics postulates them for all particles, charged or neutral. This difficulty is not as big as it seems because the random motion of the particles, which is the physical idea behind the equations, is a consequence of the coupling between the particles and the random background radiation. It is probable that the equations can be derived independently of the nature and the strength of the coupling provided that this is small. [Note that Eqs. (4.16) have been here derived in the limit $e\rightarrow 0$, which physically means a very small, although not strictly zero, electrostatic coupling.] In the following, we consider more specifically the differences between the predictions of stochastic and quantum electrodynamics for the nonrelativistic motion of a single charged particle in a static potential in the limit $e\rightarrow 0$. In this case, quantum electrodynamics becomes ordinary quantum mechanics (QM) and stochastic electrodynamics leads to the theory just developed, which will be called SM (stochastic mechanics) for short.

In some sense, SM is more general than QM. In fact, Eqs. (4.16) have been derived after some arbitrary postulate was made, namely $g=0$ in Eq. (4.7), while in QM those equations are the basic ones. A consequence of the restrictions imposed on SM by the additional postulate is that only some probability distributions in phase space can be given as initial conditions. This eliminates, in particular, probability distributions violating the Heisenberg uncertainty relations, a constraint which is not inherent to stochastic electrodynamics. The problem of associating probability distributions in phase space with vectors in the Hilbert space is not trivial. As is well known,[1] this problem cannot be solved in quantum mechanics (i.e., there is no rule compatible with all the postulates of quantum mechanics which allows a probability distribution in phase space to be associated with every state-vector). In sharp contrast, the problem *must* have a solution in stochastic electrodynamics, this one being a fully classical theory. As a guess (which will be justified in subsequent papers[9]), we postulate that the probability distribution associated with a state vector $|\psi\rangle$ is such that the expectation value of any (polynomial) function $M(X,P)$ of the position coordinate and the momentum of the particle is given by

$$\langle M(X,P)\rangle = \langle\psi|SM(\hat{x},\hat{p})|\psi\rangle, \tag{4.17}$$

where S is the symmetrizer defined in Eq. (2.7). It is easy to see that this fully determines the probability distribution associated with the state vector $|\psi\rangle$. It must be emphasized that not every vector gives rise to a (positive definite) probability distribution fulfilling Eq. (4.17). On the other hand, there are (many) state vectors which have this property. For example, it can be shown[9] that all Gaussian wavepackets fulfil this condition.

In another sense, QM is more general than SM, because in QM every vector of the Hilbert space is assumed to correspond to a possible state of the motion (actually, experimental evidence has forced the elimination of some vectors in special cases giving rise to "super-selection rules"), while in SM only those vectors which can be associated with probability distributions in phase space represent physical states. In other words, states are represented by probability distributions in phase space in SM and by vectors of the Hilbert space in QM. There is a subset of the (projective) Hilbert space which corresponds one-to-one to some subset of the set of probability distributions in phase space. Only if all states needed to interpret the experiments belong to this subset will the predictions of SM and QM agree. This gives, in principle, a procedure to test experimentally SM versus QM. In the particular domain considered in the present paper—a single charged particle in a static potential—it seems very probable that no experiment test of QM can be performed. On the contrary, we must analyze whether the most significant experiments may also be interpreted by SM. A remark must be made before. Until now, only pure states have been considered. In QM mixed states are also defined, with which density matrices (rather than vectors) are associated. It seems very probable that this can also be made in SM and we do not consider it any more.

Most experiments in the atomic domain are related to scattering or spectroscopy. The scattering experiments are studied in QM by means of wavepackets. The predictions of SM will coincide with the ones of QM if (a) all wavepackets needed to represent the initial conditions of the scattering are physical according to SM [i.e., they can be associated to probability distributions in phase space through Eq. (4.17)] and (b) each one of these wavepackets, which represents a physical state at time t_0, remains a physical state at times $t>t_0$. Before studying these points in detail (which will not be made in this paper) it is not possible to say whether the predictions of SM agree with those of QM for scattering experiments. In some cases, the possibility of interpretation according SM seems difficult in an intuitive basis. In fact, it is not easy to imagine how the action of the background radiation may influence the motion of a particle to give rise to a wave-like behaviour, such as the one observed in the scattering of electrons by crystals. A possible intuitive explanation is that some paths of the electron in the periodic potential of the crystal are much more probable than the other ones, due to a resonance between the motion of the electron and some normal modes of the background radiation. This idea has some similarity with the hypothesis of de Broglie about the guidance of particles by the associated waves. The difference is that in stochastic electrodynamics there are no matter waves, just electromagnetic radiation.

In order to analyze a spectroscopic experiment let us consider the absorbtion of light by an atom. Actually, the absorbtion of light cannot be studied with the equations previously derived, but assuming that Eqs. (4.16) can be also derived in SM for time-dependent Hamiltonians, we would have an approximate theory equivalent to the semiclassical theory of radiation of QM. With these conditions, the probability distributions of the coordinates and momenta of the electrons change according to Eq. (4.16) and this is all both QM and SM predict about the atom. For example, in the electric dipole approximation, the absorbtion of radiation produces a change of the mean dipole moment of the atom, which is given by

$e\langle\psi|\hat{\mathbf{r}}(t)|\psi\rangle$. In particular, it may not be necessary to assume the existence of excited stationary states, which may not be physical according to SM. In other words, in QM all solutions of the eigenvalue equation

$$H|\phi_j\rangle = E_j|\phi_j\rangle \qquad (4.18)$$

are associated with physical states, but in SM this is not the case, for it is not possible in general to find a (nonnegative) probability distribution associated with every $|\phi_j\rangle$. In SM only the differences between pairs of eigenvalues of Eq. (4.18) have a direct physical meaning: They represent some characteristic frequencies for the motion of the particle in the potential in resonant interaction with the background radiation. The eigenvectors of the Hamiltonian operator are only needed as a practical aid for the solution of the Heisenberg equation of motion, but it is possible that all experiments may be interpreted without assuming that all these vectors represent physical states. Obviously, a more careful analysis of the significant experiments is needed.

Finally, let us analyze another difference between SM and QM related to the presence of the symmetrizer S in Eq. (4.17). This has far-reaching consequences for the expectation values. For instance, consider the energy of the ground state. In QM the ground state of the particle is represented by the eigenvector of Eq. (4.18) which has the lowest eigenvalue. Probably it can be shown that in SM a (nonnegative) probability distribution can be associated with this eigenvector for any potential, in which case, this will represent a physical state. The corresponding eigenvalue represents the mean energy of the state in both QM and SM, but there is a difference between these theories in that the energy is not sharply defined according to SM. In fact, the operator associated with the square of the energy is not $\hat{H}^2$, but $S(\hat{H}^2)$ according to Eq. (4.17). This can be written

$$S(\hat{H}^2) = S(\hat{p}^2/2m + V(\hat{x}))^2 = \hat{H}^2 + (\hbar^2/2m)\frac{d^2V(\hat{x})}{d\hat{x}^2}, \qquad (4.19)$$

where the last equality can be derived by repeated use of the commutation relations. Hence, the dispersion of the energy of a state is given by the square root of

$$\Delta E^2 \equiv \langle\psi|S(\hat{H}^2)|\psi\rangle - \langle\psi|\hat{H}|\psi\rangle^2 = \langle\psi|\hat{H}^2|\psi\rangle - \langle\psi|\hat{H}|\psi\rangle^2 + (\hbar^2/2m)\langle\psi|\, d^2\hat{V}/d\hat{x}^2\,|\psi\rangle, \qquad (4.20)$$

and only the last term remains for the ground state. An obvious necessary (but not sufficient) condition for a vector $|\psi\rangle$ to represent a physical state is that the right-hand side of Eq. (4.20) be nonnegative. It should be noted that a dispersionless energy would be compatible only with probability distributions in phase space which vanish outside the energy surface, which would be a very strong restriction.

As a summary, the present derivation of the quantum postulates in a limited domain does not prove that the full quantum theory can be replaced by a classical stochastic one, but it seems that this possibility must be considered seriously. It is obvious that many problems remain. It would be necessary to generalize the theory to systems of particles (both charged and neutral) and to relativistic motion, and also to develop a general theory of stochastic fields. Besides, the interpretation of experiments according to the theory should be carefully studied. All this will be dealt with elsewhere.

ACKNOWLEDGMENT

I should like to acknowledge Dr. J.F. Cariñena for helpful discussions.

[1] See, for example, L.E. Ballentine, Rev. Mod. Phys. 42, 358 (1970).

[2] Two main approaches have been developed. One of them, more phenomenological, is reviewed by E. Santos, in *Irreversibility in the Many-Body Problem*, edited by J. Biel and J. Rae (Plenum, New York, 1972). For the other (stochastic electrodynamics) see Ref. 4.

[3] H.G. Hall and R.E. Collins, J. Math. Phys. 12, 100 (1971), and references therein.

[4] E. Santos, Nuovo Cimento 19 B, 57 (1974), and references therein.

[5] See, for example, D. Ruelle in *Statiscal Mechanics and Quantum Field Theory*, edited by C. DeWitt and R. Stora (Gordon and Breach, New York, 1971), p. 234. The theory of C^*-algebras can be found in the same book and references therein.

[6] See, for example, Yu V. Prohorov and Yu. A. Rozanov, *Probability Theory* (Springer-Verlag, Berlín, 1969), p. 119.

[7] See, for example, P.A.M. Dirac, *Principles of Quantum Mechanics* (Clarendon, Oxford, 1958).

[8] The theory of Brownian motion can be found, for example, in *Selected Papers on Noise and Stochastic Processes*, edited by N. Wax (Dover, New York, 1954).

[9] E. Santos, "Foundations of stochastic electrodynamics," Nuovo Cimento (to be published).

The analytic noncharacteristic Cauchy problem for nonlightlike isometries in vacuum space-times

R. Berezdivin

Departamento de Física, Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela

(Received 31 January 1974; revised manuscript received 22 April 1974)

The analytic noncharacteristic problem for the existence of a spacelike isometry in vacuum space-time, given its existence on the hypersurfaces, on the Cauchy data, is posed and solved using the ADM equations. The timelike case is also solved. In both cases the isometry will locally exist.

I. INTRODUCTION

Isometries in general relativity have proved to be useful in finding exact solutions and in classifying space—times. The Cauchy problem for isometries has not, to our knowledge, been studied although it is of some relevance and inherently of interest. It would, when solved, give the conditions for a space—time to posses a local isometry once an isometry is detected on the Cauchy data: An isometry at each of two infinitesimally close hypersurfaces would, under certain conditions, imply the existence of a local isometry on space—time.

Cosmological observations could be related to this problem if the Cauchy hypersurfaces are characteristic: E.g., we know the universe to be, on a large scale, homogeneous and isotropic on our light cone; we could, if the proper conditions are satisfied, conclude it to be homogeneous and isotropic on space—time and thus get at the Robertson—Walker metrics. For spacelike hypersurfaces the result would be mathematically relevant,[1] possibly useful when performing calculations about the metric's evolution, and related to noncosmological observations. In both cases the calculations involved in the proofs of the theorems could be starting points for the calculation of the evolution of, e.g., small anisometries.

The related problem of the existence of a timelike isometry above a hypersurface, given its existence below it, has been solved.[2] It was necessary to rule out shock waves for the isometry above the hypersurface to exist so that for a general metric[1] C the conclusion that the isometry exists above the hypersurface is not true. To rule out such shock waves, we impose the condition of analyticity. The necessary conditions for the result remain to be found.

We use the ADM equations of evolution, as written in terms of Lie derivatives which we found to be well suited for the treatment of, in vacuum, the evolution of isometries from spacelike hypersurfaces. The characteristic Cauchy problem requires a different method. Our method was also not appropiate for lightlike isometries.

In Sec. II we briefly review the ADM equations of evolution and constraint, set the notation, and project spacelike Lie derivatives onto spacelike hypersurfaces; in Sec. III we derive the evolution equations for the isometry and solve the equations imposing conditions, explained in Sec. IV, which define the Killing vector field outside the initial hypersurface, completing the result for spacelike isometries in a theorem. The timelike case is treated in Sec. IV.

II. EVOLUTION AND LIE DERIVATIVES

The ADM equations of evolution and constraint[3,4] in empty regions of space—time can be written in a geometrical fashion appropriate for the treatment of isometries in terms of Lie derivatives as[5]

$$\partial_t g_{ij} = NK_{ij} + \mathcal{L}_X g_{ij}, \tag{1}$$

$$\partial_t K_{ij} = NS_{ij} - 2NR_{ij} + 2(\mathrm{Hess}N)_{ij} + \mathcal{L}_X K_{ij}, \tag{2}$$

$$(K^l{}_l g^{ij} - K^{ij})_{;j} = 0, \tag{3a}$$

and

$$\tfrac{1}{2}(K^l{}_l - K_{ij}K^{ij}) + 2R = 0. \tag{3b}$$

In this paper g_{ij} is the metric on the spacelike hypersurfaces labelled $t=\mathrm{const}$, all Latin indices ($i, j, k, \cdots = 1, 2, 3$) are raised and lowered with it and covariant derivatives are taken with respect to it. Any four-dimensional tensor, or component of one, not clearly four dimensional, will be labelled by a 4 (e.g., ${}_4\eta$, η^α, η^0, ${}_4\eta^k$, T_{0i}, ${}_4T_{ij}$, T_{00}) and Greek indices run from 0 to 3. N is the lapse function, X_i the shift vector,[6] $K_{ij} = 2\eta_{\alpha;\beta}\delta^\alpha_i\delta^\beta_j$ is twice the spatial projection of the usual extrinsic curvature of the $t=\mathrm{const}$ hypersurfaces with η_α the normals, R_{ij} is the Ricci tensor of g_{ij} and $R = R^i{}_i$. Also,

$$S_{ij} = K_{il}K^l{}_j - \tfrac{1}{2}(g^{lk}K_{lk})K_{ij}, \tag{4a}$$

$$(\mathrm{Hess}N)_{ij} = N_{,i;j}, \tag{4b}$$

and $\mathcal{L}_M P_{ij}$, for any M^k and P_{ij}, is the lie derivative of P_{ij} with respect to M^k, $\mathcal{L}_M P_{ij} = P_{ij,k}M^k + P_{ik}M^k{}_{,j} + P_{kj}M^k{}_{,i}$. The metric is

$$g_{\alpha\beta}\,dx^\alpha\,dx^\beta = (X^iX_i - N^2)dt^2 + 2X_i\,dx^i\,dt + g_{ij}\,dx^i\,dx^j. \tag{5}$$

Equations (3) are constraints which hold at all times if they hold initially by virtue of the Bianchi identities, and Eqs. (1) and (2) are the evolution equations.

Now, the Lie derivative of a four-dimensional tensor $P_{\alpha\beta}$ may be written as

$$\begin{aligned}\mathcal{L}_{4\eta} P_{\alpha\beta} &= P_{\alpha\beta,\gamma}\,\eta^\gamma + P_{\alpha\gamma}\eta^\gamma{}_{,\beta} + P_{\gamma\beta}\,\eta^\gamma{}_{,\alpha}\\ &= P_{\alpha\beta,0}\eta^0 + P_{\alpha 0}\eta^0{}_{,\beta} + P_{0\beta}\,\eta^0{}_{,\alpha}\\ &\quad + P_{\alpha\beta,k}\,{}_4\eta^k + P_{\alpha k}\,{}_4\eta^k{}_{,\beta} + P_{k\beta}\,{}_4\eta^k{}_{,\alpha},\end{aligned} \tag{6}$$

so that

$$\mathcal{L}_{{}_4\eta}\,{}_4P_{ij} = {}_4P_{ij,0} + {}_4P_{i0}\eta^0{}_{,j} + {}_4P_{0j}\eta^0{}_{,i} + \mathcal{L}_{{}_4\eta^k}\,{}_4P_{ij}. \tag{7}$$

We will identify ${}_4\eta_i$ and η_i, since the covariant spatial components of a four-tensor is a three-tensor on the t

hypersurfaces.[3] Similarly for any other tensor [such an identification is used in writing (1) → (3)]. We next find the relation between ${}_4\eta^k$ and η^k. We have[3]

$$ {}_4\eta^k = {}_4g^{kj}\eta_j + g^{k0}\eta_0 = (g^{kj} - X^kX^j/N^2)\eta_j + (X^k/N^2)\eta_0 $$

and

$$ \eta_0 = g_{00}\eta^0 + X_i\, {}_4\eta^i = (X_kX^k - N^2)\eta^0 + X_i[(g^{ij} - X^iX^j/N^2)\eta_j + (X^i/N^2)\eta_0], $$

so that

$$ {}_4\eta^k = \eta^k - X^k\eta^0. \tag{8} $$

Then, if we impose, as we shall later do for the spacelike case, $\eta^0 = 0$, (7) gives us

$$ \mathcal{L}_{{}_4\eta}\, {}_4P_{ij} = \mathcal{L}_\eta P_{ij}, \tag{9} $$

and (6) gives

$$ \mathcal{L}_{{}_4\eta} P_{0i} = P_{0i,k}\eta^k + P_{ik}\eta^k{}_{,0} + P_{k0}\eta^k{}_{,i} \tag{10} $$

and

$$ \mathcal{L}_{{}_4\eta} P_{00} = P_{00,k}\eta^k + P_{0k}\eta^k{}_{,0} + P_{k0}\eta^k{}_{,0} \tag{11} $$

Similarly, we would also have

$$ \mathcal{L}_{{}_4\eta} P^{00} = P^{00}{}_{,k}\eta^k. \tag{12} $$

Notice that then if (9) and (12) are zero, we will have, after a short calculation, if $g^{00} \neq 0$, that $\mathcal{L}_{{}_4\eta} g_{00} = 0$.

III. EVOLUTION OF ISOMETRY: SPACELIKE CASE

We will choose a coordinate system off the initial hypersurfaces such that g_{ij} and K_{ij} determine whether there is an isometry. If we start with two spacelike Killing vectors at $t=0$ and $t=dt$ but not lying in these hypersurface, we may find two hypersurfaces $x^{0'} = 0$ and $x^{0'} = dx^{0'}$ where they do lie, and by a change of coordinates call these the $t=0$ and $t=dt$ hypersurfaces and get the Cauchy data there. The coordinate system we shall construct will be such that $\eta^0 = 0$ so that the vector field η^α will have integral curves lying on the hypersurfaces $t = \text{const}$. This will be made clear in the next section. Let us here take it as given and see what (1) and (2) predict.

We first commute $\mathcal{L}_\eta$ with the time derivative operators. For any three-tensor P_{ij},

$$ \mathcal{L}_\eta(\partial_t P_{ij}) = \partial_t(\mathcal{L}_\eta P_{ij}) - [P_{ij,k}(\partial_t\eta^k) + P_{ik}(\partial_t\eta^k)_{,j} + P_{kj}(\partial_t\eta^k)_{,i}], \tag{13} $$

so that

$$ \mathcal{L}_\eta(\partial_t P_{ij}) = \partial_t(\mathcal{L}_\eta P_{ij}) - \mathcal{L}_{\partial_t\eta} P_{ij}. \tag{14} $$

We also use[7]

$$ \mathcal{L}_\eta \mathcal{L}_X - \mathcal{L}_X \mathcal{L}_\eta = \mathcal{L}_{\mathcal{L}_\eta X} = \mathcal{L}_{[\eta, X]} $$

and

$$ \mathcal{L}_{\eta+\gamma} = \mathcal{L}_\eta + \mathcal{L}_\gamma. $$

Then, from (1) and (2), after taking Lie derivatives, we obtain

$$ \partial_t(\mathcal{L}_\eta g_{ij}) = (\mathcal{L}_\eta N)K_{ij} + (\mathcal{L}_\eta K_{ij})N + \mathcal{L}_X(\mathcal{L}_\eta g_{ij}) + \mathcal{L}_{[\eta,X]+\partial_t\eta}\, g_{ij} \tag{15} $$

and

$$ \partial_t(\mathcal{L}_\eta K_{ij}) = (\mathcal{L}_\eta N)(S_{ij} - 2R_{ij}) + 2\mathcal{L}_\eta(N_{,i;j}) + N\mathcal{L}_\eta(S_{ij} - 2R_{ij}) + \mathcal{L}_X(\mathcal{L}_\eta K_{ij}) + \mathcal{L}_{[\eta,X]+\partial_t\eta} K_{ij}. \tag{16} $$

Hence, choosing η on the hypersurfaces by (we shall see exactly how in the next section)

$$ \mathcal{L}_\eta N = 0 \tag{17} $$

and

$$ [\eta, X] + \partial_t\eta = 0, \tag{18} $$

we will have

$$ \partial_t(\mathcal{L}_\eta g_{ij}) = N(\mathcal{L}_\eta K_{ij}) + \mathcal{L}_X(\mathcal{L}_\eta g_{ij}) \tag{19} $$

and

$$ \partial_t(\mathcal{L}_\eta K_{ij}) = \mathcal{L}_\eta(N_{,i;j}) + N\mathcal{L}_\eta(S_{ij} - 2R_{ij}) + \mathcal{L}_X(\mathcal{L}_\eta K_{ij}). \tag{20} $$

It will be clear, from Appendix A, that the first two terms of (20) are homogeneous in $\mathcal{L}_\eta g_{ij}$ or $\mathcal{L}_\eta K_{ij}$ and have a term in $\mathcal{L}_\eta N$ which by (17) is zero. Let us call these terms H_{ij}. It is shown below that if $\mathcal{L}_\eta g_{ij} = \mathcal{L}_\eta K_{ij} = 0$ at $t=0$, they will remain, if analytic on t, zero on a neighborhood of $t=0$. Of course, from (19) we see that this is equivalent to setting $\mathcal{L}_\eta g_{ij} = \partial_t(\mathcal{L}_\eta g_{ij}) = 0$ on $t=0$, since $\mathcal{L}_X(\mathcal{L}_\eta g_{ij})$ must be zero on $t=0$ as it is an internal derivative on $t=0$. Notice that

$$ (\mathcal{L}_\eta g_{ij})(t=dt) = \mathcal{L}_{\eta(t=dt)}\, g_{ij}(t=dt) $$

so that two infinitesimally close Killing η's on two infinitesimally close hypersurfaces are given by $\mathcal{L}_\eta g_{ij} = 0$ and $\partial_t(\mathcal{L}_\eta g_{ij}) = 0$ on $t=0$. Now for the demonstration that $\mathcal{L}_\eta g_{ij}$ remains zero.

At $t=0$, H_{ij} is zero as it is homogeneous in terms which are zero. $\mathcal{L}_X(\mathcal{L}_\eta g_{ij}$ or $\mathcal{L}_\eta K_{ij}) = 0$ as $\mathcal{L}_X$ involves operations on $t=0$ and $\mathcal{L}_X(0) = 0$. Thus, from (19) and (20), on $t=0$, $\partial_t(\mathcal{L}_\eta g_{ij}) = \partial_t(\mathcal{L}_\eta K_{ij}) = 0$. Then since $(\partial_t H_{ij})(0)$ must be zero, as it involves terms in $\mathcal{L}_\eta g_{ij}$ or $\mathcal{L}_\eta K_{ij}$ or $\partial_t(\mathcal{L}_\eta g_{ij})$ or $\partial_t(\mathcal{L}_\eta K_{ij})$ which are all zero, if we take ∂_t of (19) and (20) we obtain

$$ \partial_t^2(\mathcal{L}_\eta g_{ij}) = (\partial_t N)\mathcal{L}_\eta K_{ij} + N\partial_t(\mathcal{L}_\eta K_{ij}) + \mathcal{L}_X(\partial_t(\mathcal{L}_\eta g_{ij})) + \mathcal{L}_{\partial_t X}(\mathcal{L}_\eta g_{ij}) \tag{21} $$

and

$$ \partial_t^2(\mathcal{L}_\eta K_{ij}) = \partial_t H_{ij} + \mathcal{L}_X(\partial_t(\mathcal{L}_\eta K_{ij})) + \mathcal{L}_{\partial_t X}(\mathcal{L}_\eta K_{ij}), \tag{22} $$

so that $\partial_t^2(\mathcal{L}_\eta g_{ij}) = \partial_t^2(\mathcal{L}_\eta K_{ij}) = 0$ at $t=0$ (all terms on the right are zeroes or internal derivatives of zeroes). Then, again, $\partial_t^2 H_{ij} = 0$, and we may continue the process indefinitely to get $\partial_t^n(\mathcal{L}_\eta g_{ij}) = 0$ for all n. If $\mathcal{L}_\eta g_{ij}$ is analytic in t, we obtain $\mathcal{L}_\eta g_{ij} = 0$ on a neighborhood (call it R) of $t=0$, and $\mathcal{L}_\eta K_{ij} = 0$ on R.

IV. COORDINATE CONDITIONS AND CONSTRUCTING A KILLING VECTOR

Let us start from the end. If ${}_4\eta$, lying on a spacelike hypersurface which we call $t=0$, satisfies $\mathcal{L}_{{}_4\eta} g_{\alpha\beta}$

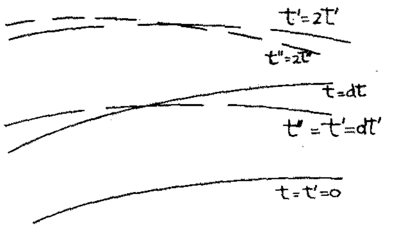


FIG. 1. $-t'=\text{const}$ and $t''=\text{const}$ are two families of spacelike hypersurfaces. We choose the time $=$ constant hypersurfaces $t=0$, $t'=dt'$, $t''=dt''$, and so on.

$=\mathcal{L}_{{}_4\eta}K_{\alpha\beta}=0$ at $t=0$, or equivalently since (17) and (18) must hold at $t=0$ (see below), if $\mathcal{L}_{{}_4\eta}g_{\alpha\beta}=\partial_t(\mathcal{L}_{{}_4\eta}g_{\alpha\beta})=0$ on $t=0$, the Cauchy problem here, and if we choose (17) and (18) and $\eta^0=0$ (see below), then using the result of Sec. III and (9) we must have $\mathcal{L}_{{}_4\eta}g_{ij}=\mathcal{L}_{{}_4\eta}K_{ij}=0$ on R. Then by (12) and (17) $\mathcal{L}_{{}_4\eta}g^{00}=0$, and (10) gives

$$\mathcal{L}_{{}_4\eta}\,g_{0i}=X_{i,k}\eta^k+g_{ik}(\partial_t\eta^k)+X_k\eta^k{}_{,i}=\mathcal{L}_\eta X_i+g_{ik}\partial_t\eta^k,$$

or, multiplying by g^{il},

$$g^{il}(\mathcal{L}_{{}_4\eta}g_{0i})=\mathcal{L}_\eta X^l+\partial_t\eta^l,$$

which by (18) is zero. Then, from the concluding remarks of Sec. II we must have $\mathcal{L}_{{}_4\eta}g_{\alpha\beta}=0$ on R.

Let us now clarify (17) and (18), and the choice of ${}_4\eta$. Given a space–time analytic (at least in the coordinates chosen below) in a region about $t=0$, with coordinates (t,x^i) and metric (g_{ij},X_i,N) and such that $\mathcal{L}_{{}_4\eta}g_{\alpha\beta}=\mathcal{L}_{{}_4\eta}K_{\alpha\beta}=0$ on $t=0$ with ${}_4\eta$ lying on $t=0$, then (18) gives $\eta^i(t=dt)$ as in Fig. 1. This may not satisfy (17) on $t=dt$. We may, however, define another hypersurface $t'=dt'(x^i,t)=\text{const}$ with $t'=0$ identically $t=0$, that is, a change of the time variable off $t=0$. This may change $N(t=0)$ and $X_i(t=0)$, but it leaves $g_{ij}(t=0)$ and $K_{ij}(t=0)$ unchanged.[3] We may, of course, also change the x^i coordinates to $x^{i'}$ on $t'=dt'$. We may, this way, construct $\eta^i(t'=dt')$ from (18) such that (17) is also satisfied at $t'=dt'$. Since one may map the initial value problem on any coordinate system to one in the Gaussian normal coordinate system,[5] we may, for example, find $t'=dt'$ such that $X_i=0$, $N^2=+1$, on $t'=0$ and then (18) gives $\eta^i(dt')=\eta^i(0)$ and (17) holds on 0; the K_{0i} and K_{00} conditions will also still hold: $K_{00}=K_{0i}=0$ on 0 (see Appendix B) and, by (10), (11), and (18), $\mathcal{L}_{{}_4\eta}K_{0i}=\mathcal{L}_{{}_4\eta}K_{00}=0$ on 0. Next we may redefine a next hypersurface $t''=2\,dt''$, with $t''=t'$ on dt', such that N, X_i and η^i satisfy (17) and (18) on $t''=2dt''$. This may be continued indefinitely to fill out the neighborhood R of $t=0$, defining $\eta^0=0$ so that ${}_4\eta$ lies on the then time $=$ constant hypersurfaces. The result that $\mathcal{L}_{{}_4\eta}g_{\alpha\beta}$ equals zero is coordinate invariant and so our result is the following.

Theorem: If there exists, in an analytic (in a Gaussian normal, or analytically related to it, coordinate system) region of space–time, a spacelike hypersurface $t=0$ such that $\mathcal{L}_{{}_4\eta}g_{\alpha\beta}=0$, and $\mathcal{L}_{{}_4\eta}K_{\alpha\beta}=0$ or $\partial_t(\mathcal{L}_{{}_4\eta}g_{\alpha\beta})=0$, on $t=0$ for some spacelike vector field ${}_4\eta$ lying on $t=0$, then there will exist a neighborhood of $t=0$ where there will exist a Killing vector, that is, R admits an isometry, which reduces to ${}_4\eta$ on $t=0$.

Notice that $\mathcal{L}_\eta$ of the constraint equations (3) give identically zero for $\mathcal{L}_\eta g_{ij}=\mathcal{L}_\eta K_{ij}=0$ so that (3) will not cause any difficulty.

V. TIMELIKE CASE

Again, the Cauchy isometry, from (24), (27), and (28) below, is equivalent to $\mathcal{L}_{{}_4\eta}g_{\alpha\beta}=\mathcal{L}_{{}_4\eta}K_{\alpha\beta}=0$ initially. Choosing t such that $\eta^\alpha=\delta^\alpha_t$, so that $\mathcal{L}_{{}_4\eta}P_{\alpha\beta}\equiv\partial_tP_{\alpha\beta}$, let us consider the following.

A. $t=0$ is the initial hypersurface:

$$\begin{aligned}&\text{Then } \partial_t g_{ij}=0 \quad \text{on } t=0 \text{ and } t=dt,\\ &\partial_tN=\partial_tX_i=0 \quad \text{on } t=0 \text{ and } t=dt,\end{aligned}\tag{23}$$

and, taking time derivatives in (1) and (2) and using (14), we have

$$\partial^2_tg_{ij}=(\partial_tN)K_{ij}+N\partial_tK_{ij}+\mathcal{L}_X(\partial_tg_{ij})+\mathcal{L}_{\partial_tX}g_{ij}\tag{24}$$

and

$$\begin{aligned}\partial^2_tK_{ij}=&(\partial_tN)(S_{ij}-2R_{ij})+N\partial_t(S_{ij}-2R_{ij})\\ &+2\partial_t(N_{,i;j})+\mathcal{L}_X(\partial_tK_{ij})+\mathcal{L}_{\partial_tX}K_{ij}.\end{aligned}\tag{25}$$

From (23), $(\partial^2_tg_{ij})(t=0)=0$, so from (24), $(\partial_tK_{ij})(0)=0$, and then from (25), $(\partial^2_tK_{ij})(0)=0$; back to (24), since $\partial^2_tN=\partial^2_tX=0$ on $t=0$, we get $\partial^3_tg_{ij}=0$, and from (25), $\partial^3_tK_{ij}=0$. We next redefine the coordinate system such that $\partial^3_tN=\partial^3_tX=0$, as before in the spacelike case by a change of coordinates, and we obtain this way, continuing the process, that the neighborhood R, under the same analyticity conditions as before, will admit an isometry.

B. $t=0$ is not the initial hypersurface, it is $x^0=0$: simply by reparametrizing along η^α, that is, a change of coordinates off $t=0$, we may call $\eta^\alpha=\delta^\alpha_0$ and all follows as in A. Our conclusion is the following

Theorem: If there exists, in a analytic (in a Gaussian normal, or analytically related to it, coordinate system) region R of vacuum space–time, a spacelike hypersurface $t=0$ such that $\mathcal{L}_{{}_4\eta}g_{\alpha\beta}=0$, and $\mathcal{L}_\eta K_{\alpha\beta}=0$ or $\partial_t(\mathcal{L}_{{}_4\eta}g_{\alpha\beta})=0$, on $t=0$ for some timelike vector field ${}_4\eta$, then there will exist a neighborhood of $t=0$ where there will exist a Killing vector that is, R admits an isometry, which reduces to ${}_4\eta$ on $t=0$.

ACKNOWLEDGMENTS

I thank L. Herrera and C. Aragone for some helpful discussions.

APPENDIX A: H_{ij} $(G_{ij}, K_{ij}, L_\eta g_{ij},\ \mathcal{L}_\eta K_{ij},\ \mathcal{L}N)$

$2\mathcal{L}_\eta(N_{,i;j})+N\mathcal{L}_\eta(S_{ij}-2R_{ij})$ equals, by using the commutation relation between covariant and Lie differentiation,[10] Eqs. (4), and writing[11] $\mathcal{L}_\eta R_{ij}$ and $\mathcal{L}_\eta\Gamma^i_{jk}$ in terms of $\mathcal{L}_\eta g_{ij}$,

$$\begin{aligned}H_{ij}=&2[\mathcal{L}_\eta(N_{,i})]_{;j}-2(\mathcal{L}_\eta\Gamma^k_{ij})N_{,k}+N\{(\mathcal{L}_\eta K_{il})K^l{}_j\\ &+K^m{}_i(\mathcal{L}_\eta K_{mj})+K_{mj}K_{il}(\mathcal{L}_\eta g^{ml})-\tfrac{1}{2}[(\mathcal{L}_\eta g^{lk})K_{lk}K_{ij}\\ &+g^{lk}(\mathcal{L}_\eta K_{lk})K_{ij}+K^l{}_l\mathcal{L}_\eta K_{ij}]-2\mathcal{L}_\eta R_{ij}\end{aligned}$$

so that

$$\begin{aligned}H_{ij}=&2(\mathcal{L}_\eta N)_{,i;j}\\ &-g^{lk}[(\mathcal{L}_\eta g_{il})_{;j}+(\mathcal{L}_\eta g_{lj})_{;i}-(\mathcal{L}_\eta g_{ij})_{;l}]N_{,k}\end{aligned}$$

$$\begin{aligned}&+N\{K^l{}_j\mathcal{L}_\eta K_{il}+K^m{}_i\mathcal{L}_\eta K_{mj}+K_{mj}K_{il}\mathcal{L}_\eta g^{ml}\\&-\tfrac{1}{2}[K_{lk}(\mathcal{L}_\eta g^{lk})K_{ij}+g^{lk}K_{ij}\mathcal{L}_\eta K_{lk}+K^l{}_l\mathcal{L}_\eta K_{ij}]\\&+g^{km}[(\mathcal{L}_\eta g_{km})_{;j}+(\mathcal{L}_\eta g_{mj})_{;k}-(\mathcal{L}_\eta g_{jk})_{;m}]_{;i}\\&-g^{km}[(\mathcal{L}_\eta g_{im})_{;j}+(\mathcal{L}_\eta g_{mj})_{;i}-(\mathcal{L}_\eta g_{ij})_{;m}]_{;k},\end{aligned}\qquad\text{(A1)}$$

where we have also used $\mathcal{L}_\eta(N_{,i})=(\mathcal{L}_\eta N)_{,i}$, and we may further use $\mathcal{L}_\eta g^{mn}=-g^{ml}g^{kn}\mathcal{L}_\eta g_{lk}$.

APPENDIX B: K_{00}, K_{0i}, AND COORDINATE CONDITIONS

With $K_{\alpha\beta}=2\eta_{\alpha;\beta}$ as we have used, with $\eta_\alpha=N\delta^0_\alpha$,

$$K_{0i}=-2N\Gamma^0_{0i}=-Ng^{\alpha 0}[g_{\alpha 0,i}+g_{\alpha i,0}-g_{0i,\alpha}]$$

so

$$K_{0i}=g_{00,i}/N-(X^j/N)(X_{j,i}-X_{i,j}+g_{ij,0}),\qquad\text{(B1)}$$

and similarly

$$K_{00}=(1/N)g_{00,0}-(X^j/N)(2\partial_t X_j-g_{00,j}).\qquad\text{(B2)}$$

Thus, one may see that $K_{00}=K_{0i}=0$ when $X_j=0$, $N^2=1$.

[1] See, for example, A. Z. Petrov, *Einstein Spaces* (Pergamon, New York, 1969). Also, e.g., O. Heckmann and E. Schücking, in *Gravitation, An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).

[2] A. Papapetrou, J. Math Phys. **6**, 1405 (1965) and Mme. I. Moret-Bailly, Ann. Inst. Henri Poincaré A **IX**, 395 (1968).

[3] R. Arnowitt, S. Deser, and C.W. Misner, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).

[4] J.A. Wheeler, in *Relativity, Groups and Topology*, edited by B. Dewitt and C. Dewitt (Gordon and Breach, New York, 1964).

[5] A.E. Fischer and J.E. Marsden, J. Math Phys. **13**, 546 (1972).

[6] We use the conventions and numerics of Ref. 5, except that instead of X_i we use $-X_i$ so that $g_{\alpha\beta}$ and $g^{\alpha\beta}$ are as in Ref. 3.

[7] R. L. Bishop and S.I. Goldberg, *Tensor Analysis on Manifolds* (Macmillan, New York, 1968). These, as well as Leibnitz rule, hold for all objects and "vectors" here considered such as N and $\partial_t\eta$.

[8] Y. Bruhat, in Ref. 3.

[9] I.G. Petrovsky, *Lectures on Partial Differential Equations* (Interscience, New York, 1950).

[10] K. Yano. *The Theory of Lie Derivatives and Its Applications* (North-Holland, Amsterdam, 1955).

[11] G.H. Katzin, J. Levine, and W.R. Davis, J. Math. Phys. **10**, 617 (1969).

Concerning a criterion for the validity of the first order smoothing approximation

I. Lerche

Enrico Fermi Institute and Department of Physics, University of Chicago, Chicago, Illinois 60637

This paper presents some statistically exact variational principles for problems in random function theory for the purpose of obtaining a criterion which will test the validity of an approximate method known variously in the many different fields of its application as first order smoothing theory, first order cumulant discard, quasilinear theory, or the adiabatic approximation. The hydromagnetic dynamo equations are used here, as particular mathematical instances of general mathematical points. The calculations show that when the random equation under investigation is self-adjoint, and when the quasilinear approximation to it is also self-adjoint, then the exact and approximate solutions will, almost surely, agree. When either (or both) the true equation or the quasilinear approximate equation is not self-adjoint, then the exact and approximate solutions will, almost surely, disagree.

I. INTRODUCTION

A large portion of the more formal work on hydromagnetic dynamos, plasma turbulence, cosmic ray diffusion, etc. is based on a popular mathematical maneuver known variously as first-order smoothing theory, first order cumulant discard, quasilinear theory, first order random Born approximation, or the adiabatic approximation.[1-3] First order smoothing conjectures that the terms which are nonlinear in the random functions do not differ much, or for very long, from their mean values, so that the difference can be neglected. The general validity of first order smoothing has not been established.

A number of authors have been concerned about the validity of the method, which neglects all mode—mode coupling (see review by Frisch[3] and references therein). A number of authors[4-7] have shown that mode—mode coupling is important in many cases in plasma physics, although there are also some situations in which the coupling can apparently be neglected with impunity.[8,9] Kraichnan[10] has established the invalidity of first order smoothing in special cases. Herring[11] gives detailed comparisons of the quasilinear, quasinormal, and Kraichnan's direct interaction approximation with numerical solutions of Boussinesq, convection, and finds that only the direct interaction approximation agrees closely with the numerical solutions.

Our own interest in the problem originally arose in connection with the hydromagnetic dynamo and the origin of the large-scale turbulent and ordered magnetic fields in astrophysical bodies. Consequently, the particular mathematical examples which we present to illustrate general mathematical points are chosen from that domain—although any other domain would have served the same purpose equally well.

The question to answer is: Does there exist a general criterion which will indicate, *ahead of any detailed calculations*, when first order smoothing will provide an accurate solution to the "true" problem? That there are *post facto* criteria is obvious, for a calculation can be worked through using both first order smoothing and also using statistically exact methods and comparison of the resulting answers can be made. This has, of course, been done.[9] However, it suffers from the disadvantage of being particular (to a given problem and equations) rather than general, and it normally involves some considerable calculation. What we seek is a *general* criterion that requires little work and which is not specific to any one problem or equation. We believe that we have found such a criterion and we present it here.

In Sec. II we shall first show how the criterion can be developed for any given set of equations (we shall use the dynamo equations as illustrative). In Sec. III we will set up the general criterion in the form of a "theorem."

II. KINEMATIC DYNAMOS AND QUASILINEAR THEORY

In their simplest form (strong shear, rectangular geometry, and large dynamo number) the dynamo equations, describing the generation of field by shear and cyclonic turbulence, are[12,13]

$$\frac{\partial B_y}{\partial t} = G\frac{\partial A}{\partial x}, \tag{1}$$

$$\frac{\partial A}{\partial t} = \Gamma B_y, \tag{2}$$

where $G \equiv dV_y/dz$ represents the large-scale shear and the Γ represents the cyclonic velocity component. In this example the vector potential $A(x,t)$ is in the y direction and the z component of field (sheared by G) is $\partial A/\partial x$. We shall consider the irregularities in the field produced by a random variation in Γ.

A. General considerations

1. Γ a random function of time

Consider first the case

$$G = G_0, \quad \Gamma = \Gamma_0[1 + \epsilon\,\delta\Gamma(t)], \tag{3}$$

where Γ_0, G_0, ϵ are constants and $\delta\Gamma(t)$ has zero mean value, $\langle\delta\Gamma\rangle = 0$. Then the two dynamo equations can be written

$$\frac{\partial^2 B_y}{\partial t^2} - \Gamma_0 G_0[1 + \epsilon\,\delta\Gamma(t)]\,\frac{\partial B_y}{\partial x} = 0. \tag{4}$$

Let $h = t/T$, where T is the correlation time for $\delta\Gamma$. Then let

$$B_y = B(h)\exp(ikx). \tag{5}$$

Equation (4) reduces to

$$\frac{d^2B}{dh^2} - ik\Gamma_0 G_0 T^2[1 + \epsilon\delta\Gamma(h)]B = 0. \tag{6}$$

First order smoothing theory writes $B=\langle B\rangle+\delta B$ so that, upon averaging Eq. (6), we obtain the statistically *exact* equation

$$\frac{d^2\langle B\rangle}{dh^2}-ik_0\Gamma_0G_0T^2[\langle B\rangle+\epsilon\langle\delta\Gamma(h)\delta B(h)\rangle]=0. \tag{7}$$

If this is subtracted from Eq. (6), we obtain the statistically exact equation

$$\begin{aligned}\frac{d^2\delta B}{dh^2}&-ik\Gamma_0G_0T^2[\delta B+\epsilon\delta\Gamma\langle B\rangle]\\&=ik\Gamma_0G_0T^2_\epsilon[\delta\Gamma\delta B-\langle\delta\Gamma\delta B\rangle].\end{aligned} \tag{8}$$

First order smoothing theory decrees that $\delta\Gamma\delta B$ never differs by much, or not for long, from $\langle\delta\Gamma\delta B\rangle$ so that the right-hand side of Eq. (8) can be neglected. Then we have the inexact equation

$$\frac{d^2\delta B}{dh^2}-ik\Gamma_0G_0T^2\delta B=ik\Gamma_0G_0T^2\epsilon\,\delta\Gamma\langle B\rangle, \tag{9}$$

which is then solved simultaneously with the exact equation (7) for the average field.

2. Γ *a random function of space*

Consider the case

$$G=G_0,\quad \Gamma=\Gamma_0[1+\epsilon\delta\Gamma(x)]. \tag{10}$$

Let $y=x/L$, where L is the correlation length for $\delta\Gamma$. Then let

$$B_y=B(y)\exp(-i\omega t), \tag{11}$$

when Eqs. (1) and (2) yield

$$\omega^2LB+\Gamma_0G_0[1+\epsilon\delta\Gamma(y)]\frac{dB}{dy}=0. \tag{12}$$

Again with $B=\langle B\rangle+\delta B$ we obtain

$$\frac{d\langle B\rangle}{dy}+\frac{\omega^2L}{\Gamma_0G_0}\langle B\rangle+\Gamma_0G_0\epsilon\left\langle\delta\Gamma(y)\frac{d\delta B}{dy}\right\rangle=0, \tag{13}$$

together with

$$\begin{aligned}\frac{d\delta B}{dy}&+\frac{\omega^2L}{\Gamma_0G_0}\,\delta B+\epsilon\delta\Gamma\,\frac{d\langle B\rangle}{dy}\\&=-\epsilon\left(\delta\Gamma\,\frac{d\delta B}{dy}-\left\langle\delta\Gamma\,\frac{d\delta B}{dy}\right\rangle\right).\end{aligned} \tag{14}$$

Once again first order smoothing sets the terms on the right-hand side of Eq. (14) to zero. We then have the inexact equation

$$\frac{d\delta B}{dy}+\frac{\omega^2L}{\Gamma_0G_0}\,\delta B=-\epsilon\,\delta\Gamma\,\frac{d\langle B\rangle}{dy}, \tag{15}$$

which is solved simultaneously with the exact equation (13) for the average field.

The normal modes that result under the first order smoothing approximation for the average field, $\langle B\rangle$, have been compared and contrasted[9] with those obtaining under a statistically exact analysis of Eqs. (6) and (12).

It is found that when Γ is a random function of time (case 1) precise and exact agreement of the normal mode frequencies obtains whether one uses the statistically exact treatment or the first order smoothing approximation. However, when Γ is a random function of space (case 2), no agreement is found except in the limit $\epsilon=0$. For ϵ small, but not precisely zero, there is strong disagreement between the first order smoothing result and the statistically exact result (but see Lerche and Parker[9] for a detailed analysis).

B. Lagrangians and adjoint fields

1. Γ *a random function of time*

Given Eq. (6), we have a Lagrangian

$$L=\int dh\left(\frac{dB^\dagger}{dh}\;\frac{dB}{dh}+ik\Gamma_0G_0T^2BB^\dagger(1+\epsilon\delta\Gamma)\right), \tag{16}$$

so that if L is varied extremally with respect to $B^\dagger$ we recover Eq. (6), while if L is varied extremally with respect to B we obtain the equation adjoint to Eq. (6) as

$$\frac{d^2B^\dagger}{dh^2}-ik\Gamma_0G_0T^2(1+\epsilon\delta\Gamma)B^\dagger=0. \tag{17}$$

Thus Eq. (6) is self-adjoint, $B=B^\dagger$, for a given $\delta\Gamma(h)$. Now it can be argued that, since $\delta\Gamma(h)$ is taken to be a random function of time, h, we should write a Lagrangian for the average field, $\langle B\rangle$, in order to compare it, and its adjoint, with the quasilinear results. So write L of expression (16) as $L=L_0+\Delta L$, where

$$\begin{aligned}L_0=\int dh&\left(\frac{d\langle B\rangle}{dh}\;\frac{d\langle B^\dagger\rangle}{dh}+\left\langle\frac{d\delta B}{dh}\;\frac{d\delta B^\dagger}{dh}\right\rangle\right.\\&+ik\Gamma_0G_0T^2(\langle B^\dagger\rangle\langle B\rangle+\langle\delta B^\dagger\delta B\rangle\\&\left.+\epsilon\langle B^\dagger\rangle\langle\delta\Gamma\delta B\rangle+\epsilon\langle B\rangle\langle\delta B^\dagger\delta\Gamma\rangle+\epsilon\langle\delta B^\dagger\delta B\delta\Gamma\rangle)\right).\end{aligned} \tag{18}$$

Suppose now that we vary the average Lagrangian L_0 extremally with respect to $\langle B^\dagger\rangle$; then we recover Eq. (7), while if L_0 is varied extremally with respect to $\langle B\rangle$, we obtain the adjoint equation

$$\frac{d^2\langle B^\dagger\rangle}{dh^2}-ik\Gamma_0G_0T^2(\langle B^\dagger\rangle+\epsilon\langle\delta B^\dagger\delta\Gamma\rangle)=0. \tag{19}$$

Suppose further that, in the ensembling process, we vary L_0 with respect to the statistical character of $\delta B^\dagger$. Then extremal variation of L_0 with respect to such changes gives

$$\frac{d^2\delta B}{dh^2}-ik\Gamma_0G_0T^2(\delta B+\epsilon\delta\Gamma\langle B\rangle)=ik\epsilon\,\Gamma_0G_0T^2\langle\delta\Gamma\delta B\rangle. \tag{20}$$

But this gives a finite average value to δB, i.e., $\langle\delta B\rangle\neq0$ whereas our premise is $\langle\delta B\rangle=0$. The reason that this occurs is, of course, that it is L that must be varied extremally for each realization and *not* L_0. And when this is done an extra term, $-ik\epsilon\Gamma_0G_0T^2\delta\Gamma\delta B$, appears on the right-hand side of Eq. (20). Then $\langle\delta B\rangle=0$.

Likewise variations of L with respect to δB give the random adjoint equation

$$\begin{aligned}\frac{d^2\delta B^\dagger}{dh^2}&-ik\Gamma_0G_0T^2(\delta B^\dagger+\epsilon\delta\Gamma\langle B^\dagger\rangle)\\&=ik\epsilon\Gamma_0G_0T^2(\langle\delta\Gamma\delta B^\dagger\rangle-\delta\Gamma\delta B).\end{aligned} \tag{21}$$

Then by inspection we see that $\langle B^\dagger\rangle=\langle B\rangle$ and $\delta B^\dagger=\delta B$ so that the Lagrangian approach preserves the self-adjoint nature of Eq. (6).

Note further that if we neglect the term $\langle\delta B^\dagger\delta B\delta\Gamma\rangle$ in

Eq. (18), then extremal variations of L_0 *alone* with respect to $\langle B\rangle$, $\langle B^\dagger\rangle$, and the statistical character of δB and $\delta B^\dagger$ give *all* of the first order smoothing approximation equations—Eqs. (10), (19), (20), and (21) [with the right-hand sides of Eqs. (20) and (21) set to zero].

Note also that even when the right-hand sides of Eqs. (20) and (21) are set to zero, it still follows that $\langle B\rangle = \langle B^\dagger\rangle$ and $\delta B = \delta B^\dagger$. Thus in this case the first order smoothing approximation preserves the self-adjointness of the fundamental equation (6). And it is this case where the first order smoothing normal modes of $\langle B\rangle$ are precisely those obtained under a rigorous and statistically exact treatment. We shall return to this point later for it provides the clue as to when first order smoothing theory can be expected to yield accurate answers.

2. Γ *a random function of space*

Given Eq. (12), we have a Lagrangian

$$L = \int dy \left(\omega^2 L(\Gamma_0 G_0)^{-1} B B^\dagger + (1+\epsilon\delta\Gamma) B^\dagger \frac{dB}{dy}\right). \tag{22}$$

Extremal variation of L with respect to $B^\dagger$ yields Eq. (12), while extremal variation of L with respect to B yields

$$\frac{d}{dy}[B^\dagger(1+\epsilon\delta\Gamma)] - \omega^2 L(\Gamma_0 G_0) B^\dagger = 0 \tag{23}$$

or, with

$$\begin{aligned} &B^\dagger(1+\epsilon\delta\Gamma) = Q^\dagger, \\ &(1+\epsilon\delta\Gamma)\frac{dQ^\dagger}{dy} - \frac{\omega^2 L}{\Gamma_0 G_0} Q^\dagger = 0, \end{aligned} \tag{24}$$

so that

$$Q^\dagger(i\omega) = B(\omega).$$

Thus, in this case, Eq. (12) is not self-adjoint and neither is the Lagrangian (22) for B, $B^\dagger$.

Again, if we write $B^\dagger = \langle B^\dagger\rangle + \delta B^\dagger$, $B = \langle B\rangle + \delta B$, we can write $L = L_0 + \Delta L$ with

$$\begin{aligned} L_0 = \int dy \bigg(&\omega^2 L(\Gamma_0 G_0)^{-1}(\langle B\rangle\langle B^\dagger\rangle + \langle\delta B\delta B^\dagger\rangle) \\ &+ \langle B^\dagger\rangle \frac{d\langle B\rangle}{dy} + \epsilon\langle B^\dagger\rangle \left\langle \delta\Gamma\frac{d\delta B}{dy}\right\rangle \\ &+ \epsilon\frac{d\langle B\rangle}{dy}\langle\delta\Gamma\delta B^\dagger\rangle + \delta B^\dagger\frac{d\delta B}{dy} + \epsilon\left\langle\delta B^\dagger\delta\Gamma\frac{d\delta B}{dy}\right\rangle\bigg). \end{aligned} \tag{25}$$

Then proceeding as for case 1 from extremal variations of L_0 with respect to $\langle B\rangle$ and $\langle B^\dagger\rangle$, we obtain

$$\omega^2 L(G_0\Gamma_0)^{-1}\langle B\rangle + \frac{d\langle B\rangle}{dy} + \epsilon\left\langle\delta\Gamma\frac{\partial\delta B}{\partial y}\right\rangle = 0, \tag{26a}$$

$$\omega^2 L(G_0\Gamma_0)^{-1}\langle B^\dagger\rangle - \frac{d\langle B^\dagger\rangle}{dy} - \epsilon\frac{d}{dy}\langle\delta\Gamma\delta B^\dagger\rangle = 0. \tag{26b}$$

Likewise if we vary $L_0 + \Delta L$ with respect to statistical variations of δB and $\delta B^\dagger$, we obtain the random equations

$$\frac{\omega^2 L}{\Gamma_0 G_0}\delta B + \frac{d\delta B}{dy} + \epsilon\delta\Gamma\frac{d\langle B\rangle}{dy} = \epsilon\left(\delta\Gamma\frac{d\delta B}{dy} - \left\langle\delta\Gamma\frac{d\delta B}{dy}\right\rangle\right), \tag{27a}$$

$$\begin{aligned} &\frac{\omega^2 L}{\Gamma_0 G_0}\delta B^\dagger - \frac{d\delta B^\dagger}{dy} - \epsilon\frac{d}{dy}(\delta\Gamma\langle B^\dagger\rangle) \\ &\quad = -\epsilon\frac{d}{dy}(\delta\Gamma\delta B^\dagger - \langle\delta\Gamma\delta B^\dagger\rangle). \end{aligned} \tag{27b}$$

Then by inspection of Eqs. (26)—(27) we have

$$\delta B^\dagger(i\omega)(1+\epsilon\delta\Gamma) - \epsilon\langle\delta B^\dagger(i\omega)\delta\Gamma\rangle = \delta B(\omega), \tag{28a}$$

$$\langle B^\dagger(i\omega)\rangle + \epsilon\langle\delta B^\dagger(i\omega)\delta\Gamma\rangle = \langle B(\omega)\rangle. \tag{28b}$$

Once again note that Eq. (26b) is adjoint to Eq. (26a) and Eq. (27b) is adjoint to Eq. (27a), but that Eqs. (26a) and (27a) are not self-adjoint.

Note further that if we neglect the term $\langle\delta B^\dagger\delta\Gamma\partial\delta B/\partial y\rangle$ in the averaged Lagrangian (25), then extremal variations of the averaged Lagrangian *alone* with respect to $\langle B\rangle$, $\langle B^\dagger\rangle$, and statistical variations of δB and $\delta B^\dagger$, produce Eqs. (26)—(27)—with the right-hand side of Eqs. (27a) and (27b) set to zero. And these are, of course, the equations of first order smoothing theory—which are also not self-adjoint.

It is this case where there is such a marked disagreement between the normal modes of the average field derived using the first order smoothing theory and the statistically exact normal modes of the average field.

C. Lagrangians for statistically averaged fields

It can, of course, be argued that we should not really compare a Lagrangian approach for a particular realization of $\delta\Gamma$ with Lagrangians for averaged quantities. Instead it can be argued that we should compare Lagrangians for statistically averaged fields with those obtaining under the first order smoothing approximation.

In order to answer any such argument we shall now, for the sake of completeness, give the results of such an approach. It yields the same results as subsections A and B above as is, of course, expected.

1. Γ *a random function of time*

From Lerche and Parker, Eqs. (14) and (15), we have

$$\frac{\partial R}{\partial h} = \frac{\partial}{\partial\delta\Gamma}(\delta\Gamma R) + \frac{\partial^2 R}{\partial\delta\Gamma^2} + S, \tag{29}$$

$$\frac{\partial S}{\partial h} = \frac{\partial}{\partial\delta\Gamma}(\delta\Gamma S) + \frac{\partial^2 S}{\partial\delta\Gamma^2} + ik\Gamma_0 G_0(1+\epsilon\delta\Gamma)RT^2, \tag{30}$$

with

$$\langle B\rangle = \int_{-\infty}^{\infty} d\delta\Gamma\, R(h,\delta\Gamma) \tag{31a}$$

and

$$\left\langle\frac{dB}{dh}\right\rangle = \int_{-\infty}^{\infty} d\delta\Gamma\, S(h,\delta\Gamma). \tag{31b}$$

The coefficients in Eqs. (29) and (30) are independent of time, h, so that the solutions have an exponential time dependence. Note also that Eqs. (29) and (30) are homogeneous in R and S. Thus they have a set of normal modes. These are the statistically exact modes which were compared with the first order smoothing theory modes in the paper by Lerche and Parker.[9] Here we are interested in seeing what a Lagrangian approach has to tell us about Eqs. (29) and (30) and their adjoints.

In order to derive, in a useful manner, a Lagrangian for Eqs. (29) and (30), it is advantageous to remove the first order derivatives in Eqs. (29) and (30) by writing

$$R = r\exp[-\tfrac{1}{4}(\delta\Gamma)^2], \quad S = s\exp[-\tfrac{1}{4}(\delta\Gamma)^2]$$

to obtain

$$\frac{\partial r}{\partial h} = \frac{\partial^2 r}{\partial\delta\Gamma^2} + \tfrac{1}{2}r[1 - \tfrac{1}{2}(\delta\Gamma)^2] + s, \tag{32a}$$

$$\frac{\partial s}{\partial h} = \frac{\partial^2 s}{\partial\delta\Gamma^2} + \tfrac{1}{2}s[1 - \tfrac{1}{2}(\delta\Gamma)^2] + ik\Gamma_0G_0(1+\epsilon\delta\Gamma)rT^2. \tag{32b}$$

Then with $r = r\exp(i\nu h)$, $s = s\exp(i\nu h)$ Eq. (32) give

$$i\nu r = \frac{\partial^2 r}{\partial\delta\Gamma^2} + \tfrac{1}{2}r[1 - \tfrac{1}{2}(\delta\Gamma)^2] + s, \tag{33a}$$

$$i\nu s = \frac{\partial^2 s}{\partial\delta\Gamma^2} + \tfrac{1}{2}s[1 - \tfrac{1}{2}(\delta\Gamma)^2] + ik\Gamma_0G_0(1+\epsilon\delta\Gamma)rT^2. \tag{33b}$$

Consider then the Lagrangian

$$L = \int d\delta\Gamma \biggl(s^\dagger\{\tfrac{1}{2}r[1 - 2i\nu - \tfrac{1}{2}(\delta\Gamma)^2] + s\} - \frac{\partial r}{\partial\delta\Gamma}\frac{\partial s^\dagger}{\partial\delta\Gamma} + r^\dagger\{\tfrac{1}{2}s[1 - 2i\nu - \tfrac{1}{2}(\delta\Gamma)^2] + ik\Gamma_0G_0T^2(1+\epsilon\delta\Gamma)r\} - \frac{\partial r^\dagger}{\partial\delta\Gamma}\frac{\partial s}{\partial\delta\Gamma}\biggr). \tag{34}$$

Extremal variations of L with respect to $r^\dagger$ and $s^\dagger$ give Eq. (33), while extremal variations of L with respect to r and s give, respectively, the adjoint equations

$$\frac{\partial^2 s^\dagger}{\partial\delta\Gamma^2} + \tfrac{1}{2}s^\dagger(1 - 2i\nu - \tfrac{1}{2}(\delta\Gamma)^2) + ik\Gamma_0G_0T^2(1+\epsilon\delta\Gamma)r^\dagger = 0, \tag{35}$$

$$\frac{\partial^2 r^\dagger}{\partial\delta\Gamma^2} + \tfrac{1}{2}r^\dagger(1 - 2i\nu - \tfrac{1}{2}(\delta\Gamma)^2) + s^\dagger = 0. \tag{36}$$

Thus by inspection of Eqs. (33), (35), and (36) we see that

$$r^\dagger = r, \quad s^\dagger = s. \tag{37}$$

Hence the pair of equations for r, s is a self-adjoint pair.

2. Γ *a random function of space*

From Lerche and Parker,[9] Eq. (29), we have

$$\frac{\partial U}{\partial y} = \frac{\partial}{\partial\delta\Gamma}(\delta\Gamma U) + \frac{\partial^2 U}{\partial\delta\Gamma^2} - \omega^2LU[\Gamma_0G_0(1+\epsilon\delta\Gamma)]^{-1}, \tag{38}$$

with

$$\langle B\rangle = \int_{-\infty}^{\infty} d\delta\Gamma\, U(y,\delta\Gamma). \tag{39}$$

Write

$$U = T\exp(iky)\exp[-\tfrac{1}{4}(\delta\Gamma)^2], \tag{40}$$

when Eq. (38) gives

$$ikT = \frac{\partial^2 T}{\partial\delta\Gamma^2} + \tfrac{1}{2}T\{1 - \tfrac{1}{2}(\delta\Gamma)^2 - \omega^2L[\Gamma_0G_0(1+\epsilon\delta\Gamma)]^{-1}\}. \tag{41}$$

Consider the Lagrangian

$$L = \int d\delta\Gamma\biggl(-\frac{\partial T}{\partial\delta\Gamma}\frac{\partial}{\partial\delta\Gamma}[T^\dagger(1+\epsilon\delta\Gamma)] + \tfrac{1}{2}TT^\dagger\{[1 - 2ik - \tfrac{1}{2}(\delta\Gamma)^2](1+\epsilon\delta\Gamma) - \omega^2L(\Gamma_0G_0)^{-1}\}\biggr). \tag{42}$$

Extremal variations of L with respect to $T^\dagger$ yield Eq. (41), while extremal variations of L with respect to T yield the adjoint equation

$$\frac{\partial^2}{\partial\delta\Gamma^2}[T^\dagger(1+\epsilon\delta\Gamma)] + \tfrac{1}{2}T^\dagger\{[1 - 2ik - \tfrac{1}{2}(\delta\Gamma)^2](1+\epsilon\delta\Gamma) - \omega^2L(\Gamma_0G_0)^{-1}\} = 0 \tag{43}$$

so that, by inspection of Eqs. (41) and (43), we have

$$T = T^\dagger(1+\epsilon\delta\Gamma). \tag{44}$$

Thus equation (41) is not self-adjoint *except* with the inclusion of a weighting factor $(1+\epsilon\delta\Gamma)^{-1}$ in the adjoint equation. But since the range of integration of $\delta\Gamma$ is $0 \leqslant |\delta\Gamma| \leqslant \infty$, the weighting factor either is singular (if included in the adjoint equation) or is not positive definite [if included in equation (41)], *except* for the singular limiting case of $\epsilon = 0$ when the weighting factor is unity. And then $T(\epsilon = 0) = T^\dagger\delta$ and $(\epsilon = 0)$ so that T is self-adjoint. It is precisely the case of $\epsilon = 0$ where agreement results between the first order smoothing theory dispersion relation and the statistically exact dispersion relation. For $\epsilon \neq 0$, no matter how small, strong disagreement obtains.

So the results of these particular investigations suggest two things:

First, when the full equation is self-adjoint, *and when the first order smoothing approximation preserves the self-adjoint character*, then the first order smoothing approximation gives the same dispersion relation as the statistically exact treatment.

Second, when the full equation is not self-adjoint, *and when the first order smoothing approximation either turns the non-self-adjoint equation into one which is self-adjoint or replaces the "true" non-self-adjoint equation by another equation which is also not self-adjoint*, then the dispersion relation obtained using the first order smoothing approximation differs markedly from that obtaining under a statistically exact treatment.

III. A GENERAL CRITERION FOR VALIDITY OF FIRST ORDER SMOOTHING THEORY

Suppose that, in general, we had an equation of the form

$$\mathcal{L}y = \delta\mathcal{L}y, \tag{45}$$

where $\mathcal{L}$ is an ordered operator and $\delta\mathcal{L}$ is a random operator. Then with $y = \langle y\rangle + \delta y$ we have the statistically exact pair of equations

$$\mathcal{L}\langle y\rangle = \langle\delta\mathcal{L}\,\delta y\rangle \tag{46}$$

and

$$\mathcal{L}\,\delta y - \delta\mathcal{L}\langle y\rangle = \delta\mathcal{L}\,\delta y - \langle\delta\mathcal{L}\,\delta y\rangle. \tag{47}$$

We can construct the Lagrangian

$$\mathcal{L} = \int d\Omega\, y^\dagger[\mathcal{L}y - \delta\mathcal{L}y] \tag{48}$$

over the space, Ω, appropriate to the operator field $\mathcal{L} + \delta\mathcal{L}$. And then extremal variations of $\mathcal{L}$ with respect to $y^\dagger$ yield equation (45) while extremal variations of $\mathcal{L}$ with respect to y yield the exact adjoint equation

$$\mathcal{L}^\dagger y^\dagger = \delta\mathcal{L}^\dagger y^\dagger. \tag{49}$$

With $y^\dagger = \langle y^\dagger\rangle + \delta y^\dagger$, Eq. (49) yields the statistically exact pair of equations

$$\underline{L}^\dagger\langle y^\dagger\rangle = \langle\delta\underline{L}^\dagger\delta y^\dagger\rangle, \tag{50a}$$

$$\underline{L}^\dagger\delta y^\dagger - \delta\underline{L}^\dagger\langle y^\dagger\rangle = \delta\underline{L}^\dagger\delta y^\dagger - \langle\delta\underline{L}^\dagger\delta y^\dagger\rangle. \tag{50b}$$

Suppose then that we write the Lagrangian L as $L_0 + \Delta L$ with

$$\begin{aligned} L_0 = \int d\Omega(\langle y^\dagger\rangle\underline{L}\,\langle y\rangle + \langle\delta y^\dagger\underline{L}\,\delta y\rangle \\ - \langle y^\dagger\rangle\langle\delta\underline{L}\,\delta y\rangle - \langle\delta y^\dagger\delta\underline{L}\,\rangle\langle y\rangle - \langle\delta y^\dagger\delta\underline{L}\,\delta y\rangle). \end{aligned} \tag{51}$$

Proceeding as in Sec. II we see that variations of L_0 with respect to $\langle y\rangle$ and $\langle y^\dagger\rangle$ give Eqs. (46) and (50a), while variations of $L_0 + \Delta L$ with respect to the statistical character of δy and $\delta y^\dagger$ give Eqs. (47) and (50b).

If we neglect the term $\langle\delta y^\dagger\delta\underline{L}\,\delta y\rangle$ in Eq. (51) (the first order smoothing approximation), then variations of L_0 *alone* with respect to $\langle y\rangle$, $\langle y^\dagger\rangle$, and the statistical character of δy and $\delta y^\dagger$ yield Eqs. (46), (47), and (50) with the right-hand sides of Eqs. (47) and (50b) set to zero.

Now if the equation for y is self-adjoint and if neglecting $\langle\delta y^\dagger\delta\underline{L}\,\delta y\rangle$ preserves the self-adjoint character of the equation, we can argue as follows. It is always possible to choose a set of trial functions for $\langle y\rangle$, $\langle y^\dagger\rangle$, etc., such that

$$\int d\Omega\langle\delta y^\dagger\delta\underline{L}\,\delta y\rangle = 0. \tag{52}$$

Then since the equations stay self-adjoint under such a choice we have the equivalent of a Rayleigh–Ritz principle which guarantees that a set of trial functions will come closer and closer (in a monotone manner) to yielding the "true" eigenvalues (See Lerche and Parker, Sec. II for a specific case).

If, however, neglecting $\langle\delta y^\dagger\delta\underline{L}\,\delta y\rangle$ gives rise to a non-self-adjoint approximation to the original self-adjoint equation no such Rayleigh–Ritz type of statement is available. The approximate eigenvalues found under such conditions may be far removed from the "true" eigenvalues, and increasing the number of trial functions can make the difference larger rather than smaller.

Likewise if the true equation is not self-adjoint and the first order smoothing approximation replaces it by a self-adjoint equation, then while a Rayleigh–Ritz type of statement *is* available for the approximate equation, no such type of statement is available for the true equation. Accordingly, once again, any approximate eigenvalues can be far removed from the true eigenvalues.

Finally if the true equation is not self-adjoint, and if the first order smoothing approximation replaces it by an equation which is also not self-adjoint, no useful statement is available on the accuracy of the approximate eigenvalue—except to say that it is probably wrong (see Lerche and Parker,[9] Sec. III for a specific case).

In summary, then, we give the following criterion:

Theorem: (i) If the true equation is self-adjoint and if the first order smoothing approximate equation is also self-adjoint, then almost surely, almost everywhere the first order smoothing dispersion relation will be the same as the correct dispersion relation.

(ii) If both (or either) the true equation and the first order smoothing approximate equation are not self-adjoint, then almost surely, almost everywhere the first order smoothing dispersion relation will differ substantively from the correct dispersion relation.[14]

To put the point another way: An approximation which changes the topological character of an equation must be in error somewhere.

ACKNOWLEDGMENTS

The work reported here was supported by the United States Air Force, Air Force Systems Command under Contract F-19628-72-C-0056.

[1] A. M. Yaglom, *An Introduction to the Theory of Stationary Random Functions* (Prentice-Hall, Englewood Cliffs, N.J., 1962).

[2] W.C. Meecham, J. Geophys. Res. 69, 3175 (1964).

[3] U. Frisch, "Wave Propagation in Random Media," in *Probabilistic Methods in Applied Mathematics, Vol.* 1, edited by A.T. Bharucha-Reid (Academic, New York, 1968).

[4] I.B. Bernstein and F. Engelmann, Phys. Fluids 9, 937 (1966).

[5] A. Dolinsky and R. Goldman, Phys. Fluids 10, 1251 (1967).

[6] K.Y. Fu, Plasma Phys. 15, 57 (1973).

[7] D.A. Tidman and N.A. Krall, *Shock Waves in Collisionless Plasmas* (Wiley-Interscience, New York, 1971).

[8] C.F. Kennel and H.E. Petschek, J. Geophys. Res. 71, 1 (1966).

[9] I. Lerche and E.N. Parker, J. Math. Phys. 14, 1949 (1973).

[10] R.H. Kraichnan, J. Math. Phys. 2, 124 (1961).

[11] J.R. Herring, Phys. Fluids 12, 39 (1969).

[12] E.N. Parker, Astrophys. J. 122, 293 (1955).

[13] E.N. Parker, Astrophys. J. 162, 665 (1970).

[14] That there may exist particular non-self-adjoint cases where the first order smoothing approximation gives correctly the true dispersion relation is not germane to the above criterion, for such situations are themselves special rather than general.

Interaction and transformation of electromagnetic and gravitational waves in Einstein-Maxwell field

Tatsuo Tokuoka

Department of Aeronautical Engineering, Kyoto University, Kyoto, Japan

(Received 11 February 1974; final revised manuscript received 7 June 1974)

The interaction and the transformation of the electromagnetic and gravitational waves and the variation of those amplitudes are analyzed theoretically, where the Einstein-Maxwell field is assumed. Two waves are defined such that the electromagnetic vector potentials and the gravitational tensor potentials and their first derivatives are continuous while their second derivatives are discontinuous at a wave front. It is proved that the concomitant electromagnetic and gravitational waves can interact with each other if an external electromagnetic field exists. The global behaviors of interaction and transformation of the two concomitant waves are also investigated, where a Lorentz metric space-time is assumed. The variation formulas for amplitude are derived. The amplitudes may grow or decay according to a factor depending on the mean and Gaussian curvatures of an initial wave front and to a factor depending on the external electromagnetic field. When there is no external electromagnetic field, there is no interaction between the two waves and they propagate independently, while when an external field exists, the electromagnetic wave may transform into the gravitational wave and vice versa. In a case of a weak constant external electromagnetic field, where a flat space may be assumed, the amplitude of a wave varies sinusoidally with respect to the distance measured along its ray. The length of interchangeability is defined by the length in which a wave is completely transformed into its dual wave, and it is inversely proportional to the strength of an external electromagnetic field.

1. INTRODUCTION

In the classical theories of continua we have, in general, three kinds of waves or discontinuity propagations, that is, *harmonic oscillations*, *characteristics*, and *singular surfaces*. The theory of harmonic oscillation is restricted to the case of linear oscillation or an infinitesimal disturbance, and the theory of characteristics cannot, in general, be a consequence of the general principles of physics. On the other hand, Christoffel, Hugoniot, Hadamard, and Duhem viewed waves in continuous media as propagating singular surfaces. Since then a great deal of thought has been devoted to the analysis of the singular surface for the waves in the many kinds of continua. In many cases, the propagation speeds derived from the above three methods turn out to be exactly the same values. No explanation of this remarkable agreement is known. For a general reference on the theory of singular surface, refer to Truesdell and Toupin.[1]

When Maxwell[2] established a dynamical theory of electromagnetic field, he showed theoretically the existence of the *electromagnetic wave*. Shortly after framing the general relativity, Einstein[3] showed the propagation of an infinitesimal disturbance of gravitational potentials in a Lorentz metric space—time, and he called it the *gravitational wave*. For general references on the gravitational waves refer to Weber[4] and Zakharov.[5]

Lichnerowicz[6] surveyed the electromagnetic and gravitational waves as the characteristic manifolds. The singular surfaces were also regarded as those wavefronts. An electromagnetic wave was defined by Trautman[7] as a surface across which the fields are continuous, but their first derivatives may suffer jump discontinuities, and a gravitational wave was defined by Trautman[7] and Thomas[8] as a surface across which the gravitational potentials and their first derivatives are continuous, but their second derivatives may have jumps. Starting from the somewhat different definition from theirs Tokuoka[9] proved the reality of the *transverse gravitational wave*.

Although Maxwell's field equations hold, in principle, within the framework of the special relativity, there is a theory, called *Einstein—Maxwell field theory*, in which the Maxwell field is assumed to be concomitant to the Einstein—Riemann space—time. While the existence and the propagation of the electromagnetic and gravitational waves were investigated fairly, there are a few recent peculiar surveys with respect to the interaction between those two waves. Johnston *et al.*[10,11] studied the feasibility of the conversion of gravitational radiation into electromagnetic radiation when an infalling neutral object perturbs the background field of a black hole. Basing this paper on Einstein—Maxwell field equations and the theory of singular surface, we shall investigate the interaction and the transformation of the electromagnetic and gravitational waves.

2. DEFINITIONS OF WAVES AND COMPATIBILITY CONDITIONS

For the Einstein—Maxwell field in free Einstein—Riemann space—time $\mathcal{E}$ the gravitational potentials $g_{\alpha\beta}$ and the Minkowski electromagnetic antisymmetric tensor $F_{\alpha\beta}$ are governed by the equations in Gaussian system of units

$$F_{\alpha\beta,\gamma}+F_{\beta\gamma,\alpha}+F_{\gamma\alpha,\beta}=0, \tag{2.1}$$

$$F_{\alpha}{}^{\rho}{}_{;\rho}=0, \tag{2.2}$$

$$R_{\alpha\beta}=-(2\kappa/c^4)(F_{\alpha\rho}F^{\rho}{}_{\beta}+\tfrac{1}{4}g_{\alpha\beta}F_{\rho\sigma}F^{\rho\sigma}), \tag{2.3}$$

where lower case Greek indices run from 0 to 3 and a comma and a semicolon followed an index refer to, respectively, partial and covariant derivatives. Here the Minkowski tensor in a free space is related by the electric field **E** and magnetic field **H** as $F_{0i}=-cE_i$, $F_{12}=H_3$, $F_{23}=H_1$, $F_{31}=H_2$ in a local geodesic coordinate system with the metric differential form

$$ds^2=c^2dt^2-\delta_{ij}dx^i dx^j, \tag{2.4}$$

which can be chosen generally in a point considered in

$\mathcal{E}$, where and henceforth lower case Latin indices run from 1 to 3.

Thomas[12] proved that if the $g_{\alpha\beta}$ are of class C^1, a coordinate transformation of class C^1 may reduce to

$$ds^2 = V^2(x^\alpha)\,dt^2 - a_{ij}\,dx^i\,dx^j, \tag{2.5}$$

where $d\sigma^2 = a_{ij}\,dx^i dx^j$ is the positive-definite differential form of the three-dimensional space $\mathcal{R}_3$. Equations (2.1) assure that it is exact and admit a four-dimensional vector potential ϕ_α, such that

$$F_{\alpha\beta} = \phi_{\alpha,\beta} - \phi_{\beta,\alpha}. \tag{2.6}$$

A three-dimensional regular hypersurface Σ in $\mathcal{E}$ can be interpreted as a representation of a wave. The form (2.5) permits us to view Σ as the successive positions of a two-dimensional wavefront $\mathcal{S}(t)$ in $\mathcal{R}_3$. Now we define an *electromagnetic wave*.[7]

Definition 1: When the vector potentials ϕ_α satisfy the following two conditions, we say an electromagnetic wave exists:

(i) ϕ_α and their first-order coordinate derivatives are continuous over a surface $\mathcal{S}$.

(ii) at least one component of ϕ_α is discontinuous at $\mathcal{S}$.

Also we define a *gravitational wave*.[9]

Definition 2: When the gravitational potentials $g_{\alpha\beta}$ satisfy the following two conditions, we say a *gravitational wave* exists:

(i) $g_{\alpha\beta}$ and their first-order coordinate derivatives are continuous over a surface $\mathcal{S}$.

(ii) at least one component of the Riemann curvature tensor is discontinuous at $\mathcal{S}$.

The wavefront $\mathcal{S}(t)$ in $\mathcal{R}_3$ can be defined parametrically by a set of equations

$$x^i = \psi^i(u^\Gamma; t), \tag{2.7}$$

where ψ^i are assumed to be continuously differential functions of the parameters u^Γ such that the functional matrix

$$\left\| \frac{\partial \psi^i}{\partial u^\Gamma} \right\| \tag{2.8}$$

has rank 2 for all values of the surface under consideration. Here capital Greek indices take 1 and 2. Then we can eliminate u^Γ from (2.7), and we have the representation of Σ in the form

$$\psi(x^i, t) = 0. \tag{2.9}$$

Let us denote by G the velocity of propagation of the surface $\mathcal{S}$ in the direction of its unit normal ν^i. The partial derivatives $x^i_\Gamma \equiv \partial x^i/\partial u^\Gamma$ denote a tangent vector to $\mathcal{S}$. The fundamental metric tensor of the surface $\mathcal{S}$ is given by

$$a_{\Gamma\Delta} = a_{ij}\, x^i_\Gamma\, x^j_\Delta. \tag{2.10}$$

The theory of surface shows that

$$\nu_{i;\Gamma} = -a^{\Delta\Theta} b_{\Gamma\Delta}\, x_{i\Theta}, \tag{2.11}$$

where $b_{\Gamma\Delta}$ are the second fundamental form of $\mathcal{S}$ and we denote that $x_{i\Theta} \equiv a_{ij}\, x^j_\Theta$. The δ time derivative of a quantity denotes time rate of the quantity observed on a wavefront.[13] Then we have obviously

$$\frac{\partial x^i}{\partial t} = G\nu^i, \quad G = \frac{d\sigma}{dt}, \tag{2.12}$$

where σ is the arc length along a normal trajectory to the surface $\mathcal{S}$. According to Thomas[8] we depict here the compatibility conditions. The compatibility conditions of the second order in the metric (2.5) are given by

$$[f_{,ij}] = \bar{f}\nu_i\nu_j, \quad [f_{,i0}] = -G\bar{f}\nu_i, \quad [f_{,00}] = G^2\bar{f}, \tag{2.13}$$

where

$$\bar{f} \equiv [f_{,km}]\nu^k\nu^m \tag{2.14}$$

denotes the magnitude of the jump discontinuity of f, which may take any component of ϕ_α or $g_{\alpha\beta}$ and where index 0 denotes the partial derivative with respect to t. The compatibility conditions of the third order have the complicated forms. Here we depict them in a local geodesic system with the metric (2.4). We have

$$\begin{aligned}[f_{,ijk}] &= \bar{\bar{f}}\nu_i\nu_j\nu_k + \bar{f}_{,\Gamma}a^{\Gamma\Delta}(\nu_i\nu_j x_{k\Delta} + \nu_j\nu_k x_{i\Delta} + \nu_k\nu_i x_{j\Delta}) \\ &\quad - \bar{f}a^{\Gamma\Delta}a^{\Theta\Lambda}b_{\Gamma\Theta}(\nu_i x_{j\Delta}x_{k\Lambda} + \nu_j x_{k\Delta}x_{i\Lambda} + \nu_k x_{i\Delta}x_{j\Lambda}),\end{aligned} \tag{2.15}$$

$$\begin{aligned}[f_{,ij0}] &= -G\bar{\bar{f}}\nu_i\nu_j - G\bar{f}_{,\Gamma}a^{\Gamma\Delta}(\nu_i x_{j\Delta} + \nu_j x_{i\Delta}) \\ &\quad + G\bar{f}a^{\Gamma\Delta}g^{\Theta\Lambda}b_{\Gamma\Theta}x_{i\Delta}x_{j\Lambda} + \frac{\delta\bar{f}}{\delta t}\nu_i\nu_j,\end{aligned} \tag{2.16}$$

$$[f_{,i00}] = G^2\bar{\bar{f}}\nu_i + G^2\bar{f}_{,\Gamma}\,a^{\Gamma\Delta}x_{i\Delta} - 2G\frac{\delta\bar{f}}{\delta t}\nu_i, \tag{2.17}$$

$$[f_{,000}] = -G^3\bar{\bar{f}} + 3G^2\frac{\delta\bar{f}}{\delta t}, \tag{2.18}$$

where we define that

$$\bar{\bar{f}} \equiv [f_{,kmn}]\nu^k\nu^m\nu^n. \tag{2.19}$$

3. EXISTENCE AND PROPAGATION OF WAVES

Some particular analyses of this item viewed from the theory of singular surface were reported by Trautman[7] and Tokuoka.[9] Here the brief summary of their results is depicted for the sake of our analysis.

Taking the differences of the field equations (2.2) and (2.3) on two contiguous sides of the wave surface $\mathcal{S}$, referring to Definitions 1 and 2, we have

$$[F_{\alpha\rho,\sigma}]g^{\rho\sigma} = 0, \tag{3.1}$$

$$[R_{\alpha\beta}] = 0, \tag{3.2}$$

which give the necessary propagation conditions of the electromagnetic and gravitational waves, respectively.

Applying the compatibility conditions of second order (2.13) to (3.1) and (3.2), we have the *characteristic condition*

$$g^{\rho\sigma}\,\xi_\rho\,\xi_\sigma = 0, \tag{3.3}$$

which gives the propagation velocity

$$G = V, \tag{3.4}$$

where

$$\xi_0 = G, \quad \xi_i = -\nu_i, \quad \xi^0 = G/V^2, \quad \xi^i = \nu^i. \tag{3.5}$$

Equation (3.3) is a null surface in $\mathcal{E}$, and the hypersurface Σ of (2.9) can be expressed as

$$g^{\rho\sigma}\psi_{,\rho}\psi_{,\sigma}=0. \tag{3.6}$$

Here we propose

Definition 3: The *rays* of electromagnetic and gravitational waves are defined by the characteristics of the wave surface (3.6).

We can easily say that the ray of a wave is null geodesic. Therefore, we can assert that electromagnetic and gravitational waves, which have an identical wavefront at an instance, can propagate concomitantly with a common velocity along a common ray.

Without loss of generality the coordinate system in $\mathcal{R}_3$ may be chosen so that the x^1 axis is parallel to the normal of the wavefront $\mathcal{S}$ and the x^K axes lie upon it. Thus the differential form (2.5) reduces to

$$ds^2=V^2dt^2-(dx^1)^2-a_{KM}dx^K dx^M, \tag{3.7}$$

where capital Latin indices take 2 and 3.

A. Electromagnetic wave

A transverse electromagnetic wave having the amplitude $\overline{\phi}_K$ is a real wave and propagate with the velocity (3.4), while we can say that

$$\overline{\phi}_0=\overline{\phi}_1=0, \tag{3.8}$$

which can be obtained by the gauge transformation.

Relations $\overline{F}_{\alpha\beta}=\overline{\phi}_\alpha\xi_\beta-\overline{\phi}_\beta\xi_\alpha$ and (3.8) show that

$$\overline{E}_1=\overline{H}_1=0, \tag{3.9}$$

$$\overline{E}_2=\overline{H}_3=\overline{\phi}_2,\quad \overline{E}_3=-\overline{H}_2=\overline{\phi}_3. \tag{3.10}$$

B. Gravitational wave

A transverse gravitational wave having the amplitudes $\overline{g}_{KM}$ is a real wave and propagates with the velocity (3.4), while a longitudinal wave having amplitudes $\overline{g}_{00}$, $\overline{g}_{01}$, and $\overline{g}_{11}$ and shear waves having amplitudes $\overline{g}_{0K}$ and $\overline{g}_{1K}$ are imaginary waves. Furthermore, there is a condition on $\overline{g}_{KM}$, that is,

$$a^{PQ}\overline{g}_{PQ}=0. \tag{3.11}$$

The imaginary waves can be canceled out by an appropriate transformation, and in any case we can put

$$\overline{g}_{00}=\overline{g}_{01}=\overline{g}_{11}=\overline{g}_{0K}=\overline{g}_{1K}=0. \tag{3.12}$$

4. LOCAL PROPERTIES OF INTERACTION AND VARIATION

Differentiating the field equations (2.2) and (2.3) with respect to x^γ and taking the differences of them on two contiguous sides of the wave surface $\mathcal{S}$, we have the differential equations for the variation of wave amplitudes of the electromagnetic and gravitational waves. In order to simplify our analysis, let us take a local geodesic system, which has the differential form (2.4) at a point in $\mathcal{E}$ under consideration. In this system all of the derivatives of first order of the gravitational potentials vanish, and we can obtain that

$$g^{\rho\sigma}\{[F_{\alpha\rho,\sigma\gamma}]-g^{\mu\nu}(F_{\mu\rho}[[\alpha\sigma,\nu]_{,\gamma}]+F_{\alpha\mu}[[\rho\sigma,\nu]_{,\gamma}])\}=0, \tag{4.1}$$

$$\begin{aligned}\tfrac{1}{2}g^{\rho\sigma}([g_{\alpha\beta,\rho\sigma\gamma}]-[g_{\alpha\rho,\sigma\beta\gamma}]-[g_{\beta\rho,\sigma\alpha\gamma}]+[g_{\rho\sigma,\alpha\beta\gamma}])\\ =-(2\kappa/c^4)g^{\rho\sigma}(F_{\alpha\rho}[F_{\sigma\beta,\gamma}]+F_{\beta\rho}[F_{\sigma\alpha,\gamma}]\\ +\tfrac{1}{2}g_{\alpha\beta}g^{\mu\nu}F_{\rho\mu}[F_{\sigma\nu,\gamma}]).\end{aligned} \tag{4.2}$$

Although full range of values of α, β, and γ may be assumed in (4.1) and (4.2), it will suffice for our purpose, however, to limit our attention to them for which $\alpha=K$, $\beta=M$, and $\gamma=0$. Taking account of this and the conditions (3.8), (3.11), and (3.12) and applying the compatibility conditions of the third order (2.16), (2.17), and (2.18), we have that

$$\begin{aligned}\frac{\delta\overline{\phi}_K}{\delta t}-\tfrac{1}{2}c\,\overline{\phi}_K a^{\Gamma\Delta}a^{\Theta\Lambda}b_{\Gamma\Theta}\delta^{ij}x_{i\Delta}x_{j\Lambda}\\ -\tfrac{1}{2}c(\overline{\phi}_{i,\Gamma}\nu_j-\overline{\phi}_i a^{\Theta\Lambda}b_{\Gamma\Theta}x_{j\Lambda})a^{\Gamma\Delta}x_{K\Delta}\delta^{ij}\\ =\tfrac{1}{2}\delta^{PQ}\overline{g}_{KP}(F_{0Q}+cF_{1Q}),\end{aligned} \tag{4.3}$$

and

$$\begin{aligned}\frac{\delta\overline{g}_{KM}}{\delta t}-c\,\overline{g}_{KM}a^{\Gamma\Delta}a^{\Theta\Lambda}b_{\Gamma\Theta}\delta^{ij}x_{i\Delta}x_{j\Lambda}\\ -c(\overline{g}_{Ki,\Gamma}x_{M\Delta}-\overline{g}_{Ki}a^{\Theta\Lambda}b_{\Gamma\Theta}x_{M\Lambda})a^{\Gamma\Delta}\delta^{ij}\nu_j\\ -c(\overline{g}_{Mi,\Gamma}x_{K\Delta}-\overline{g}_{Mi}a^{\Theta\Lambda}b_{\Gamma\Theta}x_{K\Lambda})a^{\Gamma\Delta}\delta_{ij}\nu_j\\ =-\frac{2\kappa}{c^4}\{\overline{\phi}_K(F_{0M}+cF_{1M})+\overline{\phi}_M(F_{0K}+cF_{1K})\},\end{aligned} \tag{4.4}$$

where we used $\nu_1=1$ and $\nu_2=\nu_3=0$. From the relations (2.10), (2.11), (3.8), and (3.12), we can show that the third term in the left side of (4.3) and the third and fourth terms in the left side of (4.4) vanish identically. Then referring to

$$\frac{\delta\overline{\phi}_K}{\delta t}=c\frac{d\overline{\phi}_K}{d\sigma},\quad \frac{\delta\overline{g}_{KM}}{\delta t}=c\frac{d\overline{g}_{KM}}{d\sigma}, \tag{4.5}$$

where σ is the arc length along a normal trajectory to $\mathcal{S}$, we have the set of differential equations for the amplitudes

$$\frac{d\overline{\phi}_K}{d\sigma}=\Omega\overline{\phi}_K-\tfrac{1}{2}\delta^{PQ}\overline{g}_{KP}K_Q, \tag{4.6}$$

$$\frac{d\overline{g}_{KM}}{d\sigma}=\Omega\overline{g}_{KM}+\frac{2\kappa}{c^4}(\overline{\phi}_K K_M+\overline{\phi}_M K_K-\delta_{KM}\delta^{PQ}\overline{\phi}_P K_Q),$$

where (4.7)

$$K_K\equiv(1/c)F_{K0}+F_{K1} \tag{4.8}$$

or

$$K_2\equiv E_2-H_3,\quad K_3\equiv E_3+H_2, \tag{4.9}$$

and

$$\Omega\equiv\tfrac{1}{2}a^{\Gamma\Delta}b_{\Gamma\Delta} \tag{4.10}$$

is the mean curvature of the wavefront $\mathcal{S}$. The equations of variation (4.6) and (4.7) show that *electromagnetic and gravitational waves propagate independently with each other if and only if the external electromagnetic fields are absent or those fields are longitudinal along the rays of waves, while, if the transversal components of the external fields are present and* $K_K\neq0$, *the interaction between two waves must occur.*

5. CONDITION OF SPACE-TIME

Now we will investigate the condition on the differential metric form (3.7), where the x^1-axis and $x^1 = \text{constant}$ are defined, respectively, as a ray of a wave and a wave front. The equations of a geodesic can be given, in general, by the Euler–Lagrange equations of the variational principle,[14]

$$\delta \int \left\{ V^2 \left(\frac{dt}{dq}\right)^2 - 1 - a_{PQ} \frac{dx^P}{dq} \frac{dx^Q}{dq} \right\} dq = 0, \tag{5.1}$$

where q is a parameter defined along a geodesic. We have then

$$\frac{d^2t}{dq^2} + V^{-1} \frac{\partial V}{\partial t} \left(\frac{dt}{dq}\right)^2 + 2V^{-1} \frac{\partial V}{\partial x^1} \frac{dt}{dq} \frac{dx^1}{dq} + 2V^{-1} \frac{\partial V}{\partial x^P} \frac{dt}{dq} \frac{dx^P}{dq} + \frac{\partial a_{PQ}}{\partial t} \frac{dx^P}{dq} \frac{dx^Q}{dq} \tag{5.2}$$

$$\frac{d^2x^1}{dq^2} - V \frac{\partial V}{\partial x^1} \left(\frac{dt}{dq}\right)^2 + \frac{1}{2} \frac{\partial a_{PQ}}{\partial x^1} \frac{dx^P}{dq} \frac{dx^Q}{dq} = 0, \tag{5.3}$$

$$\frac{d^2x^K}{dq^2} + Va^{KP} \frac{\partial V}{\partial x^P} \left(\frac{dt}{dq}\right)^2 + a^{KP} \frac{\partial a_{PQ}}{\partial t} \frac{dt}{dq} \frac{dx^Q}{dq} + a^{KP} \frac{\partial a_{PQ}}{\partial x^1} \frac{dx^1}{dq} \frac{dx^Q}{dq} + \left\{ {K \atop P\ Q} \right\} \frac{dx^P}{dq} \frac{dx^Q}{dq} = 0, \tag{5.4}$$

where $\{{}_P{}^K{}_Q\}$ denotes a Christoffel symbol of the second kind composed from the metric tensor a_{KM}. The condition that the x^1 axis is a null geodesic is given by

$$V\frac{dt}{dq} - \frac{dx^1}{dq} = 0, \quad \frac{dx^K}{dq} = 0. \tag{5.5}$$

Equation (5.4) and the second equation of (5.5) yield that

$$\frac{\partial V}{\partial x^K} = 0. \tag{5.6}$$

Here we assumed that $dt/dq \neq 0$. Differentiating the first equation of (5.5) with respect to q and eliminating d^2t/dq^2 and d^2x^1/dq^2 from (5.2) and (5.3), we have

$$\frac{\partial V}{\partial x^1} = 0. \tag{5.7}$$

On the other hand, if we assume $\partial V/\partial x^1 = \partial V/\partial x^K = 0$, the geodesic equations (5.2), (5.3), and (5.4) reduce, respectively, to

$$\frac{d^2t}{dq^2} + V^{-1} \frac{dv}{dt} \left(\frac{dt}{dq}\right)^2 = 0, \tag{5.8}$$

$$\frac{d^2x^1}{dq^2} = 0, \tag{5.9}$$

$$\frac{d^2x^K}{dq^2} = 0. \tag{5.10}$$

Integrating them, we have

$$V\frac{dt}{dq} = \alpha, \quad \frac{dx^1}{dq} = \beta, \quad \frac{dx^K}{dq} = \gamma, \tag{5.11}$$

where α, β, and γ are constants. Putting $\alpha = \beta \neq 0$, $\gamma = 0$, we can obtain the null geodesic of x^1 axis, (5.5).

Therefore, we can say that the x^1 axis in the metric (3.7) is a null geodesic if V is independent of x^1 and x^K. Changing the time scale from t to

$$\tau = (1/c) \int_0^t V(t)\, dt, \tag{5.12}$$

we have the differential metric

$$ds^2 = c^2 d\tau^2 - (dx^1)^2 - a_{KM} dx^K dx^M. \tag{5.13}$$

From the metric form we can easily see that the wavefront $\mathcal{S}(\tau)$ in an instance τ is parallel in the sense that the distance σ on each point of $\mathcal{S}(\tau)$ measured along the x^1 axis from an initial wavefront $\mathcal{S}(\tau_0)$ at $\tau = \tau_0$ has a constant value.[12] For a case that $\mathcal{R}_3$ is a three-dimensional Euclidean metric space, we have a formula for the mean curvature, i.e.,

$$\Omega = \frac{\Omega_0 - \sigma K_0}{1 - 2\sigma\Omega_0 + \sigma^2 K_0}, \tag{5.14}$$

where Ω_0 and K_0 are, respectively, the mean curvature and the Gaussian curvature of the initial wavefront $\mathcal{S}(\tau_0)$.[15]

6. GLOBAL PROPERTIES OF INTERACTION AND TRANSFORMATION

In this section we assume that a background external electromagnetic field is so weak that the space–time in front of the wave is flat approximately, and a Lorentz metric form (2.4) can hold there. Then the differential equations of the variation of amplitudes (4.6) and (4.7) and the formula (5.14) may hold globally. The differential equations can be reduced to

$$\frac{d\bar{\mathbf{h}}^*}{d\sigma} = \mathbf{M}\bar{\mathbf{h}}^*, \tag{6.1}$$

where we assumed that

$$\bar{\mathbf{h}}^* \equiv \bar{\mathbf{h}}(1 - 2\sigma\Omega_0 + \sigma^2 K_0)^{1/2}, \tag{6.2}$$

$$\bar{h}_1 \equiv 2C\bar{\phi}_2, \quad \bar{h}_2 \equiv 2C\bar{\phi}_3$$

$$\bar{h}_3 \equiv \bar{g}_{22} = -\bar{g}_{33}, \quad \bar{h}_4 \equiv \bar{g}_{23}, \tag{6.3}$$

$$\mathbf{M} \equiv \begin{Vmatrix} 0 & -\mathbf{N}^T \\ \mathbf{N} & 0 \end{Vmatrix}, \quad \mathbf{N} \equiv C \begin{Vmatrix} K_2 & -K_3 \\ K_3 & k_2 \end{Vmatrix}, \tag{6.4}$$

$$C \equiv \kappa^{1/2}/c^2, \tag{6.5}$$

and $\mathbf{N}^T$ denotes the transpose of $\mathbf{N}$. Here $C = 2.874 \times 10^{-25}$ $\text{cm}^{-1/2} g^{-1/2}$ sec is called the *transformation constant.*

Direct integration of (6.1) yields

$$\bar{\mathbf{h}}^* = \bar{\mathbf{h}}_0 \exp(\tilde{\mathbf{M}}), \tag{6.6}$$

where we put

$$\tilde{\mathbf{M}} \equiv \int_0^\sigma \mathbf{M}(\sigma')\, d\sigma' = C\tilde{K}\tilde{\mathbf{P}}, \tag{6.7}$$

$$\tilde{\mathbf{P}} \equiv \begin{Vmatrix} 0 & -\tilde{\mathbf{R}}^T \\ \tilde{\mathbf{R}} & 0 \end{Vmatrix}, \quad \tilde{\mathbf{R}} \equiv \frac{1}{\tilde{K}} \begin{Vmatrix} \tilde{K}_2 & -\tilde{K}_3 \\ \tilde{K}_3 & \tilde{K}_2 \end{Vmatrix}, \tag{6.8}$$

$$\tilde{K} \equiv (\tilde{K}_2{}^2 + \tilde{K}_3{}^2)^{1/2}, \quad \tilde{K}_K \equiv \int_0^\sigma K_K(\sigma')\, d\sigma'. \tag{6.9}$$

Expanding $\exp(\tilde{\mathbf{M}})$ and referring to

$$\tilde{\mathbf{M}}^2 = -C^2\tilde{K}^2\mathbf{1}, \tag{6.10}$$

where $\mathbf{1}$ is the 4×4 unit matrix, we have

$$\exp(\tilde{\mathbf{M}}) = \mathbf{1}\cos(C\tilde{K}) + \tilde{\mathbf{P}}\sin(C\tilde{K}). \tag{6.11}$$

Then we have the solutions

$$\bar{\boldsymbol{\phi}} = (1 - 2\sigma\Omega_0 + \sigma^2 K_0)^{-1/2} \times\{\bar{\boldsymbol{\phi}}_0 \cos(C\tilde{K}) - (2C)^{-1}\tilde{\mathbf{R}}^T \bar{\mathbf{g}}_0 \sin(C\tilde{K})\}, \qquad (6.12)$$

$$\bar{\mathbf{g}} = (1 - 2\sigma\Omega_0 + \sigma^2 K_0)^{-1/2} \times\{\bar{\mathbf{g}}_0 \cos(C\tilde{K}) + 2C\,\tilde{\mathbf{R}}\bar{\boldsymbol{\phi}}_0 \sin(C\tilde{K})\}, \qquad (6.13)$$

where

$$\bar{\boldsymbol{\phi}} \equiv \left\| \begin{matrix} \bar{\phi}_2 \\ \bar{\phi}_3 \end{matrix} \right\|, \quad \bar{\mathbf{g}} \equiv \left\| \begin{matrix} \bar{g}_{22} \\ \bar{g}_{23} \end{matrix} \right\| \qquad (6.14)$$

denote the two-dimensional vectors on a plane being tangent to the wavefront at a point under consideration and $\bar{\boldsymbol{\phi}}_0$ and $\bar{\mathbf{g}}_0$ denote, respectively, the initial values of $\bar{\boldsymbol{\phi}}$ and $\bar{\mathbf{g}}$ on $\mathcal{S}(\tau_0)$. Here the matrix $\tilde{\mathbf{R}}$ defined in (6.8) is an orthogonal matrix and indicates a rotion around the x^1 axis with the angle given by

$$\tilde{\theta} \equiv \tan^{-1}(\tilde{K}_3/\tilde{K}_2), \qquad (6.15)$$

then we have

$$\tilde{\mathbf{R}} = \left\| \begin{matrix} \cos\tilde{\theta} & -\sin\tilde{\theta} \\ \sin\tilde{\theta} & \cos\tilde{\theta} \end{matrix} \right\| \qquad (6.16)$$

and $\tilde{\mathbf{R}}^T$ gives a rotion with the angle $\tilde{\theta}$ in opposite direction.

The factor $(1 - 2\sigma\Omega_0 + \sigma^2 K_0)^{-1/2}$ in the *variation formulas* (6.12) and (6.13) denote the dependence of the amplitude on the form of an initial wavefront. For example, if we treat a centrifugal spherical wave, although we have no completely spherically symmetric waves, we can put

$$\Omega_0 = -1/R_0, \quad K_0 = 1/R_0{}^2, \qquad (6.17)$$

where R_0 is the radius of an initial wavefront. That factor is equal to $R_0/(R_0 + \sigma)$. Under the definition that the intensity of a wave defined by the Definitions 1 and 2 means the square of its amplitude, we have *the inverse square law* of the intensity of a wave.

The variation formulae have two terms. The first terms show the dependence on the *own waves*, and the second terms on the *dual waves*, where the electromagnetic and gravitational waves are called to be dual each other. Then we can say that *if there is no external fields, there is no transformation from the electromagnetic wave to the gravitational wave and vice versa.*

In the case of constant external fields, $\tilde{K}$ is proportional to the propagation distance σ, and $\tilde{\theta}$ reduces to a constant value, that is,

$$\tilde{K} = \{(E_2 - H_3)^2 + (E_3 + H_2)^2\}^{1/2}\sigma \equiv K\sigma, \qquad (6.18)$$

$$\tilde{\theta} = \tan^{-1}\frac{E_3 + H_2}{E_2 - H_3}. \qquad (6.19)$$

The first and second terms of the variation formulas vary, respectively, as cosine and sine forms with respect to the propagation distance of a wave. The *wave length* for variation is given by

$$\lambda = 2\pi/CK. \qquad (6.20)$$

The *length of interchangeability* l is defined by a length where a wave is completely transformed into its dual wave. We have

$$l = \lambda/4 = \pi/2CK. \qquad (6.21)$$

The wave length and the length of interchangeability are inversely proportional to the strength of the external field. Their numerical values in $K = 1$ gauss are

$$\lambda = 2.311 \times 10^7 \text{ light years,} \qquad l = 5.778 \times 10^6 \text{ light years.} \qquad (6.22)$$

If the region occupied by an external electromagnetic field is broad and/or if the intensity of that field is strong, the space—time in front of the wave can not be regarded as a flat space—time. In this case we must take into careful consideration for the formal application of the results obtained in this section. Strictly, the global mixing behavior of the waves must be analyzed in a curved space—time, but such analysis looks for future study.

[1] C. Truesdell and R. Toupin, *The Classical Field Theories*, The Encyclopedia of Physics (Springer-Verlag, Berlin, 1960), Vol. III/1, Chap. C.

[2] J. C. Maxwell, Phil. Trans. **155**, 459 (1865).

[3] A. Einstein, S.B. Preuss, Akad. Wiss., 688 (1916); 154 (1918).

[4] J. Weber, *General Relatively and Gravitational Waves* (Interscience, New York, 1961).

[5] V. D. Zakharov, *Gravitational Waves in Einstein's Theory*, translated by R.N. Sen (Halsted, New York, 1973).

[6] A. Lichnerowicz, *Théories rélativistes de la gravitation et de l'électromagnétisme* (Masson, Paris, 1955).

[7] A. Trautman, Bull. Acad. Polon. Sci. **5**, 273 (1957).

[8] T. Y. Thomas, J. Math. Anal. Appl. **3**, 315 (1961).

[9] T. Tokuoka, Arch. Ratl. Mech. Anal. **51**, 285 (1973).

[10] M. Johnston, R. Ruffini, and F. Zerilli, Phys. Rev. Lett. **31**, 1317 (1974).

[11] M. Johnston, R. Ruffini, and M. Peterson, Lett. Nuovo Cimento **9**, 217 (1974).

[12] T. Y. Thomas, Tensor (N.S.) **14**, 169 (1963).

[13] T. Y. Thomas, *Plastic Flow and Fracture in Solids* (Academic, New York, 1961), Chap. II.

[14] R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity* (McGraw-Hill, New York, 1965).

[15] T. Y. Thomas, *Concepts from Tensor Analysis and Differential Geometry* (Academic, New York, 1965), Chap. IV.

Radio wave propagation in nonuniform multilayered cylindrical structures—Generalized field transforms*

E. Bahar

Electrical Engineering Department, University of Nebraska, Lincoln, Nebraska 68508

(Received 7 January 1974; revised manuscript received 16 May 1974)

Bessel type transforms, formulated recently, provide suitable bases for the expansion of the electromagnetic fields due to line sources in the vicinity of cylindrical structures of variable curvature. However, these expansions are restricted to open structures with convex cylindrical boundaries that are characterized by surface impedances which are independent of excitation. In this paper, we derive transforms that are suitable for the expansion of electromagnetic fields in multilayered cylindrical structures with both concave and convex boundaries. The innermost medium of the structure is assumed to be a good conductor. These field transforms are shown to be related to Fourier-type transforms derived for parallel layered structures. They can be used to obtain rigorous, full wave solutions, to problems of propagation in irregular waveguide ducts in the earth's crust, the ionosphere, and in the cavity between the earth and the ionosphere.

1. INTRODUCTION

In order to provide suitable bases for the expansion of the electromagnetic fields in more realistic models of pertinent propagation problems over a wide frequency range, we derive generalized field transforms for multilayered cylindrical structures. This work is an extension of the analysis carried out earlier for open cylindrical structures with convex boundaries characterized by surface impedances that are independent of excitation.[1] The field transforms formulated in this paper are suitable for the derivation of rigorous, full wave solutions, to problems of propagation in nonuniform multilayered cylindrical structures with convex as well as concave boundaries. The innermost medium of the structure (earth's core) is assumed to be a good conductor, and the electromagnetic fields at a reference surface $r=a$ within the innermost medium are considered negligible. Thus for convenience we assume that the surface $r=a$ is perfectly conducting.

When the boundaries of the cylindrical or spherical waveguide structures are characterized by electromagnetic parameters $(\mu/\epsilon)^{1/2}\to 0$ or $(\epsilon/\mu)^{1/2}\to$ or by a constant surface impedance (independent of excitation) the characteristic (basis) functions used in the transforms are orthogonal over the cross section of the guiding structure.[2] However, for the general case considered in this paper the basis functions are shown to be orthogonal over the region $a\leqslant r\leqslant\infty$. The basis functions are normalized without recourse to direct integration over the cross section of the multilayered structure.

The field transforms derived can be used to solve problems of radiowave propagation in nonuniform multilayered cylindrical structures.[3] For spherical structures such as the earth–ionosphere waveguide, where azimuthal symmetry is preserved, it has been shown that a two-dimensional treatment based on a nonuniform cylindrical model is appropriate.[4,5]

Thus, the field transforms can also be applied to special problems of propagation in any of the nonuniform layers of the earth–ionosphere structure.

2. FORMULATION OF THE PROBLEM

A. magnetic line source $\bar{J}_m$ (analogous to an electric line current $\bar{J}$) parallel to the axis of a multilayered circular cylindrical structure (Fig. 1) excites vertically polarized waves that are independent of the variable z. Thus, for an $\exp(i\omega t)$ time dependence, the horizontal magnetic field $H_z\bar{a}_z$ satisfies the wave equation

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial H_z}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2 H_z}{\partial\phi^2}+k^2H_z=i\omega\epsilon J_m \tag{2.1a}$$

in which

$$J_m=K\delta(r-r_0)\delta(\phi-\phi_0)/r, \tag{2.1b}$$

and K is the intensity of the source measured in volts and $\delta(\alpha-\alpha_0)$ is the Dirac delta function. The wavenumber k for the ith layer is

$$k_i=\omega(\mu_i\epsilon_i)^{1/2},\quad i=0,\ldots,m \tag{2.1c}$$

where μ_i and ϵ_i are the permeability and complex permittivity for the ith layer. The dual problem, excitation of horizontally polarized waves by electric line sources,

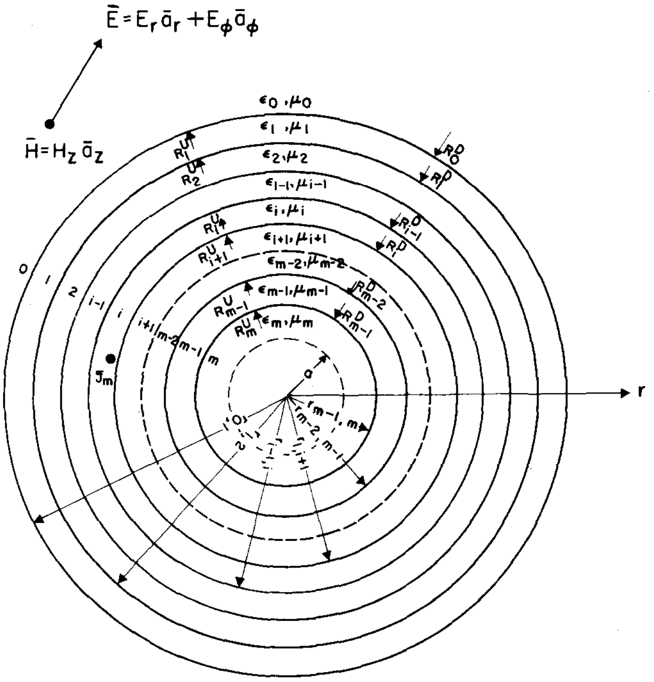


FIG. 1. Magnetic line source parallel to a multilayered circular cylindrical structure.

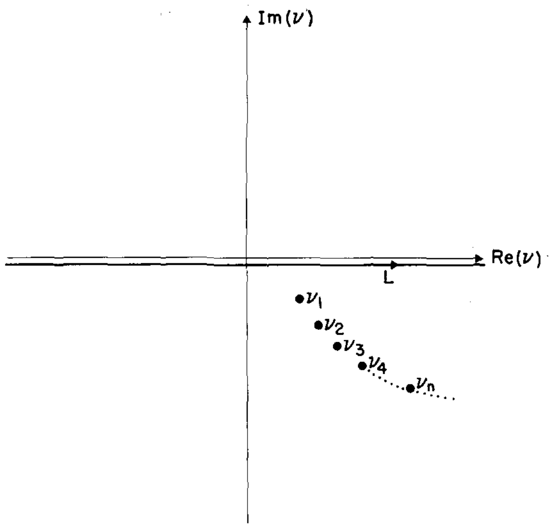


FIG. 2. Path of integration in the ν plane.

can be treated in a similar manner. Suitable expressions for the Dirac delta function $\delta(\phi-\phi_0)$ are obtained by considering the solution to the differential equation

$$\left(\frac{d^2}{d\phi^2}+\lambda\right)\Phi=\delta(\phi-\phi_0) \qquad (2.2a)$$

in conjunction with the periodic boundary conditions

$$\Phi(\phi+2\pi)=\Phi(\phi) \text{ and } \frac{d}{d\phi}\Phi(\phi+2\pi)=\frac{d}{d\phi}\Phi(\phi), \quad \phi\neq\phi_0. \qquad (2.2b)$$

Thus the Green's function $\Phi(\lambda,\phi)$ is[6]

$$\Phi(\lambda,\phi)=\cos[\nu(|\phi-\phi_0|-\pi)]/2\nu\sin\nu\pi, \quad |\phi-\phi_0|<2\pi, \qquad (2.3a)$$

where

$$\nu=\sqrt{\lambda}, \quad 0\leqslant\arg\lambda<2\pi \qquad (2.3b)$$

and the appropriate expansions for the Dirac delta function $\delta(\phi-\phi_0)$ are

$$\delta(\phi-\phi_0)=\frac{1}{2\pi i}\oint\Phi(\lambda,\phi)\,d\lambda \qquad (2.4a)$$

$$=\frac{1}{2\pi i}\int_L\frac{\cos[\nu(\phi-\phi_0-\pi)]}{\sin\nu\pi}\,d\nu \qquad (2.4b)$$

$$=\frac{1}{2\pi}\sum_n^{\infty}(2-\delta_{on})\cos n(\phi-\phi_0). \qquad (2.4c)$$

The contour in (2.4a) is $|\lambda|\to\infty$, $0\leqslant\arg\lambda\leqslant 2\pi$, and for (2.4b) the path of integration L is along the real axis as shown in Fig. 2. In (2.4c) δ_{mn} is the Kronecker delta.

We now seek a solution for $H_z(r,\phi)$ of the form

$$H_z(r,\phi)=\frac{1}{2\pi i}\int_L H(r,\nu)\,\frac{\cos[\nu(\phi-\phi_0-\pi)]}{\sin\nu\pi}\,d\nu. \qquad (2.5)$$

Substituting (2.5) for H_z and (2.4b) for $\delta(\phi-\phi_0)$ into (2.1a), we obtain the differential equation for $H(r,\nu)$

$$\left(r\frac{d}{dr}r\frac{d}{dr}+(kr)^2-\nu^2\right)H(r,\nu)=C\,r\,\delta(r-r_0) \qquad (2.6a)$$

where

$$C=i\omega\epsilon_i K \text{ and } \epsilon_i=\epsilon(r_0), \quad r_{i-1,i}>r_0>r_{i,i+1}. \qquad (2.6b)$$

Since the tangential electric and magnetic fields are continuous at each interface where $r=r_{i,i+1}$ (Fig. 1), $H(r,\nu)$ satisfies the boundary conditions

$$H(r^+_{i,i+1},\nu)=H(r^-_{i,i+1},\nu) \qquad (2.7a)$$

and

$$\frac{1}{\epsilon_i}\frac{\partial}{\partial r}H(r^+_{i,i+1},\nu)=\frac{1}{\epsilon_{i+1}}\frac{\partial}{\partial r}H(r^-_{i,i+1},\nu). \qquad (2.7b)$$

Hence the solution to the Green's function $H(r,\nu)$ [(2.6)] is

$$H(r,\nu)=A_\nu\begin{cases}\psi^D_\nu(r)\cdot\psi^U_\nu(r_0), & r<r_0\\ \psi^U_\nu(r)\cdot\psi^D_\nu(r_0), & r>r_0\end{cases} \qquad (2.8a)$$

where $H^{(1)}_\nu(z)$ and $H^{(2)}_\nu(z)$ are Hankel functions of the first and second kind and

$$A_\nu=C\pi iR^{Dr}_i/4[1-R^{Dr}_iR^{Ur}_i], \qquad (2.8b)$$

$$\psi^D_\nu(r)=\frac{1}{R^{Dr}_i}\times\begin{cases}\prod_{p=1}^{i-n}(T^{DH}_{i+1-p}/T^D_{i-p})S^1_{i+1-p,i-p}[H^{(1)}_\nu(k_nr)+R^{Dr}_nH^{(2)}_\nu(k_nr)], \\ \quad n=i-1,\ldots,0\\ H^{(1)}_\nu(k_ir)+R^{Dr}_iH^{(2)}_\nu(k_ir),\\ \prod_{p=1}^{n-i}(T^D_{i+p-1}/T^{DH}_{i+p})S^1_{i+p-1,i+p}[H^{(1)}_\nu(k_nr)+R^{Dr}_nH^{(2)}_\nu(k_nr)],\\ \quad n=i+1,\ldots,m\end{cases} \qquad (2.9a)$$

and

$$\psi^U_\nu(r)=\begin{cases}\prod_{p=1}^{i-n}(T^U_{i+1-p}/T^{UH}_{i-p})S^2_{i+1-p,i-p}[H^{(2)}_\nu(k_nr)+R^{Ur}_nH^{(1)}_\nu(k_nr)], \\ \quad n=i-1,\ldots,0\\ H^{(2)}_\nu(k_ir)+R^{Ur}_iH^{(1)}_\nu(k_ir),\\ \prod_{p=1}^{n-i}(T^{UH}_{i+p-1}/T^U_{i+p})S^2_{i+p-1,i+p}[H^{(2)}_\nu(k_nr)+R^{Ur}_nH^{(1)}_\nu(k_nr)],\\ \quad n=i+1,\ldots,m.\end{cases} \qquad (2.9b)$$

The reflection coefficient at $r=r_{i,i+1}$ looking in the $-r$ direction is R^D_i and the reflection coefficient at $r=r_{i-1,i}$ looking in the $+r$ direction is R^U_i.

Thus,

$$R^D_i=\frac{R_{i+1,i}+R^{DH}_{i+1}(1+R_{i+1,i}+R_{i,i+1})}{1-R^{DH}_{i+1}R_{i,i+1}}, \quad i=0,1\cdots m-1 \qquad (2.10a)$$

$$R^U_i=\frac{R_{i-1,i}+R^{UH}_{i-1}(1+R_{i-1,i}+R_{i,i-1})}{1-R^{UH}_{i-1}R_{i,i-1}}, \quad i=1,2\cdots m, \qquad (2.10b)$$

in which $R_{i,i+1}$ and $R_{i+1,i}$ are for $i=0, 1\ldots m-1$

$$R_{i,i+1}=-\frac{\eta_{i+1}\ln'H^{(2)}_\nu(k_{i+1}r_{i,i+1})-\eta_i\ln'H^{(2)}_\nu(k_ir_{i,i+1})}{\eta_{i+1}\ln'H^{(1)}_\nu(k_{i+1}r_{i,i+1})-\eta_i\ln'H^{(2)}_\nu(k_ir_{i,i+1})} \qquad (2.11a)$$

and

$$R_{i+1,i}=-\frac{\eta_{i+1}\ln' H_\nu^{(1)}(k_{i+1}r_{i,i+1})-\eta_i\ln' H_\nu^{(1)}(k_i r_{i,i+1})}{\eta_{i+1}\ln' H_\nu^{(1)}(k_{i+1}r_{i,i+1})-\eta_i\ln' H_\nu^{(2)}(k_i r_{i,i+1})} \tag{2.11b}$$

where $\ln' u(x)=(du/dx)/u$. For $r_{i,i+1}\rightarrow\infty$, $R_{i,i+1}=-R_{i+1,i}$ and (2.11a) and (2.11b) reduce to the familiar Fresnel reflection coefficients for two semi-infinite media. The ith medium intrinsic impedance is

$$\eta_i=(\mu_i/\epsilon_i)^{1/2}. \tag{2.11c}$$

For $i=0,1,2\cdots m-1$

$$R_i^{Dr}=R_i^D H_\nu^{(1)}(k_i r_{i,i+1})/H_\nu^{(2)}(k_i r_{i,i+1}). \tag{2.12a}$$

Similarly, for $i=1,2,3\cdots m$

$$R_i^{Ur}=R_i^U H_\nu^{(2)}(k_i r_{i-1,i})/H_\nu^{(1)}(k_i r_{i-1,i}). \tag{2.12b}$$

In view of the radiation condition, and since the surface $r=a$ is perfectly conducting

$$R_0^{Ur}=0 \text{ and } R_m^{Dr}=-H_\nu^{(1)\prime}(k_m a)/H_\nu^{(2)\prime}(k_m a). \tag{2.12c}$$

Furthermore,

$$R_i^{DH}=R_i^{Dr}H_\nu^{(2)}(k_i r_{i-1,i})/H_\nu^{(1)}(k_i r_{i-1,i}),\quad i=1,2,\ldots,m,$$
$$R_i^{UH}=R_i^{Ur}H_\nu^{(1)}(k_i r_{i,i+1})/H_\nu^{(2)}(k_i r_{i,i+1}),\quad i=0,1,\ldots m-1, \tag{2.12d}$$

$$T_i^D=1+R_i^D,\quad T_i^{DH}=1+R_i^{DH},\quad T_i^U=1+R_i^U,$$
$$T_i^{UH}\quad=1+R_i^{UH}, \tag{2.12e}$$

and

$$S^1_{n-1,n}=H_\nu^{(1)}(k_{n-1}r_{n-1,n})/H_\nu^{(1)}(k_n r_{n-1,n})=1/S^1_{n,n-1}, \tag{2.13a}$$
$$S^2_{n-1,n}=H_\nu^{(2)}(k_{n-1}r_{n-1,n})/H_\nu^{(2)}(k_n r_{n-1,n})=1/S^2_{n,n-1}. \tag{2.13b}$$

It follows from the properties of the Hankel functions that $H(r,\nu)$ is an even function of ν and the expansion (2.5) is justified.

Substituting (2.8) into (2.5) the magnetic field can be expressed as

$$H_z(r,\phi)=\frac{C}{8}\int_L\frac{\cos[\nu(\phi-\phi_0-\pi)]}{\sin\nu\pi[(1/R_i^{Dr})-R_i^{Ur}]}[\psi_\nu^D(r)\psi_\nu^U(r_0)$$
$$+\{\psi_\nu^U(r)\psi_\nu^D(r_0)-\psi_\nu^D(r)\psi_\nu^U(r_0)\}U(r-r_0)]\,d\nu, \tag{2.14}$$

in which $U(r-r_0)$ is the unit step function. The location of the poles of the integrand in the lower half plane (below the path of integration L) is given by the solutions $\nu=\nu_n$ $(n-1,2,3\cdots)$ of the modal equation

$$(1/R_i^{Dr})-R_i^{Ur}=0,\quad \mathrm{Im}(\nu_n)\leqslant 0. \tag{2.15a}$$

From (2.9), it follows that for $\nu=\nu_n$ (see Fig. 2)

$$\psi_\nu^D(r)=\psi_\nu^U(r)\equiv\psi_\nu(r). \tag{2.15b}$$

Hence, the factor of $U(r-r_0)$ in (2.14) vanishes at $\nu=\nu_n$. Furthermore, the expansion (2.4) is symmetric in ϕ and ϕ_0 and over the semicircle $|\nu|\rightarrow\infty$ in the lower half plane

$$\frac{\cos[\nu(|\phi-\phi_0|-\pi)]}{\sin\nu\pi}\rightarrow\exp|\phi-\phi_0|\,\mathrm{Im}(\nu). \tag{2.16a}$$

Therefore, it follows that (2.14) reduces to

$$H_z(r,\phi)=\frac{iK}{2}\int_L\frac{\cos[\nu(|\phi-\phi_0|-\pi)]}{\sin\nu\pi}\,\psi_\nu^D(r)\psi_\nu^U(r_0)\,\frac{d\nu}{D(\nu)}, \tag{2.16b}$$

in which

$$D(\nu)=4[(1/R_i^{Dr})-R_i^{Ur}]/\omega\epsilon_i. \tag{2.16c}$$

Since medium m is a good conductor, for $|k_m(r_{m-1,m}-a)|\gg 1$, $R_m^{DH}\rightarrow 0$. Therefore, the expression for $H_z(r,\phi)$ [(2.16b)] for $r\geqslant r_{m-1,m}$ does not depend on the particular value of a, as expected.

3. SPECTRAL REPRESENTATIONS FOR $\delta(r-r_0)$

To obtain the appropriate complete expansions for $\delta(r-r_0)$, we substitute (2.16b) for H_z into the differential equation (2.1). Noting that the Green's function $\Phi(\lambda,\phi)$ [(2.3a)], satisfies (2.2a) and that ψ_ν^U and ψ_ν^D [(2.9)] are solutions to Bessel's differential equation, it follows that

$$\delta(r-r_0)=\int_L\psi_\nu^D(kr)\psi_\nu^U(kr_0)Z_\nu(r)\,\frac{d\nu}{D(\nu)} \tag{3.1a}$$

in which the transverse wave impedance $Z_\nu(r)$ is

$$Z_\nu(n)=\nu/\omega\epsilon r=\eta\nu/kr. \tag{3.1b}$$

From the completeness relationship (3.1), we obtain the following transform for the magnetic field

$$H_z(r,\phi)=\int_L H(\nu,\phi)\psi_\nu^D(r)\,\frac{C(\nu)}{D(\nu)}\,d\nu, \tag{3.2a}$$

where

$$H(\nu,\phi)=\int_a^\infty H_z(r,\phi)\psi_\nu^U(r)\,\frac{Z_\nu(r)}{C(\nu)}\,dr \tag{3.2b}$$

and $C(\nu)$ is a normalization coefficient.

From Maxwell's equations, we have

$$\frac{1}{r}\,\frac{\partial H_z}{\partial\phi}=i\omega\epsilon E_r, \tag{3.3a}$$

$$-\frac{\partial H_z}{\partial r}=i\omega\epsilon E_\phi, \tag{3.3b}$$

and

$$\frac{1}{r}\left(\frac{\partial}{\partial r}(rE_\phi)-\frac{\partial}{\partial\phi}E_r\right)=-i\omega\mu H_z-J_m. \tag{3.3c}$$

Hence, the basis function for the transverse component of the electric field E_r is $-\psi_\nu^D(r)Z_\nu(r)$ and the appropriate transform pair is

$$E_r(r,\phi)=-\int_L E(\nu,\phi)Z_\nu(r)\psi_\nu^D(r)\,\frac{C(\nu)}{D(\nu)}\,d\nu, \tag{3.4a}$$

$$E(\nu,\phi)=-\int_a^\infty E_r(r,\phi)\psi_\nu^U(r)\,\frac{dr}{C(\nu)}. \tag{3.4b}$$

The orthogonality relationship corresponding to (3.1) is

$$\delta(\mu-\nu)=\int_a^\infty\psi_\mu^D(r)\psi_\nu^U(r)Z_\nu(r)\,\frac{dr}{D(\nu)}. \tag{3.5}$$

The general expressions for the transform pairs (3.2) and (3.4) correspond to the Bessel transform derived earlier for open cylindrical structures characterized by convex, surface impedance boundaries.[1]

The corresponding Watson type transforms[7] may be obtained by deforming the path of integration L in the lower half ν plane. Thus accounting for the residues at the poles $\nu=\nu_n$ [(2.15a)] and using (2.15b), we obtain

$$\delta(r-r_0)=-2\pi i\sum_{n=1}^{\infty}\psi_\nu(r)\psi_\nu(r_0)Z_\nu(r)/(\partial D(\nu)/\partial\nu), \quad \nu=\nu_n, \tag{3.6}$$

and

$$H_z(r,\phi)=\sum_n H(\nu,\phi)\psi_\nu(r)/N_\nu, \tag{3.7a}$$

$$H(\nu,\phi)=\int_a^\infty H_z(r,\phi)\psi_\nu(r)Z_\nu(r)\,\frac{dr}{M_\nu}, \tag{3.7b}$$

$$E_r(r,\phi)=-\sum_n E(\nu,\phi)\psi_\nu(r)Z_\nu(r)/N_\nu, \tag{3.8a}$$

$$E(\nu,\phi)=-\int_a^\infty E_r(r,\phi)\psi_\nu(r)\frac{dr}{M_\nu}, \tag{3.8b}$$

in which the product of the normalization coefficients M_ν and N_ν is

$$M_\nu N_\nu=\frac{i}{2\pi}\,\frac{\partial D(\nu)}{\partial\nu}, \quad \nu=\nu_n. \tag{3.9}$$

The orthonormal relationship corresponding to (3.6) is

$$\delta_{n,m}=\int_a^\infty \psi_\mu(r)\psi_\nu(r)Z_\nu(r)\,\frac{dr}{M_\mu N_\nu} \tag{3.10a}$$

in which $\mu=\mu_m$ and $\nu=\nu_n$ $(m,n=1,2,3\cdots)$ are solutions of the modal equation (2.15a). The orthogonal relationship (3.10) $(n\neq m)$ can be verified by direct integration of (3.10a). Integrating over each layer separately, we get

$$\int_a^\infty \psi_\nu(r)\psi_\mu(r)Z_\nu(r)\,dr=\sum_{i=1}^{m}\frac{1}{\mu^2-\nu^2}\left[\frac{\nu r}{\omega\epsilon}\left(\psi_\mu(r)\frac{\partial\psi_\nu(r)}{\partial r}-\psi_\nu(r)\,\frac{\partial\psi_\mu(r)}{\partial r}\right)\right]_{r_{i-1,i}^-}^{r_{i-1,i}^+}-\frac{1}{\mu^2-\nu^2}\left[\frac{\nu r}{\omega\epsilon}\left(\psi_\mu(r)\,\frac{\partial\psi_\nu(r)}{\partial r}-\psi_\nu(r)\,\frac{\partial\psi_\mu(r)}{dr}\right)\right]_a^\infty. \tag{3.10b}$$

In view of the boundary conditions (2.7) and (2.12c), it can be shown that (3.10b) vanishes. It is interesting to note that the basis functions $\psi_\nu(r)$ are orthogonal over the range $a<r<\infty$, and not over the finite region $r_{i-1,i}<r<r_{i,i+1}$. If, however, the boundaries $r=r_{i-1,i}$ and $r=r_{i,i+1}$ are characterized by the surface impedances Z_s^U and Z_s^D, respectively, the exact boundary conditions (2.7) are replaced by the approximate boundary conditions

$$i\eta\ln' H(r,\nu)=\begin{cases} Z_s^U, & r=r_{i-1,i},\\ -Z_s^D, & r=r_{i,i+1}.\end{cases} \tag{3.11a}$$

The corresponding reflection coefficients are

$$R_i^U=\frac{i\eta_i\ln' H_\nu^{(2)}(k_i r_{i-1,i})-Z_s^U}{-i\eta_i\ln' H_\nu^{(2)}(k_i r_{i-1,i})+Z_s^D} \tag{3.11b}$$

and

$$R_i^D=\frac{-i\eta_i\ln' H_\nu^{(1)}(k_i r_{i,i+1})-Z_s^D}{i\eta_i\ln' H_\nu^{(2)}(k_i r_{i,i+1})+Z_s^D}\,. \tag{3.11c}$$

In this case, it can be readily verified that the basis functions are orthogonal over the region $r_{i-1,i}<r<r_{i,i+1}$. The value of the product $M_\nu N_\nu$, can also be derived by direct integration of (3.10a) for $m=n$; however, this tedious manipulation has been avoided and the result is given in closed form (3.9). For the special case when ϵ_i and μ_i are the same for all i, and $a\to 0$, it can be shown that the transform (3.2) reduces to the Kontorowich Lebedev transform.[1,8]

When the radius of the cylindrical structure is very large and the effects of curvature can be neglected, the field transforms (3.2) and (3.3) can be shown to merge with the generalized Fourier type transforms derived recently for multilayered structures.[9] To this end, we set

$$\nu\equiv\beta R \quad\text{and}\quad u_i\equiv(k_i^2-\beta^2)^{1/2}, \quad \mathrm{Im}(\beta)\leqslant 0, \quad \mathrm{Im}(u_i)\leqslant 0, \tag{3.12a}$$

$$r=y+R, \quad r_{i-1,k}=h_{i-1,i}+R, \quad x=R\phi. \tag{3.12b}$$

The normalization coefficient $C(\nu)$ is chosen to be

$$C(\nu)=2\pi\left(\frac{\nu}{\omega\epsilon_0 R}\right)H_\nu^{(2)}(k_0R)/R_0^{Dr} \quad\text{and}\quad D(\nu)=4/R_0^{Dr}\,\omega\epsilon_0. \tag{3.12c}$$

Taking the limit $R/\lambda\to\infty$, (3.2a) and (3.2b) reduce to

$$H_z(y,x)=\int_{-\infty}^{\infty} H(\beta,x)\psi^{Dh}(y,u)\,\frac{\beta d\beta}{u_0} \tag{3.13a}$$

and

$$H(\beta,x)=\int_{-\infty}^{\infty}H_z(y,x)\psi^{Uh}(y,u)Z(u,y)\,dy, \tag{3.13b}$$

in which

$$\psi^{Dh}(y,u)\to[\exp(iu_0y)+R_0^{Dh}\exp(-iu_0y)]/R_0^{Dh}, \quad y\geqslant h_{0,1} \tag{3.14a}$$

$$\psi^{Uh}(y,u)\to R_0^{Dh}\exp(-iu_0y)/2\pi Z_0(u), \quad y\geqslant h_{0,1}, \tag{3.14b}$$

$$Z(u,y)=\beta/\omega\epsilon, \quad Z_0(u)=\beta/\omega\epsilon_0, \tag{3.14c}$$

and R_0^{Dh} is the reflection coefficient for a horizontally stratified structure

$$R_0^{Dh}=\lim_{R\to\infty}R_0^{Dr}. \tag{3.14d}$$

The expressions for ψ^{Dh} and ψ^{Uh}, for $y<h_{0,1}$, may also be obtained by taking the limit $R\to\infty$; however, they can be written directly using (3.14) since $\psi^{Dh},\psi^{Uh},\epsilon^{-1}\partial\psi^{Dh}/\partial y$ and $\epsilon^{-1}\partial\psi^{Uh}/\partial y$ are continuous at each interface. The integrand in (3.13a) has branch points at $u_0=0$ and $u_m=0$ and poles at $1/R_0^{Dh}=0$. Thus, it can be shown that the magnetic field can be expressed in terms of a Fourier-type transform consisting of two branch cut integrals (the radiation and lateral wave terms) and a finite number of surface wave terms which are due to the residues at the poles of the integrand (3.13a).[9]

In view of the boundary conditions (2.7) for multilayered cylindrical structures and the property of the conjunct J for Bessel functions, the quantity $J(\psi_\nu^D,\psi_\nu^U)/\epsilon=\text{const}$ for $a<r<\infty$. Hence

$$J(\psi_\nu^D,\psi_\nu^U)/\epsilon=\frac{r}{\epsilon}[\psi_\nu^D\partial\psi_\nu^U/\partial r-\psi_\nu^U\partial\psi_\nu^D/\partial r]=\text{const}. \tag{3.15a}$$

On substituting (2.9) into (3.15) it follows that for $r_{i,i+1}\leqslant r\leqslant r_{i-1,i}$

$$J(\psi_\nu^D,\psi_\nu^U)/\epsilon=(1-R_i^{Dr}R_i^{Ur})J(H_\nu^{(1)},H_\nu^{(2)})/\epsilon_i R_i^{Dr}. \tag{3.15b}$$

Hence, for values of ν satisfying the modal equation (2.15a),

$$J(\psi_\nu^D, \psi_\nu^U)/\epsilon = 0, \quad \nu = \nu_n, \quad a < r < \infty. \tag{3.15c}$$

But $J(\psi_\nu^D, \psi_\nu^U)$ is proportional to $[1 - R_p^{Dr} R_p^{Ur}]$ for $p = 0, 1, 2 \cdots m$. Thus, the resonance condition

$$(1/R_p^{Dr}) - R_p^{Ur} = 0 \tag{3.15d}$$

is satisfied in each layer of the structure ($p = 0, 1, 2 \cdots m$) for the same values of $\nu = \nu_n$ ($n = 1, 2, 3 \cdots$).[9]

4. DIFFERENTIAL EQUATIONS FOR THE WAVE AMPLITUDES

It is often more convenient to express the electromagnetic fields in terms of the forward and backward propagating wave amplitudes $a(\nu, \phi)$ and $b(\nu, \phi)$, respectively. In view of the normalization chosen in (3.7) and (3.8), we set

$$H(\nu, \phi) = a(\nu, \phi) + b(\nu, \phi) \tag{4.1a}$$

and

$$E(\nu, \phi) = a(\nu, \phi) - b(\nu, \phi). \tag{4.1b}$$

Eliminating E_ϕ from Maxwell's equations (3.3), we get for the transverse field components

$$\partial H_z/\partial\phi = i\omega\epsilon r E_r \tag{4.2a}$$

and

$$\partial E_r/\partial\phi = i\omega\mu \left(r H_z + \frac{1}{k^2}\frac{\partial}{\partial r}(r\,\partial H_r/\partial r)\right) + r J_m. \tag{4.2b}$$

Expressing H_z and E_r in terms of the transforms $H(\nu, \phi)$ and $E(\nu, \phi)$ and using the orthogonality relationship (3.10a), we get the inhomogeneous telegraphist's equations

$$-dH(\nu, \phi)/d\phi = i\nu E(\nu, \phi) \tag{4.3a}$$

and

$$-dE(\nu, \phi)/d\phi = i\nu H(\nu, \phi) + K\psi_\nu(r_0)\delta(\phi - \phi_0)/M_\nu. \tag{4.3b}$$

Thus the differential equations for the wave amplitudes are

$$-da(\nu, \phi)/d\phi = i\nu a(\nu, \phi) + K\psi_\nu(r_0)\delta(\phi - \phi_0)/2M_\nu \tag{4.4a}$$

and

$$-db(\nu, \phi)/d\phi = -i\nu b(\nu, \phi) - K\psi_\nu(r_0)\delta(\phi - \phi_0)/2M_\nu. \tag{4.4b}$$

The solutions to (4.4) for $0 < \phi - \phi_0 < 2\pi$ are

$$a(\nu, \phi) = iK\psi_\nu(r_0)\exp[-i\nu(\phi - \phi_0 - \pi)]/4M_\nu \sin\nu\pi \tag{4.5a}$$

and

$$b(\nu, \phi) = iK\psi_\nu(r_0)\exp[i\nu(\phi - \phi_0 - \pi)]/4M_\nu \sin\nu\pi. \tag{4.5b}$$

Thus on substituting (4.5) into (3.7a), we get for $|\phi - \phi_0| < 2\pi$

$$H_z(r, \phi) = \pi K \sum_{n=1}^{\infty} \frac{\cos\nu(|\phi - \phi_0| - \pi)}{\sin\nu\pi} \frac{\psi_\nu(r)\psi_\nu(r_0)}{\partial D(\nu)/\partial\nu}, \quad \nu = \nu_n. \tag{4.6}$$

The above solution (4.6) can be shown to agree with (2.16b). For dissipative media and when $kr \gg 1$ and $kr_0 \gg 1$, only the direct waves between the source and observation point need be retained since

$$\frac{\cos\nu(|\phi - \phi_0| - \pi)}{\sin\nu\pi} \to i\{\exp(-i\nu|\phi - \phi_0|) + \exp[i\nu(|\phi - \phi_0| - 2\pi)]\}. \tag{4.7}$$

5. CONCLUDING REMARKS

In this paper, we have derived generalized transforms that are suitable for the expansion of electromagnetic fields in multilayered cylindrical structures. The relationship between these transforms and the Watson and Kontorowich–Lebedev transforms has been established. It has also been shown that these field transforms merge with the Fourier-type transforms derived earlier for parallel stratified media when the radius of curvature of the structure, $R \to \infty$. The characteristic (basis) functions used in the analysis are shown to be orthogonal over the region $a < r \leq \infty$ when medium m is assumed to be a good conductor and $|k_m(r_{m-1} - a)| \gg 1$. The characteristic functions are orthogonal, over the cross section of one of the layers of the structure $r_{i,i+1} \leq r \leq r_{i-1,i}$, when the approximate surface impedance concept is employed. The normalization coefficients for the basis functions are derived directly in closed form without recourse to integration over the region $a < r < \infty$.

The generalized transforms provide the basic mathematical tools for the analysis of radio wave propagation in irregular cylindrical waveguide structures in which both the thickness and the electromagnetic parameters of the structure's layers vary as a function of distance along the path of propagation.[3] In this case it is seen that the differential equations for the wave amplitudes $a(\nu, \phi)$ and $b(\nu, \phi)$ are coupled.

*The research reported in this paper was supported by the National Science Foundation and the Engineering Research Center of the University of Nebraska. The author wishes to thank W. L. Brogan and F. G. Ullman for their comments. The manuscript was typed by Mrs. E. Everett.

[1] E. Bahar, J. Math. Phys. 12, 179, 1971.

[2] J.R. Wait, *Electromagnetic Waves in Stratified Media* (Pergamon, Oxford, 1962).

[3] E. Bahar, J. Math. Phys. 15, (1974).

[4] J.R. Wait, J. Res. Natl. Bur. Stand. (U.S.) D Radio Propag. 68, 81 (1964).

[5] E. Bahar, IEEE Trans. Antennas Propag., **AP-16**, (6), 673 (1968).

[6] B. Friedman, *Principles and Techniques of Applied Mathematics* (Wiley, London, 1956).

[7] G.N. Watson, Proc. Royal Soc. A 95, 546 (1919).

[8] M.J. Kontorowich and N.N. Lebedev, J. Phys. 1, (3), 229 (1939).

[9] E. Bahar, J. Math. Phys. 14, 1024 (1973).

Radio wave propagation in nonuniform multilayered cylindrical structures—Full wave solutions*

E. Bahar

Electrical Engineering Department, University of Nebraska, Lincoln, Nebraska 68508

(Received 7 January 1974; revised manuscript received 16 May 1974)

Full wave solutions to problems of propagation in irregular cylindrical structures are derived. The electromagnetic parameters and the thickness of the layers of the structure are assumed to vary along the path of propagation. Exact boundary conditions are imposed at each interface between the layers of the structure, and the solutions are shown to satisfy the reciprocity relationship in electromagnetic theory. This analysis is a generalization of an earlier solution to the problem of propagation in open cylindrical structures characterized by convex, surface impedance boundaries. It can be employed to solve more realistic models of pertinent propagation problems over a broad frequency range. These include problems of propagation in the earth's crust and in the ionosphere and over widely varying propagation paths on the surface of the earth. For the purpose of the analysis, we employ generalized field transforms that provide suitable bases for the expansion of the electromagnetic fields in nonuniform cylindrical structures.

1. INTRODUCTION

Due to the lack of rigorous solutions to propagation problems in irregular structures, it is often necessary to idealize significantly the original problem under consideration to obtain tractable solutions. In this paper, we derive full wave solutions to the problem of propagation in irregular multilayered cylindrical structures in order to investigate more suitable models of propagation problems. It is assumed, in this analysis, that the electromagnetic parameters μ and ϵ and the thickness of the structure's layers are functions of distance along the propagation path.

Rigorous boundary conditions are imposed at each interface of the nonuniform multilayered structure and the solutions are shown to satisfy the reciprocity relationships in electromagnetic theory. Thus, these solutions are applicable over a very wide frequency range and they are valid for the region between the innermost boundary $r=r_{m-1,m}$ around a good conducting core and $r\to\infty$. Since propagation though the conducting core is negligible, a reference surface $r=a<r_{m-1,m}$ is assumed to be perfectly conducting for convenience. The region of propagation is assumed to extend to infinity since, in general, it is not justified to represent the lower ionosphere by a sharp boundary characterized by a surface impedance independent of excitation. Using a uniform, spherically stratified model of the earth—ionosphere waveguide, Tran and Polk[1] have recently demonstrated the limitation of using an impedance boundary condition derived from a planar model of the lower ionosphere. Furthermore, in irregular multilayered structures, the impedance boundary conditions introduce additional errors in the analysis.[2]

Generalized field transforms for the transverse components of the electromagnetic fields are used to facilitate the full wave analysis.[3]

2. FORMULATION OF THE PROBLEM

The ith layer of the irregular cylindrical structure we consider is characterized by the inhomogeneous complex dielectric coefficient $\epsilon_i(\phi)$ and permeability $\mu_i(\phi)$ $(i=0,1,\ldots,m)$. The interface between two adjacent layers of the structure is given by $r_{i-1,i}(\phi)$ (see Fig. 1). Thus, for $r_{i-1,i}\leq r\leq r_{i,i+1}$,

$$\epsilon(r,\phi)=\epsilon_i(\phi)\quad\text{and}\quad\mu(r,\phi)=\mu_i(\phi).$$

For axially oriented line source excitations, the electromagnetic fields are also independent of the axial variable z. We consider, in detail, only excitation by magnetic line sources $\bar{J}_m$ that radiate vertically polarized waves. Solution to the dual problem, excitation by electric line sources $\bar{J}$ (that radiate horizontally polarized waves) can be derived in a similar manner. An $\exp(i\omega t)$ time dependence is assumed for the electromagnetic fields. Thus, for a z-directed line source $J_m(r,\phi)$ of strength K (volts) located at $r=r_0$ and $\phi=\phi_0$, the nonvanishing transverse components of the electromagnetic fields are[3]

$$\partial H_z/\partial\phi=i\omega\epsilon rE_r \tag{2.1a}$$

and

$$\partial E_r/\partial\phi=i\omega\mu\left(rH_z+\frac{1\partial}{k^2\partial r}\left(r\partial H_z/\partial r\right)\right)+rJ_m \tag{2.1b}$$

in which $k=\omega(\mu\epsilon)^{1/2}$ is the wavenumber,

$$J_m(r,\phi)=K\delta(\phi-\phi_0)\delta(r-r_0)/r \tag{2.1c}$$

and $\delta(\alpha-\alpha_0)$ is the Dirac delta function.

Since the tangential components of the electric and magnetic fields at each interface $r=r_{i-1,i}$ $(i=1,2\ldots m)$ must be continuous, the exact boundary conditions at each interface of the irregular cylindrical structure are

$$H_z(r^{+}_{i-1,i},\phi)=H_z(r^{-}_{i-1,i},\phi) \tag{2.2a}$$

and

$$\bar{n}_{i-1,i}\times\left[\bar{E}(r^{+}_{i-1,i},\phi)-\bar{E}(r^{-}_{i-1,i},\phi)\right]=0 \tag{2.2b}$$

in which the unit vector normal to the interface $r=r_{i-1,i}$ is

$$\bar{n}_{i-1,i}=\left(\bar{a}_r-\frac{1}{r_{i-1,i}}\,\frac{d}{d\phi}r_{i-1,i}\,\bar{a}_\phi\right)\times\left[1+\left(\frac{1}{r_{i-1,i}}\,\frac{d}{d\phi}\,r_{i-1,i}\right)^2\right]^{-1/2}. \tag{2.2c}$$

Thus,

$$\ln' r_{i-1,i}\left[E_r(r^{+}_{i-1,i},\phi)-E_r(r^{-}_{i-1,i},\phi)\right]+E_\phi(r^{+}_{i-1,i},\phi)-E_\phi(r^{-}_{i-1,i},\phi)=0 \tag{2.2d}$$

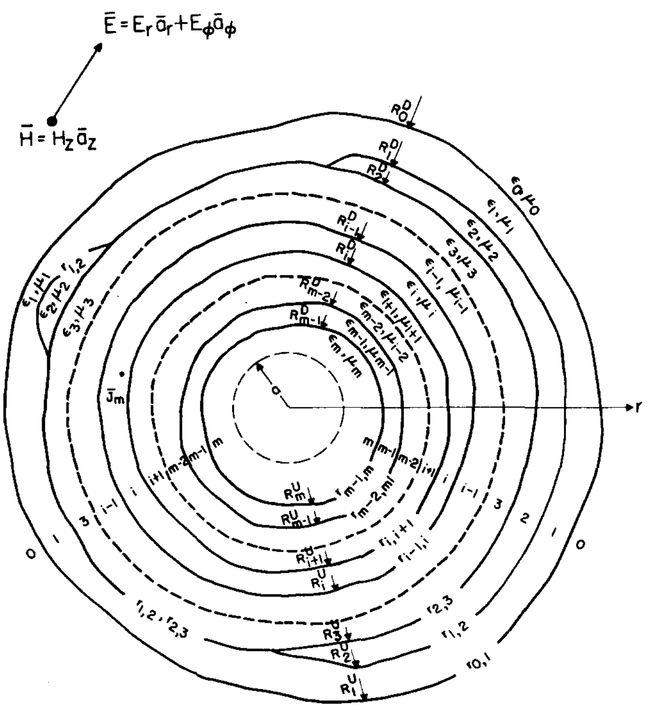


FIG. 1. Magnetic line source in irregular multilayered cylindrical structure.

where $\ln' u = (du/d\phi)/u$ and on eliminating E_ϕ we get

$$\ln' r_{i-1,i}(E_r^+ - E_r^-) + i\eta_{i-1}\frac{\partial H_z^+}{\partial(k_{i-1}r)} - i\eta_i\frac{\partial H_z^-}{\partial(k_i r)} = 0 \quad (2.2e)$$

in which $\eta_i = (\mu_i/\epsilon_i)^{1/2}$ is the intrinsic impedance for the ith layer. We assume that $r_{i-1,i}$ is finite for all $i = 1, 2 \ldots m$; therefore, the transverse components of the electromagnetic fields, E_r and H_z, are expressed in terms of their generalized Watson transforms[3]:

$$H_z(r, \phi) = \sum_1^\infty H(\nu, \phi)\psi_\nu(r)/N_\nu, \quad \nu = \nu_n, \quad n = 1, 2, 3 \cdots, \quad (2.3a)$$

$$H(\nu, \phi) = \int_a^\infty H_z(r, \phi)\psi_\nu(r)Z_\nu(r)\frac{dr}{M_\nu} \quad (2.3b)$$

and

$$E_r(r, \phi) = -\sum_{n=1}^\infty E(\nu, \phi)\psi_\nu(r)Z_\nu(r)/N_\nu, \quad \nu = \nu_n, \quad n = 1, 2, 3 \cdots, \quad (2.4a)$$

$$E(\nu, \phi) = -\int_a^\infty E_r(r, \phi)\psi_\nu(r)\frac{dr}{M_\nu}, \quad (2.4b)$$

in which

$$\psi_\nu(r) = \frac{1}{R_i^{Dr}} \times \begin{cases} \prod_{p=1}^{i-j}(T_{i+1-p}^{DH}/T_{i-p}^D)S_{i+1-p,i-p}^1[H_\nu^{(1)}(k_jr) + R_j^{Dr}H_\nu^{(2)}(k_jr)], \\ \qquad j = i-1, \cdots, 0 \\ H_\nu^{(1)}(k_ir) + R_i^{Dr}H_\nu^{(2)}(k_ir) \\ \prod_{p=1}^{j-i}(T_{i+p-1}^D/T_{i+p}^{DH})S_{i+p-1,i+p}^1[H_\nu^{(1)}(k_jr) + R_j^{Dr}H_\nu^{(2)}(k_jr)], \\ \qquad j = i+1, \cdots, m. \end{cases} \quad (2.5)$$

The order $\nu = \nu_n$ of the Hankel functions $H_\nu^{(1)}(kr)$ and $H_\nu^{(2)}(kr)$ are solutions of the modal equation

$$[(1/R_i^{Dr}) - R_i^{Ur}] = 0, \quad i = 0, 1 \cdots \text{ or } m, \quad (2.6)$$

in which

$$R_i^{Dr} = R_i^D H_\nu^{(1)}(k_i r_{i,i+1})/H_\nu^{(2)}(k_i r_{i,i+1}),$$

$$R_m^{Dr} = -H_\nu^{(1)\prime}(k_m a)/H_\nu^{(2)\prime}(k_m a), \quad (2.7a)$$

$$R_i^{Ur} = R_i^U H_\nu^{(2)}(k_i r_{i-1,i})/H_\nu^{(1)}(k_i r_{i-1,i}), \quad R_0^{Ur} = 0, \quad (2.7b)$$

and

$$R_i^{DH} = R_i^{Dr} H_\nu^{(2)}(k_i r_{i-1,i})/H_\nu^{(1)}(k_i r_{i-1,i}),$$

$$R_i^{UH} = R_i^{Ur} H_\nu^{(1)}(k_i r_{i,i+1})/H_\nu^{(2)}(k_i r_{i,i+1}). \quad (2.7c)$$

The reflection coefficient at $r = r_{i,i+1}$ looking in the $-r$ direction is R_i^D and the reflection coefficient at $r = r_{i-1,i}$ looking in the $+r$ direction is R_i^U. The explicit expressions for these coefficients are given in the companion paper.[3] The transmission coefficients are

$$T_i^D = 1 + R_i^D \quad \text{and} \quad T_i^{DH} = 1 + R_i^{DH} \quad (2.8)$$

and

$$S_{j-1,j}^1 = H_\nu^{(1)}(k_{j-1}r_{j-1,j})/H_\nu^{(1)}(k_j r_{j-1,j}) = 1/S_{j,j-1}^1. \quad (2.9)$$

For $|k_m(r_{m-1,m} - a)| \gg 1$, $R_m^{DH} \to 0$, thus for $r \geqslant r_{m-1,m}$ the expressions for the electromagnetic fields do not depend upon the particular value of a.[3] The product of the normalization coefficients M_ν and N_ν is

$$M_\nu N_\nu = \frac{i}{2\pi}\partial D(\nu)/\partial\nu, \quad (2.10a)$$

in which

$$D(\nu) = 4[1/R_i^{Dr} - R_i^{Ur}]/\omega\epsilon_i, \quad i = 0, 1, 2 \cdots \text{ or } m. \quad (2.10b)$$

The transverse wave impedance is

$$Z_\nu(r) = \nu/\omega\epsilon r. \quad (2.11)$$

In view of the normalization used in (2.3) and (2.4), the forward and backward wave amplitudes $a(\nu, \phi)$ and $b(\nu, \phi)$, respectively, are expressed as follows in terms of the magnetic and electric field transforms:

$$H(\nu, \phi) = a(\nu, \phi) + b(\nu, \phi), \quad (2.12a)$$

$$E(\nu, \phi) = a(\nu, \phi) - b(\nu, \phi). \quad (2.12b)$$

For uniform cylindrical structures the basis functions ψ_ν are functions of r only; however, for nonuniform structures, ψ_ν are also implicitly functions of ϕ through the parameters μ_i, ϵ_i and $r_{i-1,i}$.

3. THE COUPLED DIFFERENTIAL EQUATIONS FOR THE WAVE AMPLITUDES

The partial differential equations (2.1) for the transverse components of the electromagnetic fields in conjunction with the boundary conditions (2.2a) and (2.2e) are converted into ordinary differential equations for the wave amplitudes $a(\nu, \phi)$ and $b(\nu, \phi)$. To this end, we express the Dirac delta function $\delta(r - r_0)$ in terms of the transforms (2.3). Thus,

$$\delta(r - r_0) = \sum_{n=1}^\infty \psi_\nu(r)\psi_\nu(r_0)Z_\nu(r)/M_\nu N_\nu, \quad \nu = \nu_n. \quad (3.1a)$$

The orthonormal relationship between the basis functions $\psi_\nu(r)$ and $\psi_\mu(r)$ corresponding to the roots $\nu=\nu_n$ and $\mu=\nu_m$ of the modal equation is

$$\delta_{nm}=\int_a^\infty \psi_\mu(r)\psi_\nu(r)Z_\nu(r)\,\frac{dr}{M_\mu N_\nu}=\left[\frac{\nu r/\omega\epsilon}{M_\mu N_\nu(\mu^2-\nu^2)}\right.$$

$$\times\left(\psi_\mu\,\frac{\partial\psi_\nu}{\partial r}-\psi_\nu\,\frac{\partial\psi_\mu}{\partial r}\right)_a^\infty, \qquad (3.\,1\text{b})$$

where δ_{nm} is the Kronecker delta and since ψ_ν is piecewise continuous, the integration in each layer of the structure is carried out separately. Since at each interface, $r=r_{i-1,i}$, ψ_ν and $(1/\omega\epsilon)\partial\psi_\nu/\partial r$ are continuous, the basis functions satisfy the appropriate boundary conditions, (2. 2), only when $r_{i,i-1}$, ϵ_i, and μ_i are not functions of ϕ for all i. Thus, for nonuniform cylindrical structures, the individual terms of the expansions (2. 3) and (2. 4) do not satisfy the boundary conditions (2. 2). Hence, it is not permissible, in general, to interchange orders of differentiation and summation whenever it is necessary to differentiate (2. 3) and (2. 4).[2] Multiply (2. 1a) by $\psi_\nu(r)Z_\nu(r)dr/M_\nu$ and integrate with respect to r over the interval (a,∞). Thus, on noting that

$$\int_a^\infty \frac{\partial H_z}{\partial\phi}\frac{\psi_\nu Z_\nu}{M_\nu}\,dr=\frac{\partial}{\partial\phi}\int_a^\infty H_z\frac{\psi_\nu Z_\nu}{M_\nu}\,dr$$

$$-\int_a^\infty H_z\frac{\partial}{\partial\phi}\left(\frac{\psi_\nu Z_\nu}{M_\nu}\right)dr$$

$$+\sum_{i=1}^m\frac{d}{d\phi}r_{i-1,i}\;H_z\left(\frac{\psi_\nu Z_\nu}{M_\nu}\right)_{r_{i-1,i}^-}^{r_{i-1,i}^+} \qquad (3.\,2)$$

and using the orthonormal relationship (3. 1b), we get

$$-\frac{dH(\nu,\phi)}{d\phi}=i\nu E(\nu,\phi)-\sum_{p=1}^\infty C(\nu,\mu)H(\mu,\phi),\quad \mu=\mu_p,$$

$$p=1,2,3\cdots, \qquad (3.\,3)$$

in which

$$C(\nu,\mu)=\int_a^\infty\frac{\partial}{\partial\phi}\left(\frac{\psi_\nu Z_\nu}{M_\nu}\right)\frac{\psi_\mu}{N_\mu}\,dr-\sum_{i=1}^m\left(\frac{\psi_\nu Z_\nu\psi_\mu}{M_\nu N_\mu}\right)_{r_{i-1,i}^-}^{r_{i-1,i}^+}$$

$$\times\frac{d}{d\phi}r_{i-1,i}. \qquad (3.\,4)$$

Similarly, multiply (2. 1b) by $\psi_\nu dr/M_\nu$ and integrate over the interval $a\leqslant r<\infty$. Using Green's theorem in one dimension, it can be shown that

$$\int_a^\infty\frac{1}{i\omega\epsilon}\frac{\partial}{\partial r}\left(r\,\frac{\partial H_z}{\partial r}\right)\frac{\psi_\nu}{M_\nu}\,dr=\int_a^\infty\frac{1}{i\omega\epsilon}\frac{H_z}{M_\nu}\frac{\partial}{\partial r}\left(r\,\frac{\partial\psi_\nu}{\partial r}\right)dr$$

$$-\sum_{i=1}^m\left[\frac{r}{i\omega\epsilon}\left(\frac{\partial H_z}{\partial r}\frac{\psi_\nu}{M_\nu}-\frac{\partial\psi_\nu}{\partial r}\frac{H_z}{M_\nu}\right)\right]_{r_{i-1,i}^-}^{r_{i-1,i}^+}. \qquad (3.\,5)$$

The last term in (3. 5) vanishes since H_z and $-(1/\omega\epsilon)\partial\psi_\nu/\partial r$ are continuous at each interface $r=r_{i-1,i}$. Thus, on employing the orthonormal relationship (3. 1b) and the boundary condition (2. 2e), it follows that

$$-\frac{dE(\nu,\phi)}{d\phi}=i\nu H(\nu,\phi)-\sum_{p=1}^\infty D(\nu,\mu)E(\mu,\phi)+J(\nu,\phi),$$

$$\mu=\nu_p,\quad p=1,2,3\cdots, \qquad (3.\,6)$$

in which

$$D(\nu,\mu)=\int_a^\infty\frac{\partial}{\partial\phi}\left(\frac{\psi_\nu}{M_\nu}\right)\frac{\psi_\mu Z_\mu}{N_\mu}\,dr \qquad (3.\,7)$$

and

$$J(\nu,\phi)=K\psi_\nu(r_0)\delta(\phi-\phi_0)/M_\nu. \qquad (3.\,8)$$

Expressing the field transforms $H(\nu,\phi)$ and $E(\nu,\phi)$ in terms of the wave amplitudes $a(\nu,\phi)$ and $b(\nu,\phi)$ [(2. 12)] we obtain from (3. 3) and (3. 6) the ordinary coupled differential equations for the wave amplitudes

$$-\frac{da(\nu,\phi)}{d\phi}-i\nu a(\nu,\phi)$$

$$=\sum_{p=1}^\infty S_{\nu\mu}^{BA}\,a(\mu,\phi)+S_{\nu\mu}^{BB}\,b(\mu,\phi)+J(\nu,\phi)/2 \qquad (3.\,9\text{a})$$

and

$$-\frac{db(\nu,\phi)}{d\phi}+i\nu b(\nu,\phi)=\sum_{p=1}^\infty S_{\nu\mu}^{AB}\,b(\mu,\phi)+S_{\nu\mu}^{AA}\,a(\mu,\phi)$$

$$-J(\nu,\phi)/2,\quad \mu=\nu_p,\ p=1,2,3\cdots,$$

$$(3.\,9\text{b})$$

in which the transmission scattering coefficients are

$$S_{\nu\mu}^{BA}=S_{\nu\mu}^{AB}=-[C(\nu,\mu)+D(\nu,\mu)]/2 \qquad (3.\,9\text{c})$$

and the reflection scattering coefficients are

$$S_{\nu\mu}^{AA}=S_{\nu\mu}^{BB}=-[C(\nu,\mu)-D(\nu,\mu)]/2. \qquad (3.\,9\text{d})$$

4. THE SCATTERING COEFFICIENTS AND THE RECIPROCITY RELATIONSHIPS

Before obtaining explicit expressions for the scattering coefficients, we determine the relationship between the coupling coefficients $C(\nu,\mu)$ and $D(\nu,\mu)$. Differentiating (3. 1b) with respect to ϕ for $\nu=\nu_n$ and $\mu=\nu_m$, we get

$$C(\nu,\mu)+\frac{M_\mu N_\nu}{N_\mu M_\nu}D(\mu,\nu)+\delta_{n,m}\frac{d}{d\phi}\ln\left(\frac{M_\mu}{N_\mu}\right)=0. \qquad (4.\,1)$$

Hence, it is only necessary to evaluate the expression for $D(\nu,\mu)$ to determine the scattering coefficients. Furthermore, it follows that

$$\frac{M_\nu}{N_\nu}S_{\nu\mu}^{\alpha\alpha}=\frac{M_\mu}{N_\mu}S_{\mu\nu}^{\alpha\alpha}\quad\text{for}\quad\alpha=A\text{ or }B \qquad (4.\,2\text{a})$$

and

$$\frac{M_\nu}{N_\nu}S_{\nu\mu}^{BA}=-\frac{M_\mu}{N_\mu}S_{\mu\nu}^{AB}. \qquad (4.\,2\text{b})$$

In view of the normalization adopted in this paper, (4. 2) satisfies the reciprocity relationships in electromagnetic theory.[3,4]

The choice $M_\nu/N_\nu=1$ for the normalization coefficients simplifies (4. 1) and the reciprocity relationships (4. 2); however, at times it is more convenient to choose M_ν/N_ν in some other manner.[3] Noting the relationship between (3. 1b) and the expression for $D(\nu,\mu)$, (3. 7), it follows that for $\nu\neq\mu$,

$$D(\nu, \mu) = \frac{1}{M_\nu N_\mu} \sum_{i=1}^{m} \frac{1}{\mu^2 - \nu^2} \left[\left(\frac{\mu r}{\omega\epsilon}\right)\left(\psi_\mu \frac{\partial^2 \psi_\nu}{\partial\phi\partial r} - \frac{\partial\psi_\nu}{\partial\phi}\frac{\partial\psi_\mu}{\partial r}\right)\right]_{r_{i-1,i}^-}^{r_{i-1,i}^+}. \tag{4.3}$$

On differentiating (3.1b) with respect to ϕ for $\nu = \mu$, we get

$$2D(\nu,\nu) + \frac{d}{d\phi}\ln\left(\frac{M_\nu}{N_\nu}\right) + \int_a^\infty \frac{d}{d\phi}[\ln(\nu/\omega\epsilon)]\frac{\psi_\nu\psi_\nu Z_\nu}{N_\nu M_\nu}dr - \sum_{i=1}^{m}\left(\frac{dr_{i-1,i}}{d\phi}\frac{\psi_\nu\psi_\nu Z_\nu}{N_\nu M_\nu}\right)_{r_{i-1,i}^-}^{r_{i-1,i}^+} = 0, \tag{4.4a}$$

where[5]

$$\int \frac{d}{d\phi}[\ln(\nu/\omega\epsilon)]\frac{\psi_\nu\psi_\nu Z_\nu}{N_\nu M_\nu}dr = \frac{d}{d\phi}[\ln(\nu/\omega\epsilon)]\frac{(r/\omega\epsilon)}{2N_\nu M_\nu} \times\left(\psi_\mu\frac{\partial^2\psi_\nu}{\partial\nu\partial r} - \frac{\partial\psi_\nu}{\partial\nu}\frac{\partial\psi_\mu}{\partial r}\right). \tag{4.4b}$$

It is interesting to note that the expressions for the scattering coefficients (3.9) may also be derived by imposing the condition that $H_z(r,\phi)$ and $E_r(r,\phi)$ are continuous at planes $\phi = \text{const}$ separating two cylindrical structures whose electromagnetic parameters are $\epsilon_i(\phi)$, $\mu_i(\phi)$, and $\epsilon_i(\phi + \Delta\phi)$, $\mu_i(\phi+\Delta\phi)$, respectively, and with concentric boundaries at $r_{i-1,i}$ and $r_{i-1,i} + (d/d\phi)\overline{r_{i-1,i}}\,\Delta\phi$.

The general expression for $\partial\psi_\nu/\partial\phi$ in (4.3) is

$$\frac{\partial\psi_\nu}{\partial\phi} = \sum_{i=1}^{m}\frac{\partial\psi_\nu}{\partial r_{i-1,i}}\frac{dr_{i-1,i}}{d\phi} + \sum_{i=0}^{m}\left(\frac{\partial\psi_\nu}{\partial\epsilon_i}\frac{d\epsilon_i}{d\phi} + \frac{\partial\psi_\nu}{\partial\mu_i}\frac{d\mu_i}{d\phi}\right). \tag{4.5}$$

A similar expression can be written for $\partial^2\psi_\nu/\partial\phi\partial r$. Thus, as expected, the scattering coefficients are also independent of a, when medium m is a good conductor.

5. SOLUTIONS FOR THE WAVE AMPLITUDES

The forward and backward wave amplitudes $a(\nu,\phi)$ and $b(\nu,\phi)$, respectively, are solutions of the inhomogeneous first order ordinary differential equations (3.9). The effect of the source term $J(\nu,\phi)/2$ is to produce a jump in the value of $a(\nu,\phi)$ and $b(\nu,\phi)$ at $\phi=\phi_0$. On integrating (3.9) with respect to ϕ over the infinitesimal range $\phi_0^- \leq \phi \leq \phi_0^+$, we get

$$a(\nu,\phi_0^+) - a(\nu,\phi_0^-) = -K\psi_\nu(r_0)/2M_\nu \equiv -J_{\nu 0} \tag{5.1a}$$

and

$$b(\nu,\phi_0^+) - b(\nu,\phi_0^-) = K\psi_\nu(r_0)/2M_\nu \equiv J_{\nu 0}. \tag{5.1b}$$

Thus, to evaluate $a(\nu,\phi)$ and $b(\nu,\phi)$ it is convenient to set $J(\nu,\phi)=0$ in (3.9) and solve the resulting homogeneous differential equations in conjunction with (5.1). The Runge–Kutta method can be used to solve these equations numerically.[6]

We consider here iterative solutions that are very suitable when the power coupled into spurious modes is small compared to the power associated with the incident mode. In this case, the first order or WKB type solution to (3.9) is obtained by setting $S_\nu^{AA} = S_{\nu\mu}^{BB} = 0$ for all ν and μ and $S_{\nu\mu}^{AB} = S_{\nu\mu}^{BA} = 0$ for $\nu \neq \mu$. Using (4.1), it follows that

$$S_{\nu\nu}^{AB} = S_{\nu\nu}^{BA} = \frac{d}{d\phi}\ln\left(\frac{N_\nu}{M_\nu}\right)^{1/2}. \tag{5.2}$$

Thus we get for $0 < \phi - \phi_0 < 2\pi$

$$a^1(\nu,\phi) = iJ_{\nu 0}A_\nu(\phi,\phi_0)\exp[-i(\textstyle\int_{\phi_0}^{\phi}\nu(\phi')d\phi' - \theta)]/2\sin\theta \tag{5.3a}$$

and

$$b^1(\nu,\phi) = iJ_{\nu 0}B_\nu(\phi,\phi_0)\exp[i(\textstyle\int_{\phi_0}^{\phi}\nu(\phi')d\phi' - \theta)]/2\sin\theta \tag{5.3b}$$

in which

$$\theta = \frac{1}{2}\int_0^{2\pi}\nu(\phi')d\phi' \tag{5.3c}$$

and

$$A_\nu(\phi,\phi_0) = B_\nu(\phi,\phi_0) = \left(\frac{M_\nu(\phi_0)N_\nu(\phi)}{M_\nu(\phi)N_\nu(\phi_0)}\right)^{1/2}. \tag{5.3d}$$

The integral $\int_{\phi_0}^{\phi}\nu(\phi')d\phi'$ constitutes the familiar phase memory concept in slowly varying media and (5.3d) is consistent with power conservation. If the normalization coefficients M_ν and N_ν are chosen such that $M_\nu = N_\nu$, $A_\nu(\phi,\phi_0) = B_\nu(\phi,\phi_0) = 1$. For uniform structures, the phase memory integral is $\nu(\phi-\phi_0)$, $\theta = \nu\pi$, and $A_\nu(\phi,\phi_0) = B_\nu(\phi,\phi_0) = 1$.[3] For $kr \gg 1$ and $kr_0 \gg 1$, we retain only the direct waves between the source and the observation point and ignore the contributions from the waves that creep around the cylindrical structure p times, $p = 1, 2, 3\cdots$.[7] Thus,

$$a^1(\nu,\phi) \approx -J_{\nu 0}A_\nu(\phi,\phi_0)\exp\left[-i\int_{\phi_0}^{\phi}\nu(\phi')d\phi'\right]U(\phi-\phi_0) \tag{5.4a}$$

and

$$b^1(\nu,\phi) \approx -J_{\nu 0}B_\nu(\phi,\phi_0)\exp\left[i\int_{\phi_0}^{\phi}\nu(\phi')d\phi'\right]U(\phi_0-\phi) \tag{5.4b}$$

in which $U(\phi-\phi_0)$ is the unit step function.

The second order iterative solutions are obtained on substituting (5.3) or (5.4) for the wave amplitudes in the terms in the right-hand side of (3.9). Thus,

$$a^2(\nu,\phi) = G_\nu a^1(\nu,\phi) - U(\phi-\phi_0)\int_{\phi_0}^{\phi}A_\nu(\phi,\phi')\times\exp\left[-i\int_{\phi'}^{\phi}\nu(\phi'')d\phi''\right]f_1(\phi')d\phi', \tag{5.5a}$$

in which

$$f_1(\phi) = \sum_{\substack{p=1\\p\neq n}}^{\infty}S_{\nu\mu}^{BA}a(\mu,\phi) + \sum_{p=1}^{\infty}S_{\nu\mu}^{BB}b(\mu,\phi), \quad \nu=\nu_n, \quad \mu=\mu_p. \tag{5.5b}$$

The constant G_ν in (5.5a) is determined by imposing the jump condition (5.1a). When only contributions for direct waves are considered, $G_\nu = 1$ and $a^1(\nu,\phi)$ is given by (5.4a). Expressions similar to (5.5b) may be written for the backward wave amplitudes $b(\nu,\phi)$.

6. CONCLUDING REMARKS

The problem of propagation of radio waves excited by line sources in nonuniform cylindrical structures has been reduced to the solution of ordinary first-order differential equations for the forward and backward wave amplitudes. These full wave solutions which satisfy exact boundary conditions at each interface of the multilayered structure are shown to satisfy the reciprocity relationships in electromagnetic theory.

For the purpose of the analysis, generalized field transforms are employed.[3] These transforms provide complete expansions for the electromagnetic fields in terms of discrete sets of modes that are very suitable for numerical evaluation when the distance between the transmitter and receiver is large compared to a wave length.[8,9] When the separation between the transmitter and receiver is small, and the effects of curvature are negligible, it is more convenient to use a parallel stratified model to analyze the problem.[2] The relationship between the generalized field transforms for cylindrical structures and the Fourier-type transforms for parallel stratified structures has been established.

Since it is assumed in the anlaysis that the electromagnetic parameters as well as the thickness of the layers of the structure vary along the propagation path, the solutions have wide applications to problems of propagation in both open and closed irregular guiding structures.

*The research reported in this paper was supported by the National Science Foundation and the Engineering Research Center of the University of Nebraska. The author wishes to thank W. L. Brogan and F. G. Ullman for their comments. The manuscript was typed by Mrs. E. Everett.

[1] A. Tran and C. Polk, Abstracts of URSI 1972 Fall Meeting, Williamsburg, Virginia, p. 50, (1972).

[2] E. Bahar, J. Math. Phys. 14, 1030 (1973).

[3] E. Bahar, J. Math. Phys. 15, 1977 (1974).

[4] D. M. Kerns and R. W. Beatty, *Basic Theory of Waveguide Junctions and Introductory Microwave Network Analysis* (Pergamon, New York, 1967).

[5] A. Erdelyi, W. Magnus, F. Oberhittinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. 2.

[6] M. Abromowitz and I. A. Stegum, *Handbook of Mathematical Functions* , with formulas, graphs and mathematical tables, Department of Commerce, Natl. Bur. Stand., Applied Mathematics Series 55 (Washington, D.C., 1964).

[7] E. Bahar, J. Math. Phys. 12, 186 (1971).

[8] E. Bahar and G. Govindarajan, J. Geophys. Res. 78 (2) 393 (1973).

[9] E. Bahar and G. Govindarajan, IEEE Trans Microwave Theory and Techniques, **MTT-21**, (12) 819 (1973).

On the spectrum of the linear transport operator*

Edward W. Larsen

Courant Institute of Mathematical Sciences, New York University, New York, New York 10012

Paul F. Zweifel

Department of Physics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

(Received 25 February 1974)

In this paper, spectral properties of the time-independent linear transport operator A are studied. This operator is defined in its natural Banach space $L_1(D \times V)$, where D is the bounded space domain and V is the velocity domain. The collision operator K accounts for elastic and inelastic slowing down, fission, and low energy elastic and inelastic scattering. The various cross sections in K and the total cross section are piecewise continuous functions of position and speed. The two cases $v_0 > 0$ and $v_0 = 0$ are treated, where v_0 is the minimum neutron speed. For $v_0 = 0$, it is shown that $\sigma(A)$ consists of a full half-plane plus, in an adjoining strip, point eigenvalues and curves. For $v_0 > 0$, $\sigma(A)$ consists just of point eigenvalues and curves in a certain half-space. In both cases, the curves are due to purely elastic "Bragg" scattering and are absent if this scattering does not occur. Finally the spectral differences between the two cases $v_0 > 0$ and $v_0 = 0$ are discussed briefly, and it is proved that A is the infinitesimal generator of a strongly continuous semigroup of operators.

I. INTRODUCTION

Since the pioneering work of Lehner and Wing,[1] the spectrum of the linear or "neutron" transport operator has been the subject of intensive study by mathematicians, physicists, and nuclear engineers.[2-37] Knowledge of this spectrum is, as has been stressed in Ref. 1, a necessary prerequisite to the calculation of eigenfunction expansions for time-dependent problems. (In some cases, due to the existence of a half plane of "continuum" spectrum, analysis has shown that such expansions are not feasible.) In addition, knowledge of the spectrum is important in the interpretation of pulsed neutron experiments.

It would be a formidable task to summarize the research embodied in the references (which listing should not be considered all-inclusive, incidentally, but only representative). We would like to make a few general comments, however, about their contents in order to motivate our own work.

A great deal of the above work involves specific models of the transport operator (i.e., of the collision term and the geometry). Thus Ref. 1 treats a one-speed, one-dimensional equation with isotropic scattering, while Ref. 12, for another example, deals with a three-dimensional ideal gas scattering model. As the years progressed, various authors attempted to treat increasingly general problems. For example, we find Mika[14] generalizing the results of Ref. 1 to anisotropic and energy dependent operators, still in one space dimension. As another example, Bednarz[13] obtained some fairly general results but specifically excluded purely elastic scattering at thermal neutron energies. An attempt was made to remove this restriction[17,27] but these results depend upon the assumptions that the minimum neutron speed v is zero, and $v\Sigma_0(v) < v\Sigma(v) - \lambda^*$. Here Σ_0 and Σ are the elastic and total cross sections, and λ^* is the minimum value of $v\Sigma(v)$, which is required to be at $v = 0$. Some additional recent work[36] has treated elastic scattering involving a discontinuity in the total reaction rate (as a function of neutron speed), but it contains a technical error in the proof of Lemma 2.[38]

Most of the cited references employ a spectral analysis in Hilbert space similar to that in Ref. 1. However, other approaches have been used, two of which we shall mention here.

First, we refer to a paper by Jörgens,[2] where semigroup techniques are used to treat the case of a finite body in which the minimum neutron speed v_0 is positive. For this case Jörgens showed that the spectrum of the transport operator consists solely of isolated point spectrum. (Other authors, considering the case $v_0 = 0$, have asserted that the spectrum should contain a half-space $\mathrm{Re}\lambda \leq -\lambda^*$.)

Also, we refer to the homogeneous, infinite medium problems which have been treated[31] by taking a Fourier transform in the space variable and a Laplace transform in time. One obtains a problem involving both time and space eigenvalues in which a dispersion law, i.e., a functional relation between the two quantities, is sought.

The above brief overview of the previous work suggests the motivations for the present paper.

First, we employ a more general collision operator than has been considered previously. It consists of three terms: (i) a completely continuous portion, representing elastic slowing down, fission, and low energy inelastic scattering; (ii) a singular nilpotent portion, representing inelastic slowing down; and (iii) a singular portion, representing low energy elastic "Bragg" scattering.[39]

Secondly, our work is carried out in an L_1 space whereas nearly all previous work has been performed in L_2. We do this because L_1 is the natural space for the transport operator, since the integral of the angular density has physical significance. (In quantum mechanics, where $|\psi|^2$ carries physical meaning, L_2 is appropriate.) This point, incidentally, has been stressed by Case (private communications) and Ribaric,[37] who gives a lengthy discussion. See also Refs. 33 and 35.

Thirdly, we treat the two separate cases $v_0 = 0$ and $v_0 > 0$ where v_0 is the minimum neutron speed. (Thus our work for $v_0 > 0$ will generalize Jörgens' results.[2]) The differences in the spectrum for these cases is substantial, and we shall briefly discuss these differences in Sec. 6.

Finally, we feel that our methods have an advantage of being systematic. Subsequent generalizations, for example to problems of gas dynamics, should thus be made simpler.

The plan of this paper is as follows. In Sec. 2 we describe the transport operator and its domain, and we introduce various restrictions which we find necessary to impose on the collision terms mentioned above.

Section 3 is devoted to studying the spectrum of the streaming operator T, i.e., the transport operator minus collision terms. The results are embodied in Theorems 1 and 2, which state that $\sigma(T)$ (Ref. 40) consists of a half plane or only the point at infinity, depending respectively upon whether the minimum neutron speed v_0 is or is not zero. In Sec. 4 we study the one-speed transport operator denoted by $T+K_0$; i.e., K_0 is a collision operator which does not change the neutron speed. The analysis of this section depends heavily on the results of Sec. 3 and uses the concept of potentially compact operators[41] and a theorem of Gohberg,[42] sometimes called the "Smul'yan" theorem. The first conclusions (Theorem 3) are that the spectrum of $T+K_0$, for fixed v, is a pure isolated point spectrum of finite geometrical multiplicity restricted to a certain left half-space. (These results are also proved in Ref. 2. The lack of "continuum" spectrum is due to our considering a finite body; for an infinite slab, for example, the one-speed operator does have a "continuum" spectrum.[5]) Next (Theorem 5) we consider the "full" spectrum of $T+K_0$ by considering all admissable values of v. As v varies, the (point) spectrum of $T+K_0$ for fixed v shifts about to form curves; the full spectrum of $T+K_0$ consists of the closure of this set of curves plus $\sigma(T)$. We also prove (Theorem 6) that for $v_0=0$ and a sufficiently small body, all the spectrum is contained in the continuum, a result which has been argued heuristically[7,10] and proved for certain models.[12,13]

All of these results are, of course, of greater or lesser importance depending upon how meaningfully one takes a one-speed model of the transport operator. However, we use these results in Sec. 5, in which the total transport operator $T+K$ is considered and we prove that $\sigma(T+K)$ differs from $\sigma(T+K_0)$ only by the addition of point spectrum. Also, we make some estimates as to the location of this spectrum and we show that the low energy elastic scattering term K_0 can introduce lines of spectra.

Finally, we show, in Theorem 11, that the transport operator is the infinitesimal generator of a strongly continuous semigroup of operators. This theorem, in a sense, justifies this entire paper since it guarantees the existence of a semigroup which solves the initial value problem for the transport operator.

In Sec. 6 we discuss our results and indicate the direction in which future work might be aimed. We conclude with an Appendix to which some of the technical details of the proofs have been relegated.

II. DESCRIPTION OF THE TRANSPORT OEPRATOR

We take D to be an open, bounded, connected set of points $\mathbf{r}$ in three-dimensional configuration space. (If D is not convex then we require that neutrons emitted out of D be absorbed, so that the problem of reentering neutrons does not arise.) We take V to be the three-dimensional velocity space consisting of velocities $\mathbf{v}=v\mathbf{\Omega}$, with $|\mathbf{\Omega}|=1$ and $v_0\leqslant v\leqslant v_1<c$, where c is the speed of light. (Since we are dealing with a nonrelativistic equation, we require $v_1\ll c$. In a nuclear reactor, where the maximum neutron energy is about one percent of the neutron rest mass, this condition is certainly fulfilled.)

We define X as the Banach space of complex-valued, measurable functions $\psi(\mathbf{r},\mathbf{v})$, defined on $\bar{D}\times\bar{V}$, satisfying

$$\|\psi\|=\int_{\mathbf{r}\in D}\int_{\mathbf{v}\in V}|\psi(\mathbf{r},\mathbf{v})|\,d\mathbf{r}d\mathbf{v}<\infty.$$

Now we shall describe the transport operator A and its (dense) domain $X_0\subset X$.

We write A as the sum $A=T+K$ where T is the "streaming" operator and K is the "scattering" or "collision" operator.

The operator T is defined by[44]

$$(T\psi)(\mathbf{r},\mathbf{v})=-[\mathbf{v}\cdot\nabla+v\Sigma(\mathbf{r},v)]\psi(\mathbf{r},\mathbf{v}). \tag{2.1}$$

Here the gradient operator ∇ acts only on $\mathbf{r}$ and $v\Sigma(\mathbf{r},v)$ satisfies the following properties:

(a) $v\Sigma(\mathbf{r},v)$ is nonnegative, bounded, and piecewise continuous in $\mathbf{r}$ and v.

(b) If $v_0=0$, then

$$\operatorname*{ess\,inf}_{\mathbf{r}\in D}\lim_{v\to 0}v\Sigma(\mathbf{r},v)\equiv\lambda^* \tag{2.2}$$

exists, and

$$\frac{v\Sigma(\mathbf{r},v)-\lambda^*}{v}\geqslant -c_0,\quad \mathbf{r}\in\bar{D},\ 0<v\leqslant v_1, \tag{2.3}$$

where c_0 is a nonnegative constant. [If $v\Sigma(\mathbf{r},v)$ is monotone increasing in v for each $\mathbf{r}$, then c_0 automatically exists and can be taken to be 0.]

(c) For $v_0\geqslant 0$, we define the constant $\hat{\lambda}(v_0)$ by

$$\hat{\lambda}(v_0)=\operatorname*{ess\,inf}_{\substack{\mathbf{r}\in D\\ v_0\leqslant v\leqslant v_1}} v\Sigma(\mathbf{r},v). \tag{2.4}$$

We note that $\hat{\lambda}(0)\leqslant\lambda^*$, and that equality holds if $v\Sigma(\mathbf{r},v)$ is monotone increasing in v.[45]

Now we define X_0 to be the (dense) subspace of functions ψ such that $\psi(\mathbf{r},\mathbf{v})=0$ for $\mathbf{r}\in\partial D$ and $\mathbf{v}$ pointing into D, and $T\psi\in X$. T is a closed operator on X_0. Since K will be bounded on X, then $A=T+K$ will be a closed operator on X_0.

Next we discuss the scattering operator K. To do so, we write it as the sum

$$K=K_c+K_d+K_0, \tag{2.5}$$

where K_c, K_d, and K_0 are all bounded operators in X.

K_c represents "continuum" scattering and is described by an integral operator:

$$(K_c\psi)(\mathbf{r},\mathbf{v}) = \int_{\mathbf{v}' \in V} k_c(\mathbf{r},\mathbf{v}' \to \mathbf{v})\psi(\mathbf{r},\mathbf{v}')d\mathbf{v}'. \qquad (2.6)$$

The kernel k_c satisfies:

(d) k_c is nonnegative and piecewise continuous. Also, k_c is bounded except possibly for the case $v_0 = 0$, in which we allow

$$0 \le k_c(\mathbf{r},\mathbf{v}' \to \mathbf{v}) \le M_c/v^2. \qquad (2.7)$$

This "continuum" scattering corresponds physically to fission, high energy elastic slowing down, and thermal inelastic scattering.

K_d represents high energy inelastic scattering and is described by a "downshift" operator of the form

$$(K_d\psi)(\mathbf{r},\mathbf{v}) = \sum_{m=1}^{M_0} (K_d^{(m)}\psi)(\mathbf{r},\mathbf{v}),$$

$$(K_d^{(m)}\psi)(\mathbf{r},\mathbf{v}) = \int_{\mathbf{v}' \in V} k_d^{(m)}(\mathbf{r}, v', \mathbf{\Omega}' \to \mathbf{\Omega})$$

$$\times \delta[v' - \omega_n(v)]\psi(\mathbf{r},\mathbf{v}')d\mathbf{v}'. \qquad (2.8)$$

Here the operator $K_d^{(m)}$ describes an event in which a discrete energy E_m is lost by a neutron at $\mathbf{r}$ with initial speed $\omega_m(v)$ and final speed v. $\omega_m(v)$ is defined by

$$E_m = \tfrac{1}{2}N\omega_m^2(v) - \tfrac{1}{2}Nv^2,$$

where N is the mass of a neutron. The kernels $k_d^{(m)}$ satisfy:

(e) $k_d^{(m)}$ are nonnegative, piecewise continuous, and bounded:

$$0 \le k_d^{(m)}(\mathbf{r}, v', \mathbf{\Omega}' \to \mathbf{\Omega}) \le M_d^{(m)}. \qquad (2.9)$$

(f) There exists a threshold speed v_t such that, for all m, $k_d^{(m)}(\mathbf{r},\mathbf{v}' \to \mathbf{v}) = 0$ for $v' < v_t$. (v_t is the threshold speed below which the high-energy inelastic scattering described by K_d cannot occur.)

It follows from the above description of K_d and from our assumption that neutron speeds are bounded above that K_d is nilpotent, i.e., there exists an integer M_1 such that

$$K_d^{M_1} = 0. \qquad (2.10)$$

Physically, this means that after a maximum of $M_1 - 1$ consecutive high energy inelastic collisions, a neutron must have speed below v_t.

Finally, the operator K_0 in (2.5) is a "Bragg" scattering or one-speed operator for low energy neutrons described by

$$(K_0\psi)(\mathbf{r},\mathbf{v}) = \int_{\mathbf{v}'} k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega})\delta(v' \to v)\psi(\mathbf{r},\mathbf{v}')d\mathbf{v}'. \qquad (2.11)$$

The kernel k_0 satisfies:

(g) k_0 is nonnegative, piecewise continuous, and bounded except possibly for the case $v_0 = 0$, in which we allow

$$0 \le k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega}) \le M_0/v^2. \qquad (2.12)$$

(h) $k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega}) = 0$ for $v > v_t$.

We note that (f) and (h) imply

$$K_d(\lambda I - T)^{-1}K_0 = 0, \quad \lambda \in \rho(T). \qquad (2.13)$$

To end this section, we shall make some comments regarding the above assumptions.

First, the inequalities (2.7), (2.9), and (2.12) imply that $K = K_0 + K_c + K_d$ is a bounded operator. This means physically that for each neutron density ψ, the corresponding total rate of secondary neutron production $K\psi$ is uniformly bounded: $\|K\psi\| \le \|K\|\,\|\psi\|$.

Next, we note that the various kernels and cross sections have been assumed piecewise continuous. Physically, discontinuities in $\mathbf{r}$ correspond to boundaries between regions with different constituents while discontinuities in v correspond to threshold effects, either the Bragg scattering "cutoff" or the inelastic scattering threshold. To prove our results, we shall assume that "piecewise continuous" has one of the following two meanings:

(i) The various kernels and cross sections are continuous in all of their variables. (This corresponds to a body in which the constituents vary continuously with position, and no threshold effects occur in speed.)

(j) $\bar{D} = \cup_{n=1}^{M_2} \bar{D}_n$, where D_n are open sets. In D_n, the various kernels and cross sections are constant functions of $\mathbf{r}$, $k_c(\mathbf{r},\mathbf{v}' \to \mathbf{v})$ is continuous in $\mathbf{v}'$ and $\mathbf{v}$; and $v\Sigma(\mathbf{r}, v)$, $k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega})$, and $k_d^{(m)}(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega})$ are continuous in $\mathbf{\Omega}'$ and $\mathbf{\Omega}$ and piecewise continuous in v. For mathematical convenience, we take the values of these functions on ∂D_n to be the limiting value from either side of the boundary; we take Σ, k_0, and $k_d^{(m)}$ to be continuous from the right in v for $0 < v < v_1$ and continuous from the left for $v = v_1$; and require that the limits in v from the left exist for $0 < v < v_1$. (This corresponds to a composite body in which each part is homogeneous in position and threshold effects can occur in speed.)

More complicated discontinuities can be handled using our methods. However we shall not consider them here since the geometrical descriptions and proofs become very lengthy. Assumptions (i) and (j) will be explicitly needed only in Theorem 5 and in Lemmas 3 and 4 (Appendix).

III. THE STREAMING OPERATOR AND ITS SPECTRUM SPECTRUM

In this section we shall consider the operator T described in Sec. 2 and determine its spectrum for the two cases $v_0 > 0$ and $v_0 = 0$. First we consider the simpler case $v_0 > 0$.

Theorem 1: If $v_0 > 0$, then $\sigma(T) = \{\infty\}$.

Proof: For any λ we can formally solve the equation $(\lambda I - T)\phi = \psi$ for ϕ to obtain

$$(\lambda I - T)^{-1}\psi(\mathbf{r},\mathbf{v}) = \frac{1}{v}\int_{t=0}^{d(\mathbf{r},\mathbf{\Omega})} \psi(\mathbf{r} - t\mathbf{\Omega}, \mathbf{v}) \times \exp\left(-\int_0^t \frac{\lambda + v\Sigma(\mathbf{r} - s\mathbf{\Omega}, v)}{v}\,ds\right)dt, \qquad (3.1)$$

where $d(\mathbf{r},\mathbf{\Omega})$ is the distance from $\mathbf{r}$ to ∂D in the direction $-\mathbf{\Omega}$.

For each $v_0 > 0$ and λ, there exists a positive constant $M(\lambda, v_0)$ satisfying

$$\left|\frac{1}{v}\exp\left(-\int_0^t \frac{\lambda + v\Sigma(\mathbf{r}-s\mathbf{\Omega}, v)}{v}\,ds\right)\right| \leqslant M(\lambda, v_0),$$

$$\mathbf{r}\in D, \quad v_0 \leqslant v \leqslant v_1,$$
$$0 \leqslant t \leqslant d(\mathbf{r}, \mathbf{\Omega}).$$

Then by (3.1),

$$\left|(\lambda I - T)^{-1}\psi(\mathbf{r}, \mathbf{v})\right| \leqslant M(\lambda, v_0)\int_{t=0}^{d(\mathbf{r},\mathbf{\Omega})} \left|\psi(\mathbf{r}-t\mathbf{\Omega}, \mathbf{v})\right| dt.$$

Integrating this inequality over r and v, we obtain

$$\|(\lambda I - T)^{-1}\psi\| \leqslant lM(\lambda, v_0)\|\psi\|,$$

where l is the length of the longest straight line in D. Thus $\lambda \in \rho(T)$ for every finite λ. Since T is unbounded, then $\sigma(T)$ must consist solely of the point at ∞. QED

Theorem 2: If $v_0 = 0$, then $\sigma(T) = \{\lambda \,|\, \mathrm{Re}\,\lambda \leqslant -\lambda^*\}$. Further, for each $\lambda \in \sigma(T)$, there exists a sequence $\{\psi_n\} \subset X_0$ such that $\|\psi_n\| = 1$ and $\lim_{n\to\infty}(\lambda I - T)\psi_n = 0$.

Proof: For each λ, the operator $(\lambda I - T)^{-1}$ exists on $R(\lambda I - T)$ (Ref. 40) and is given by (3.1). Thus $\lambda \in \rho(T)$ iff $(\lambda I - T)^{-1}$ is a bounded operator defined on X.

First we consider $\mathrm{Re}\lambda > -\lambda^*$. Then

$$\left|\frac{1}{v}\exp\left(-\int_0^t \frac{\lambda + v\Sigma(\mathbf{r}-s\mathbf{\Omega}, v)}{v}\,ds\right)\right|$$

$$= \left|\frac{1}{v}\exp\left(-t\,\frac{\lambda+\lambda^*}{v}\right)\exp\left(-\int_0^t \frac{v\Sigma(\mathbf{r}+s\mathbf{\Omega}, v) - \lambda^*}{v}\,ds\right)\right|$$

$$\leqslant \frac{1}{v}\exp[-t(\mathrm{Re}\lambda + \lambda^*)/v]\exp(lc_0), \tag{3.2}$$

where λ^* is defined by (2.2), c_0 by (2.3), and l is the length of the longest straight line in D. By (3.1) and (3.2),

$$\left|(\lambda I - T)^{-1}\psi(\mathbf{r}, \mathbf{v})\right| \leqslant \exp(lc_0)\int_{t=0}^{d(\mathbf{r},\mathbf{\Omega})} \left|\psi(\mathbf{r}-t\mathbf{\Omega}, \mathbf{v})\right| \times \frac{\exp[-t(\mathrm{Re}\lambda + \lambda^*)/v]}{v}\,dt.$$

We integrate this inequality along the line L: $\mathbf{r}_0 + s\mathbf{\Omega}$, $\mathbf{r}_0 \in \partial D$, $0 \leqslant s \leqslant d(\mathbf{r}_0, -\mathbf{\Omega})$, to obtain

$$\int_0^{d(\mathbf{r}_0,-\mathbf{\Omega})} \left|(\lambda I - T)^{-1}\psi(\mathbf{r}_0 + s\mathbf{\Omega}, \mathbf{v})\right| ds \leqslant \frac{\exp(lc_0)}{\mathrm{Re}\lambda + \lambda^*}\int_0^{d(\mathbf{r}_0,-\mathbf{\Omega})} \times \left|\psi(\mathbf{r}_0 + t\mathbf{\Omega}, \mathbf{v})\right| dt.$$

Now we integrate over the remaining two space directions and v to get

$$\|(\lambda I - T)^{-1}\psi\| \leqslant \frac{\exp(lc_0)}{\mathrm{Re}\lambda + \lambda^*}\|\psi\|.$$

Thus $\lambda \in \rho(T)$, and

$$\|(\lambda I - T)^{-1}\| \leqslant \frac{\exp(lc_0)}{\mathrm{Re}\lambda + \lambda^*}, \quad \mathrm{Re}\lambda > -\lambda^*. \tag{3.3}$$

Next we consider $\mathrm{Re}\lambda < -\lambda^*$. Then there exists a point $\mathbf{r}_0 \in D$ and positive numbers ϵ_0, ϵ_1, and ϵ_2 such that

$$\mathrm{Re}\lambda + v\Sigma(\mathbf{r}, v) < -\epsilon_0$$

for

$$|\mathbf{r} - \mathbf{r}_0| < \epsilon_1 \text{ and } 0 \leqslant v \leqslant \epsilon_2.$$

For n such that $0 < 1/n < \epsilon_2/2$, we define

$$\phi_n(\mathbf{r}, \mathbf{v}) = \left(\frac{3}{4\pi}\right)^2 \frac{n^3}{7\epsilon_1^3} H_n(\mathbf{r}, v)\exp\left(-i\,\frac{\mathrm{Im}\lambda}{v}\,d(\mathbf{r}, \mathbf{\Omega})\right)$$

where Im denotes "imaginary part" and

$$H_n(\mathbf{r}, v) = \begin{cases} 1, & |\mathbf{r} - \mathbf{r}_0| < \epsilon_1 \text{ and } 1/n < v < 2/n \\ 0, & \text{otherwise.} \end{cases}$$

Then $\phi_n \in X$ and $\|\phi_n\| = 1$. From (3.1) we can easily verify that $(\lambda I - T)^{-1}\phi_n \in X$; also, we obtain the estimate

$$\left|(\lambda I - T)^{-1}\phi_n(\mathbf{r}, \mathbf{v})\right| \geqslant \left(\frac{3}{4\pi}\right)^2 \frac{n^3}{7\epsilon_1^3}\,\frac{1}{\epsilon_0}\left[\exp\left(\frac{\epsilon_0\epsilon_1}{4}\,n\right) - 1\right]$$

$$\text{for } \frac{1}{n} < v < \frac{2}{n}, \quad |\mathbf{r} - \mathbf{r}_0| < \epsilon_1/2.$$

Therefore,

$$\|(\lambda I - T)^{-1}\phi_n\| \geqslant \frac{1}{8\epsilon_0}\left[\exp\left(\frac{\epsilon_0\epsilon_1}{4}n\right) - 1\right].$$

Since this becomes unbounded as $n \to \infty$, then $\lambda \in \sigma(T)$. Thus $\{\lambda \,|\, \mathrm{Re}\lambda < -\lambda^*\} \subset \sigma(T)$, and since the spectrum is a closed set, then $\sigma(T)$ is as described in the statement of the theorem.

Next we let $\mathrm{Re}\lambda < -\lambda^*$ and take ϕ_n to be as defined above. We define ψ_n by

$$\psi_n = (\lambda I - T)^{-1}\phi_n / \|(\lambda I - T)^{-1}\phi_n\|.$$

Then $\|\psi_n\| = 1$ and $\lim_{n\to\infty}(\lambda I - T)\psi_n = 0$. For $\mathrm{Re}\lambda = -\lambda^*$, there exist sequences $\{\lambda_n\}$ with $\mathrm{Re}\lambda_n < -\lambda^*$ and $\lambda_n \to \lambda$, and $\{\psi_{n,m}\}$ with $\|\psi_{n,m}\| = 1$ and $(\lambda_n I - T)\psi_{n,m} \to 0$ as $m \to \infty$. We can thus construct a sequence $\psi_n = \psi_{n,m_n}$ such that $(\lambda_n I - T)\psi_n \to 0$. Then ψ_n satisfies

$$\lim_{n\to\infty}(\lambda I - T)\psi_n = \lim_{n\to\infty}[(\lambda - \lambda_n)\psi_n + (\lambda_n I - T)\psi_n] = 0.$$

This proves the second half of the theorem. QED

Thus the finite spectrum exists and is a half-space only if neutrons can exist with arbitrarily small speeds. Also the description of $\sigma(T)$ (i.e., of λ^*) depends only on the limiting value of $v\Sigma(\mathbf{r}, v)$ as $v \to 0$ and is insensitive to discontinuities or nonmonotonicity of $v\Sigma(\mathbf{r}, v)$.

Finally, we add that for $v_0 > 0$, the calculations leading to (3.3) can be modified to yield the useful inequality

$$\|(\lambda I - T)^{-1}\| \leqslant \frac{1}{\mathrm{Re}\lambda + \hat{\lambda}(v_0)}, \quad \mathrm{Re}\lambda > -\hat{\lambda}(v_0), \tag{3.4}$$

where $\hat{\lambda}(v_0)$ is defined by (2.4).

IV. THE ONE SPEED OPERATOR $T + K_0$ AND ITS SPECTRUM

In this section we shall consider the operator $T + K_0$ and determine the basic properties of its spectrum. $T + K_0$ is a one-speed operator in the sense that it commutes with functions of v alone. Thus we define an auxiliary Banach space X^0 as follows. We let S be the unit sphere in V and define X^0 as the set of all complex-valued functions $\psi(\mathbf{r}, \mathbf{\Omega})$ defined on $\bar{D} \times S$, satisfying

$$\|\psi\|^0 = \int_{\mathbf{r}\in D}\int_{|\mathbf{\Omega}|=1} |\psi(\mathbf{r}, \mathbf{\Omega})|\,d\mathbf{r}\,d\mathbf{\Omega}.$$

Then the operators $T = T(v)$, $K_0 = K_0(v)$, and $T + K_0 = A(v)$ (which, as we have indicated, depend parametrically upon v) are defined on the subspace $X_0^0 \subset X^0$ of functions such that $T(v)\psi \in X^0$ and $\psi(\mathbf{r}, \mathbf{\Omega}) = 0$ for $\mathbf{r} \in \partial D$ and $\mathbf{\Omega}$ pointing into D. Clearly, X_0^0 is independent of v.

To proceed we need the following theorem, which is due to Gohberg.[42]

Theorem (Gohberg): Let $L(\lambda)$ be an operator-valued function, holomorphic in an open connected set G, and compact for $\lambda \in G$. Then for all points $\lambda \in G$, with the possible exception of certain isolated points, the number $\alpha(\lambda)$ of linearly independent solutions of the equation

$$[I - L(\lambda)]\phi = 0$$

is constant:

$$\alpha(\lambda) = n;$$

at the isolated points mentioned,

$$\alpha(\lambda) > n.$$

Henceforth we shall denote all quantities pertaining to X^0 by a "zero" superscript. We can now prove

Theorem 3: Let $\tilde{\lambda}(v) = \inf_{\mathbf{r}} v\Sigma(\mathbf{r}, v)$. Then $\{\lambda \mid \mathrm{Re}\lambda > -\tilde{\lambda}(v) + \|K_0(v)\|^0\} \subset \rho^0[A(v)]$. Also, $\sigma^0[A(v)]$ consists entirely of isolated eigenvalues of finite geometrical multiplicity.

Proof: The proof of Theorem 1 can be modified to show that $\sigma^0[T(v)] = \{\infty\}$. Also, the calculations leading to (3. 3) can be modified to give

$$\|[\lambda I - T(v)]^{-1}\|^0 \leqslant \frac{1}{\mathrm{Re}\lambda + \tilde{\lambda}(v)} \quad \text{for } \mathrm{Re}\lambda > -\tilde{\lambda}(v),$$

and thus

$$\|[\lambda I - T(v)]^{-1} K_0(v)\|^0 \leqslant \frac{\|K_0(v)\|^0}{\mathrm{Re}\lambda + \tilde{\lambda}(v)} < 1$$
$$\text{for } \mathrm{Re}\lambda > -\tilde{\lambda}(v) + \|K_0(v)\|^0. \tag{4. 1}$$

For such λ, $\lambda I - A(v) = [\lambda I - T(v)]\{I - [\lambda I - T(v)]^{-1}K_0(v)\}$ is invertible, yielding

$$\|[\lambda I - A(v)]^{-1}\| \leqslant \frac{1}{\mathrm{Re}\lambda + \tilde{\lambda}(v) - \|K_0(v)\|^0},$$
$$\mathrm{Re}\lambda > -\tilde{\lambda}(v) + \|K_0(v)\|^0. \tag{4. 2}$$

This proves the first half of the theorem.

To prove the remainder of the Theorem, we shall consider the operator $Q(\lambda, v) \equiv K_0(v)[\lambda I - T(v)]^{-1}K_0(v)$. In the Appendix, we shall prove that Q is a compact operator on X^0. (See Lemma 3.) Therefore, $Q(\lambda, v)[\lambda I - T(v)]^{-1} = \{K_0(v)[\lambda I - T(v)]^{-1}\}^2$ is a compact, operator-valued function of λ which is holomorphic in the entire complex plane. Also, by (4. 1), $I - \{K_0(v)[\lambda I - T(v)]^{-1}\}^2$ is invertible for $\mathrm{Re}\lambda > -\tilde{\lambda}(v) + \|K_0(v)\|^0$. Thus by Gohberg's Theorem there exists at most a set of isolated values of λ in the complex plane such that $1 \in P\sigma^0\{[K_0(v)[\lambda I - T(v)]^{-1}]^2\}$, and at these points 1 is an eigenvalue of finite geometrical multiplicity. At all other values of λ, $1 \in \rho^0\{[K_0(v)[\lambda I - T(v)]^{-1}]^2\}$.

Now since $\{K_0(v)[\lambda I - T(v)]^{-1}\}^2$ is compact, then $K_0(v)[\lambda I - T(v)]^{-1}$ is potentially compact[41] and its spectrum, except possibly for the point 0, consists entirely of point spectrum.

Thus by the spectral mapping theorem, only for the above set of isolated values of λ can we have $1 \in P_\sigma^0\{K_0(v)[\lambda I - T(v)]^{-1}\}$. At such a λ value the equation $0 = \{I - K_0(v)[\lambda I - T(v)]^{-1}\}\phi = [\lambda I - A(v)][\lambda I - T(v)]^{-1}\phi$ has a finite number of solutions. Consequently, λ is an eigenvalue of $A(v)$ of finite geometrical multiplicity. At all other values of λ, $1 \in \rho^0\{K_0(v)[\lambda I - T(v)]^{-1}\}$ and for such λ the operator

$$I - K_0(v)[\lambda I - T(v)]^{-1} = [\lambda I - A(v)][\lambda I - T(v)]^{-1}$$

has a bounded inverse defined on X^0. Taking this inverse, we find

$$[\lambda I - A(v)]^{-1} = [\lambda I - T(v)]^{-1}\{I - K_0(v)[\lambda I - T(v)]^{-1}\}^{-1},$$

and consequently $\lambda \in \rho^0[A(v)]$. This proves the theorem.

QED

The next theorem is based on a result of Vidav[22]:

Theorem 4: Let $\tilde{\lambda}_0(v) \equiv \sup_{\lambda \in \sigma^0[A(v)]} \mathrm{Re}\lambda$, and let $\tilde{\lambda}(v) < \tilde{\lambda}_0(v)$. Then $\tilde{\lambda}_0(v)$ is an eigenvalue of $A(v)$, corresponding to which is a positive eigenfunction.

Proof: A simple modification of the proof of Theorem 3, Ref. 22 yields the result. QED

Thus, the picture of $\sigma^0[A(v)]$ which emerges can be graphically described by Fig. 1. In the Appendix, we show that $\|K_0(v)\|^0 = \sup_{\mathbf{r}} v\Sigma_0(\mathbf{r}, v)$, where $\Sigma_0(\mathbf{r}, v)$ is the cross section for low energy elastic collisions at speed v. Thus there is no spectrum to the right of the line

$$\mathrm{Re}\lambda = v[\sup_{\mathbf{r}} \Sigma_0(\mathbf{r}, v) - \inf_{\mathbf{r}} \Sigma(\mathbf{r}, v)].$$

We note that the above results hold for v fixed; consequently, the eigenvalues depend parametrically upon v. In general, as v varies between v_0 and v_1, some of the eigenvalues will remain stationary and some will shift and trace out curves. Thus, the set

$$\mathcal{S} \equiv \bigcup_{v_0 \leqslant v \leqslant v_1} \sigma^0[A(v)] \tag{4. 3}$$

will consist of isolated points and curved lines. Using $\mathcal{S}$, we have:

Theorem 5: Let $\mathcal{S}$ be defined by (4. 3). Then, for $v_0 \geqslant 0$,

$$\sigma(T + K_0) = \sigma(T) \cup \bar{\mathcal{S}} \tag{4. 4}$$

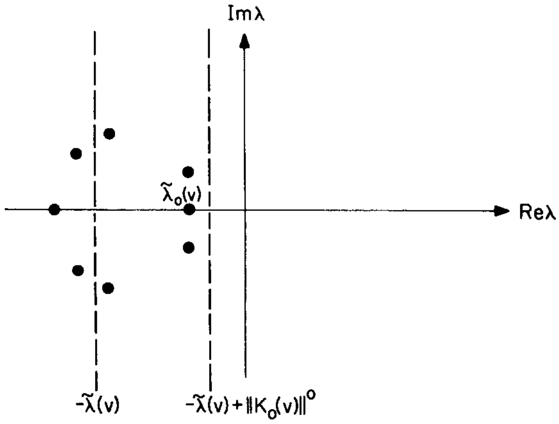


FIG. 1. $\sigma^0[A(v)]$, $\tilde{\lambda}(v) < \tilde{\lambda}_0(v)$.

and

$$\{\lambda \mid \mathrm{Re}\lambda > -\hat{\lambda}(v_0) + \|K_0\|\} \subset \rho(T + K_0). \qquad (4.5)$$

Furthermore, for each $\lambda \in \sigma(T+K_0)$, there exists a sequence $\{\psi_n\} \subset X$ such that $\|\psi_n\| = 1$, $\psi_n(\mathbf{r}, \mathbf{v}) = 0$ for $v > v_t$, and $\|(\lambda I - T - K_0)\psi_n\| \to 0$.

Proof: First we show that $\sigma(T) \cup \mathcal{S} \subset \sigma(T+K_0)$. To do this, let $\lambda \in \sigma(T)$. Then by Theorem 2, there exists a sequence $\{\phi_n\}$ with $\|\phi_n\| = 1$ such that $(\lambda I - T)\phi_n \to 0$. By Lemma 1 (see the Appendix) there exists a sequence of integers M_n such that, with

$$\psi_n(\mathbf{r}, \mathbf{v}) = \exp(iM_n \mathbf{\Omega}\cdot\mathbf{\Omega}_0)\, \phi_n(\mathbf{r}, \mathbf{v}), \qquad (4.6)$$

we have

$$\|K_0\psi_n\| \to 0.$$

Thus,

$$\|(\lambda I - T - K_0)\psi_n\| \leq \|(\lambda I - T)\phi_n\| + \|K_0\psi_n\| \to 0,$$

and so if $(\lambda I - T - K_0)^{-1}$ exists, it must be unbounded. Consequently, $\lambda \in \sigma(T+K_0)$, and so $\sigma(T) \subset \sigma(T+K_0)$.

Next, we let $\lambda \in \mathcal{S}$. Then by Theorem 3 there exists a function $\widetilde{\phi}(\mathbf{r}, \mathbf{\Omega})$ such that $\|\widetilde{\phi}\|^0 = 1$ and $(\lambda I - T - K_0)\widetilde{\phi}(\mathbf{r}, \mathbf{\Omega}) = 0$ for $v = \widetilde{v}$. If $\widetilde{v} < v_1$, we define the sequence ψ_n by

$$\psi_n(\mathbf{r}, \mathbf{v}) = \begin{cases} n/v^2 \quad \phi(\mathbf{r}, \mathbf{\Omega}), & \widetilde{v} \leq \widetilde{v} \leq v + 1/n \\ 0, & \text{otherwise.} \end{cases} \qquad (4.7)$$

Then $\|\psi_n\| = 1$. For $\widetilde{v} \leq v \leq \widetilde{v} + 1/n$,

$$\begin{aligned}(\lambda I - T - K_0)\psi_n(\mathbf{r}, \mathbf{v}) = {}& n\,\frac{v\Sigma(\mathbf{r}, v) - \widetilde{v}\Sigma(\mathbf{r}, \widetilde{v})}{v^2}\,\phi(\mathbf{r}, \mathbf{\Omega}) \\ &+ n\,\frac{v - \widetilde{v}}{v^2}\,\mathbf{\Omega}\cdot\nabla\phi(\mathbf{r}, \mathbf{\Omega}) \\ &- \frac{n}{v^2}\int_{|\mathbf{\Omega}'|=1} [k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega}) \\ &- k_0(\mathbf{r}, \widetilde{v}, \mathbf{\Omega}' \to \mathbf{\Omega})]\phi(\mathbf{r}, \mathbf{\Omega}')d\mathbf{\Omega}',\end{aligned}$$

while for $v < \widetilde{v}$ or $v > \widetilde{v} + 1/n$, $(\lambda I - T - K_0)\psi_n = 0$. Thus, integrating over $\mathbf{r}$ and $\mathbf{v}$, we obtain

$$\begin{aligned}\|(\lambda I - T - K_0)\psi_n\| \leq {}& n\int_{v=\widetilde{v}}^{\widetilde{v}+1/n} |v\Sigma(\mathbf{r}, v) - \widetilde{v}\Sigma(\mathbf{r}, \widetilde{v})|\, dv \\ &+ \frac{1}{2n}\int_{\mathbf{r}\in D}\int_{|\mathbf{\Omega}|=1} |\mathbf{\Omega}\cdot\nabla\,\phi(\mathbf{r}, \mathbf{\Omega})|\, d\mathbf{\Omega}d\mathbf{r} \\ &+ n\int_{v=\widetilde{v}}^{\widetilde{v}+1/n}\int_{r\in D}\int_{|\mathbf{\Omega}|=1}\int_{|\mathbf{\Omega}'|=1} \\ &\times |k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega}) - k_0(\mathbf{r}, \widetilde{v}, \mathbf{\Omega}' \to \mathbf{\Omega})| \\ &\times |\phi(\mathbf{r}, \mathbf{\Omega}')|\, d\mathbf{\Omega}'\, d\mathbf{\Omega}d\mathbf{r}dv.\end{aligned}$$

Since $v\Sigma(\mathbf{r}, v)$ and $k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega})$ are continuous from the right in v (see Sec. 2), then each of the above integrals will tend to zero as $n \to \infty$. Consequently, $\|(\lambda I - T - K_0)\psi_n\| \to 0$, and so $\lambda \in \sigma(T+K_0)$. This result holds if $\widetilde{v} < v_1$. If $\widetilde{v} = v_1$, we define the sequence ψ_n as in (4.7), except that we take ψ_n to be nonzero over the interval $v_1 - 1/n \leq v \leq v_1$. Then since $v\Sigma(\mathbf{r}, v)$ and k_0 are continuous from the left at v_1, the above procedure will apply. Thus we have $\mathcal{S} \subset \sigma(T+K)$.

Hence $\sigma(T) \cup \bar{\mathcal{S}} \subset \sigma(T+K_0)$, since the spectrum is closed. To prove inclusion the other way, we shall consider the equivalent inclusion

$$C[\sigma(T) \cup \bar{\mathcal{S}}] \subset \rho(T+K_0), \qquad (4.8)$$

where C means "complement." Thus, we let $\lambda \in C[\sigma(T) \cup \bar{\mathcal{S}}]$. Then the equation

$$(\lambda I - T - K_0)\phi(\mathbf{r}, \mathbf{v}) = \psi(\mathbf{r}, \mathbf{v})$$

has a solution ϕ such that $\|\phi(\mathbf{r}, \mathbf{v})\|^0 \leq \text{const}\|\psi(\mathbf{r}, \mathbf{v})\|^0$ for $v_0 \leq v \leq v_1$, even if $v_0 = 0$. Consequently, $\|\phi\| = \|(\lambda I - T - K_0)^{-1}\psi\| < \infty$ for each $\psi \in X$. By the results in Sec. 2, T is closed and K is bounded, so $T + K_0$ is closed in X. Consequently, $(\lambda I - T - K_0)^{-1}$ is closed, and so by the closed graph theorem is bounded. Thus $\lambda \in \rho(T+K_0)$. This proves (4.8), and also (4.4).

Next we use (3.4) and we repeat the same arguments which led to (4.2) to obtain

$$\|(\lambda I - T - K_0)^{-1}\| \leq \frac{1}{\mathrm{Re}\lambda + \hat{\lambda}(v_0) - \|K_0\|},$$

$$\mathrm{Re}\lambda > -\hat{\lambda}(v_0) + \|K_0\|. \qquad (4.9)$$

This proves (4.5).

Finally, we let $\lambda \in \sigma(T+K_0)$. (With no change in the spectrum, we can decrease v_1 so that $v_1 = v_t$.) If $\lambda \in \sigma(T)$, then the sequence described by (4.6) satisfies $\|\psi_n\| = 1$ and $\|(\lambda I - T - K_0)\psi_n\| \to 0$. If $\lambda \in \mathcal{S}$ then the sequence described by (4.7) satisfies these conditions. If $\lambda \in \bar{\mathcal{S}}$, then there exist sequences $\{\psi_n\} \subset \mathcal{S}$ with $\lambda_n \to \lambda$ and $\{\psi_{n,m}\}$ with $\|\psi_{n,m}\| = 1$ and $\|(\lambda_n I - T - K_0)\psi_{n,m}\| \to 0$ as $m \to \infty$. We can thus construct a sequence $\psi_n \equiv \psi_{n,m_n}$ such that $(\lambda_n I - T - K_0)\psi_n \to 0$. Then

$$(\lambda I - T - K_0)\psi_n = (\lambda - \lambda_n)\psi_n + (\lambda_n I - T - K_0)\psi_n \to 0.$$

These results hold for $v_1 = v_t$. For $v_1 > v_t$, we extend the above functions ψ_n by defining them to be zero for $v_1 \geq v > v_t$; this yields a sequence which satisfies all the conditions of the theorem.

This completes the proof of Theorem 5. QED

We remark that $\sigma(T+K_0)$ is described graphically by Fig. 2 for $v_0 > 0$ and by Fig. 3 for $v_0 = 0$. Also, as in the case of $\sigma^0[A(v)]$, $\|K_0\| = \sup_{\mathbf{r},v} v\Sigma_0(\mathbf{r}, v)$ (see Appendix). Thus there can exist no spectrum to the right of the line $\mathrm{Re}\lambda = \sup_{\mathbf{r},v} v\Sigma_0(\mathbf{r}, v) - \inf_{\mathbf{r},v} v\Sigma(\mathbf{r}, v)$. Note that the de-

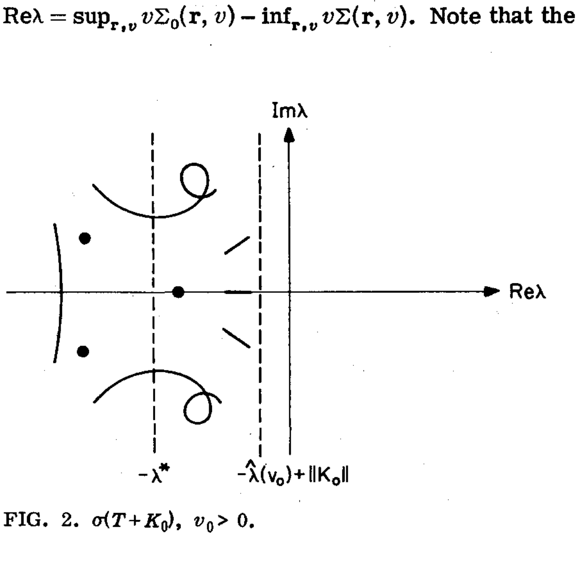


FIG. 2. $\sigma(T+K_0)$, $v_0 > 0$.

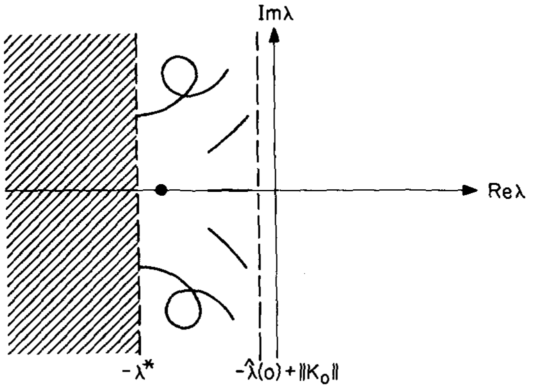


FIG. 3. $\sigma(T+K_0)$, $v_0=0$.

finition of $\Sigma_0(\mathbf{r}, v)$ is such that it may actually exceed $\Sigma(\mathbf{r}, v)$. In fact, in the extreme one-speed case,[44] $v\Sigma(\mathbf{r}, v)=1$ and $v\Sigma_0(\mathbf{r}, v)=c$.

Finally, let us consider $v_0=0$. We note that K_0 is a bounded operator on X but that the operator $v^{-1}K_0$ is in general unbounded. However, if k_0 satisfies an inequality of the form $k_0(\mathbf{r}, v, \mathbf{\Omega}' \to \mathbf{\Omega}) \leq M_0/v$ [instead of (2.12)], then $v^{-1}K_0$ will be bounded. For cases of this type, we can prove the following theorem:

Theorem 6: Let $v_0=0$ and $v^{-1}K_0$ be a bounded operator on X. If the maximum diameter l of D satisfies

$$l\exp(lc_0)\|v^{-1}K_0\| < 1, \tag{4.10}$$

then

$$\{\lambda \mid \mathrm{Re}\lambda > -\lambda^*\} \subset \rho(T+K_0). \tag{4.11}$$

Proof: Let (4.10) be satisfied and let $\mathrm{Re}\lambda > -\lambda^*$. Then by (3.1) and (3.2) we have

$$\begin{aligned}|(\lambda I-T)^{-1}K_0\psi(\mathbf{r},\mathbf{v})| &\leq \int_{t=0}^{d(\mathbf{r},\mathbf{\Omega})} |(v^{-1}K_0)\psi(\mathbf{r}-t\mathbf{\Omega},\mathbf{v})| \\ &\quad\times\exp\left(-t\,\frac{\mathrm{Re}\lambda+\lambda^*}{v}\right)\exp(lc_0)\,dt \\ &\leq \exp(lc_0)\int_{t=0}^{d(\mathbf{r},\mathbf{\Omega})} |(v^{-1}K_0)\psi(\mathbf{r}-t\mathbf{\Omega},\mathbf{v})|\,dt.\end{aligned}$$

Integrating over $\mathbf{r}$ and $\mathbf{v}$, we obtain

$$\|(\lambda I-T)^{-1}K_0\| \leq l\exp(lc_0)\|v^{-1}K_0\| < 1.$$

Therefore, the operators

$$(\lambda I-T)^{-1}(\lambda I-T-K_0) \subset I-(\lambda I-T)^{-1}K_0$$

have bounded inverses defined on X, and

$$(\lambda I-T-K_0)^{-1}=[I-(\lambda I-T)^{-1}K_0]^{-1}(\lambda I-T)^{-1}$$

is bounded, proving that $\lambda\in\rho(T+K_0)$. This verifies the inclusion (4.11). QED

Thus if the hypotheses of Theorem 6 are satisfied, then all the spectrum of A is imbedded in the "continuum" $\mathrm{Re}\lambda \leq -\lambda^*$. We shall comment on this in Sec. 6.

V. THE TRANSPORT OPERATOR $T+K$ AND ITS SPECTRUM

In this section, we shall prove that $\sigma(T+K)$ differs from $\sigma(T+K_0)$ only by the addition of point spectrum, and we shall give estimates on the location of this spectrum. If $\rho(T+K_0)$ is a connected set, then the added spectrum consists of isolated, discrete points of finite geometrical multiplicity. If $\rho(T+K_0)$ is not a connected set, as in Fig. 2 and 3, then certain connected components of $\rho(T+K_0)$ can become wholly or partially filled with point spectrum. We shall prove these results in the following theorems.

Theorem 7: $\sigma(T+K_0)\subset\sigma(T+K)$.

Proof: Let $\lambda\in\sigma(T+K_0)$. Then by Theorem 4, there exists a sequence $\{\phi_n\}$ such that $\|\phi_n\|=1$, $\phi_n(\mathbf{r},\mathbf{v})=0$ for $v>v_t$, and $(\lambda I-T-K_0)\phi_n\to 0$. By Lemma 2 (Appendix), there exists a sequence of integers $\{M_n\}$ such that, with

$$\psi_n(\mathbf{r},\mathbf{v})\equiv\phi_n(\mathbf{r},\mathbf{v})\exp(iM_nv),$$

we have $\|\psi_n\|=1$ and $K_c\psi_n\to 0$. Furthermore, by condition (f) of Sec. 2, $K_d\psi_n=0$. Therefore, by (2.5),

$$\|(\lambda I-T-K)\psi_n\| \leq \|(\lambda I-T-K_0)\phi_n\|+\|K_c\psi_n\|\to 0.$$

This proves the theorem. QED

To state the next theorem, we write $\rho(T+K_0)$ as the union of its connected components:

$$\rho(T+K_0)=\bigcup_{-\alpha\leq n\leq\alpha} S_n,$$

where α is a nonnegative integer or ∞. Each S_n is a connected, open set and is the reflection of S_{-n} across the Reλ axis. S_0 is the "largest" of these sets and contains the right half plane $\mathrm{Re}\lambda > -\hat{\lambda}(v_0)+\|K\|$. (See Theorem 9.) Figures 2 and 3 illustrate this situation. If the lines generated by $\sigma(T+K_0)$ do not form closed loops, then $\rho(T+K_0)$ is connected and is equal to S_0.

Now we can state the theorem.

Theorem 8: The set $\sigma(T+K)\cap\rho(T+K_0)$ is described by:

(i) $[\sigma(T+K)\cap\rho(T+K_0)]\subset P\sigma(T+K)$.

(ii) $\sigma(T+K)\cap S_n$ consists of eigenvalues of finite geometrical multiplicity.

(iii) $\sigma(T+K)\cap S_0$ consists of isolated points.

Proof: We define $K_1\equiv K_c+K_d$. Then the operator $Q(\lambda)\equiv K_1(\lambda I-T-K_0)^{-1}$ is a holomorphic, operator-valued function of λ in S_n. Simple algebraic manipulations allow us to rewrite $Q(\lambda)$ in the form

$$Q(\lambda)=K_1(\lambda I-T)^{-1}+K_1(\lambda I-T)^{-1}K_0(\lambda I-T-K_0)^{-1}. \tag{5.1}$$

In the Appendix we shall show that for $\lambda\in\rho(T)$, $K_1(\lambda I-T)^{-1}K_0$ and $[K_1(\lambda I-T)^{-1}]^{M_1+1}$ are compact, where M_1 satisfies (2.10). Then $Q^{M_1+1}(\lambda)$ is a holomorphic, compact operator-valued function of λ in S_n. It follows from Gohberg's theorem that either $1\in P\sigma[Q^{M_1+1}(\lambda)]$ for all $\lambda\in S_n$, or there exist at most isolated values of λ for which $1\in P\sigma[Q^{M_1+1}(\lambda)]$ and $1\in\rho[Q^{M_1+1}(\lambda)]$ for all other values of $\lambda\in S$. In either case the eigenvalue 1 has finite geometrical multiplicity.

Since $Q^{M_1+1}(\lambda)$ is compact, then $Q(\lambda)$ is potentially compact and its spectrum, except possibly for the point 0, consists entirely of point spectrum. Thus by the

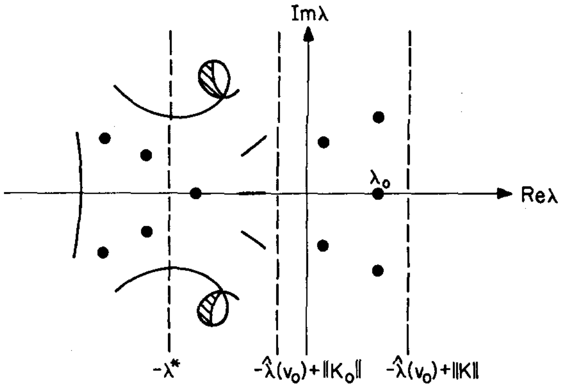


FIG. 4. $\sigma(T+K)$, $v_0>0$, $-\hat{\lambda}(v_0)<\sup\tilde{\lambda}_0(v)<\lambda_0$.

spectral mapping theorem, we must have for $\lambda\in S_n$ either $1\in\rho[Q(\lambda)]$ or $1\in P\sigma[Q(\lambda)]$. If 1 is in the point spectrum, it has finite geometrical multiplicity.

If $1\in\rho[Q(\lambda)]$, then

$$I-Q(\lambda)=(\lambda I-T-K)(\lambda I-T-K_0)^{-1}$$

has a bounded inverse defined on X, and hence

$$(\lambda I-T-K)^{-1}=(\lambda I-T-K_0)^{-1}[I-Q(\lambda)]^{-1}.$$

Thus $\lambda\in\rho(T+K)$.

If $1\in P\sigma[Q(\lambda)]$, then the equation

$$0=[I-Q(\lambda)]\phi=(\lambda I-T-K)\,[(\lambda I-T-K_0)^{-1}\phi]$$

has a finite number of solutions. Hence $\lambda\in P\sigma(T+K)$ and the geometrical multiplicity of λ is finite. This proves claims (i) and (ii).

To prove claim (iii), we note from (5.1) and (4.9) that $Q(\lambda)\to 0$ as $\mathrm{Re}\lambda\to\infty$. Hence by Gohberg's theorem there exist at most isolated values of $\lambda\in S_0$ for which $1\in P\sigma[Q^{M_1+1}(\lambda)]$. By the spectral mapping theorem, only for these λ values can $1\in P\sigma[Q(\lambda)]$, and as we showed above, only such λ can be in $\sigma(T+K)$. QED

The next theorem provides estimates on the location of $\sigma(T+K)$.

Theorm 9: $\{\lambda\,|\,\mathrm{Re}\lambda>-\hat{\lambda}(v_0)+\|K\|\}\subset\rho(T+K)$. Also, if $v_0=0$, if $v^{-1}K$ is a bounded operator on X, and if the maximum diameter l of D satisfies

$$l\exp(lc_0)\|v^{-1}K\|<1, \tag{5.3}$$

then

$$\{\lambda\,|\,\mathrm{Re}\lambda>-\lambda^*\}\subset\rho(T+K).$$

Proof: Using (3.4) and repeating the argument which led to (4.2), we obtain

$$\|(\lambda I-T-K)^{-1}\|\leqslant\frac{1}{\mathrm{Re}\lambda+\hat{\lambda}(v_0)-\|K\|},\quad \mathrm{Re}\lambda>-\hat{\lambda}(v_0)+\|K\|. \tag{5.4}$$

This proves the first part of the theorem. To prove the second part, we simply repeat the proof of Theorem 6 with K replacing K_0. QED

The next result generalizes Theorem 4, from which we borrow some notation.

Theorem 10: Let $\lambda_0\equiv\sup_{\lambda\in\sigma(A)}\mathrm{Re}\lambda$. If $\lambda_0>\sup_v\tilde{\lambda}_0(v)$ and $\lambda_0>-\hat{\lambda}(v_0)$ then λ_0 is an eigenvalue of A, corresponding to which is a positive eigenfunction.

Proof: As in Theorem 4, we refer to the proof of Theorem 3 in Ref. 22. QED

Thus, $\sigma(T+K)$ is described by Fig. 4 for $v_0>0$ and by Fig. 5 for $v_0=0$. These figures differ from Figs. 2 and 3 respectively only by the addition of point spectrum. We note that if "loops" exist, as shown in Figs. 2 and 3, then we cannot exclude the possibility that their interiors (the sets S_n with $n=0$) become partially or wholly filled with point spectrum. Also, $\|K\|$ is given in the Appendix and as in the cases of $\sigma^0[A(v)]$ and $\sigma(T+K_0)$, we deduce an absolute limit to the real part of $(T+K)$ as $\sup_{\mathbf{r},v}v\Sigma_s(\mathbf{r},v)-\inf_{\mathbf{r},v}v\Sigma(\mathbf{r},v)$.

We conclude the main body of this paper with the following result:

Theorem 11: The transport operator $A=T+K$ is the infinitesimal generator of a strongly continuous semigroup of operators.

Proof: By the results in Sec. 2, T is a closed, densely defined operator and K is bounded. Thus $T+K$ is a closed, densely defined operator satisfying (5.4), and so the conditions of the Hille–Yosida–Phillips theorem[46] are met. QED

It follows that the semigroup $\Upsilon(t)=\exp(At)$ exists and enables us to solve the initial value problem

$$\frac{\partial\psi}{\partial t}=A\psi$$

$$\psi|_{t=0}=\psi_0.$$

The behavior of the solution of this problem thus depends on the location and classification of $\sigma(A)$. In this paper we have described many of the basic properties of this spectrum for arbitrary bounded domains and, we hope, realistic and general transport operators.

VI. DISCUSSION

If we refer to the results of Sec. 5, as exhibited schematically in Figs. 4 and 5, we see that the spectrum of $T+K$ can have a rather complicated structure. This is due in part to the curves and loops which $\sigma(T+K)$ inherits from $\sigma(T+K_0)$ (Figs. 2 and 3). At present,

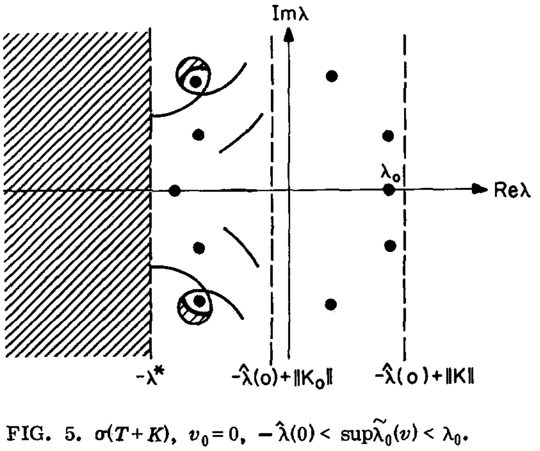


FIG. 5. $\sigma(T+K)$, $v_0=0$, $-\hat{\lambda}(0)<\sup\tilde{\lambda}_0(v)<\lambda_0$.

little is known about the spectrum of this one-speed operator. However, for a restrictive case (isotropic scattering, spatial independence of all cross sections and kernels, and analysis in L_2) it is known that the portion of the spectrum of one-speed operators which lies to the right of the line $\mathrm{Re}\lambda = -v\Sigma(v)$ is real [where $\Sigma(v)$ is the total cross section]. If this is true in general, then the curves in Figs. 2 through 5 to the right of the line $\mathrm{Re}\lambda = -\hat{\lambda}(v_0)$ will consist of line segments on the real axis, and the spectral picture for $T+K$ will simplify considerably.

Another question of obvious importance concerns the existence of a simple dominant eigenvalue, i.e., a simple eigenvalue whose real part is larger than any other λ in the spectrum, and whose eigenfunction is nonnegative. Physically one expects that such an eigenvalue will exist but this remains to be proved. Our results indicate only that under the conditions of Theorem 10, a "semidominant" real eigenvalue λ_0 with real eigenfunction exists. At present we cannot show that complex eigenvalues with real parts equal to λ_0 do not exist, nor can we say anything about the algebraic multiplicity of λ_0 or any other eigenvalue. (Vidav's proof[22] that the eigenvalues out of the continuum have finite multiplicity is valid only for geometrical multiplicity, and his proof that λ_0 is a simple eigenvalue in L_p, $p>1$, only shows that λ_0 has geometrical multiplicity one.)

Finally, we note (Theorem 9) that if $v_0=0$ and $v^{-1}K$ is a bounded operator, then for sufficiently small bodies the spectrum to the right of the line $\mathrm{Re}\lambda = -\lambda^*$ disappears. This famous "disappearance of the point spectrum into the continuum" was first predicted on a heuristic basis by Nelkin[7] and has become a part of the folklore of neutron transport theory. It turns out that this effect has never been observed experimentally.[47] Our results suggest that this is due to the absence of the continuum to the left of $\mathrm{Re}\lambda = -\lambda^*$ for $v_0>0$. In other words the case $v_0>0$ corresponds more closely to physical reality. This is hardly surprising; the Boltzman equation considered here treats the neutrons as classical particles and cannot be expected to be valid for neutron speeds so low that the neutron wavelength becomes comparable to a mean free path. To consider realistically the case $v_0=0$, another equation should be studied. The experimental evidence suggests strongly that equation would not predict the continuous spectrum we find here for the case $v_0=0$.

APPENDIX

Here we shall prove certain results which were needed earlier. We shall state these results as lemmas.

Lemma 1: Let $\psi \in X$ and let $\mathbf{\Omega}_0$ be fixed. Then $\psi_n(\mathbf{r},\mathbf{v}) \equiv \psi(\mathbf{r},\mathbf{v})\exp(in\mathbf{\Omega}\cdot\mathbf{\Omega}_0) \in X$, $\|\psi_n\| = \|\psi\|$, and $K_0\psi_n \to 0$.

Lemma 2: Let $\psi \in X$. Then $\psi_n(\mathbf{r},\mathbf{v}) \equiv \psi(\mathbf{r},\mathbf{v})\exp(\mathrm{inv}) \in X$, $\|\psi_n\| = \|\psi\|$, and $K_c\psi_n \to 0$.

Proofs: By the Riemann–Lebesgue Lemma,[48] the sequence $K_0\psi_n$ of Lemma 1 satisfies $\lim_{n\to\infty}(K_0\psi_n)(\mathbf{r},\mathbf{v}) = 0$ for almost every $\mathbf{r}$ and $\mathbf{v}$. Since, by (2.11) and (2.12),

$$|(K_0\psi_n)(\mathbf{r},\mathbf{v})| \leq M_0 \int_{\Omega'} \psi(\mathbf{r}, v\Omega')d\Omega' \equiv g(\mathbf{r}, v),$$

and since $\|g\| < \infty$, then it follows from the Lebesgue dominated convergence theorem[48] that $\|K_0\psi_n\| \to 0$. Lemma 2 is proved in a similar way. QED

Lemma 3: $K_0(v)[\lambda I - T(v)]^{-1}K_0(v)$ is, for fixed v, a compact operator in X^0.

Lemma 4: For $\lambda \in \rho(T)$, the operators $K_1(\lambda I - T)^{-1}K_0$, $K_c(\lambda I - T)^{-1}K_c$, and $K_d(\lambda I - T)^{-1}K_c$ are compact in X.

Proofs: The proofs that each of the operators of Lemmas 3 and 4 is compact are virtually identical. Thus we shall single out $K_1(\lambda I - T)^{-1}K_0$ and prove the result only for this operator, and for the more difficult case $v_0=0$.

Since $K_1 = K_c + K_d$, then by (2.13) $K_1(\lambda I - T)^{-1}K_0 = K_c(\lambda I - T)^{-1}K_0$. Thus we need to show only that $K_c(\lambda I - T)^{-1}K_0 \equiv L$ is compact for $\mathrm{Re}\lambda > -\lambda^*$.

Using (2.6), (2.10), and (3.1) we write L in the explicit form

$$(L\psi)(\mathbf{r},\mathbf{v}) = \int_{\mathbf{r}'}\int_{\mathbf{v}'} G(\mathbf{r}',\mathbf{v}',\mathbf{r},\mathbf{v})\psi(\mathbf{r}',\mathbf{v}')d\mathbf{v}'\,d\mathbf{r}'$$

where $]^{m_1+1}$,

$$G(\mathbf{r}',\mathbf{v}',\mathbf{r},\mathbf{v}) = \frac{v'}{|\mathbf{r}-\mathbf{r}'|^2} k_0\left(\mathbf{r}', v', \Omega' \to \frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}\right) \times k_c\left(\mathbf{r}, v'\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|} \to \mathbf{v}\right) \times \exp\left\{-\int_{s=0}^{|\mathbf{r}-\mathbf{r}'|} \frac{1}{v'}\left[\lambda + v'\Sigma\left(\mathbf{r} - s\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}, v'\right)\right]ds\right\}.$$

Now for each $(\mathbf{r}',\mathbf{v}') \in \bar{D}\times V$, we define the function $\psi_{\mathbf{r}',\mathbf{v}'}$ by

$$\psi_{\mathbf{r}',\mathbf{v}'}(\mathbf{r},\mathbf{v}) \equiv G(\mathbf{r}',\mathbf{v}',\mathbf{r},\mathbf{v}).$$

Then by the "Dunford–Pettis" theorem,[49] L is a compact operator if $\Psi \equiv \{\psi_{\mathbf{r}',\mathbf{v}'} \mid (\mathbf{r}',\mathbf{v}') \in \bar{D}\times V\}$ is a "compact" subset of X. Equivalently, L is compact if $\Psi \subset X$ and every infinite sequence in Ψ possesses a Cauchy subsequence.

First we show that $\Psi \subset X$. Using (2.7), (2.12), and (3.2), we obtain

$$|\psi_{\mathbf{r}',\mathbf{v}'}(\mathbf{r},\mathbf{v})| \leq \frac{M_0 M_c \exp(lc_0)\exp\{-|\mathbf{r}-\mathbf{r}_0|[(\mathrm{Re}\lambda + \lambda^*)/v']\}}{v'|\mathbf{r}-\mathbf{r}'|^2 v^2} \tag{A1}$$

Now we integrate over $\mathbf{r}$ and $\mathbf{v}$ to get

$$\|\psi_{\mathbf{r}',\mathbf{v}'}\| \leq \frac{M_0 M_c \exp(lc_0)}{v'} 4\pi v_1 \times \int_{\Omega}\int_{t=0}^{d(\mathbf{r},\Omega)} \exp\left(-t\frac{\mathrm{Re}\lambda + \lambda^*}{v'}\right)dt\,d\Omega \leq (4\pi)^2 \frac{v_1 M_0 M_c \exp(lc_0)}{\mathrm{Re}\lambda + \lambda^*}.$$

Thus $\Psi \subset X$.

Next, we consider an infinite sequence in Ψ. If the various kernels and cross sections are continuous, as described by (i) of Sec. 2, then we select a subsequence $\psi_n \equiv \psi_{\mathbf{r}'_n,\mathbf{v}'_n}$ such that $(\mathbf{r}_n', \mathbf{v}_n') \to (\mathbf{r}_0', \mathbf{v}_0')$. Letting $\psi_0 \equiv \psi_{\mathbf{r}_0,\mathbf{v}_0}$, we have, for any $\epsilon > 0$,

$$\|\psi_0-\psi_n\| = \int_{|\mathbf{r}-\mathbf{r}_0|\le\epsilon}\int_{\mathbf{v}} |\psi_0-\psi_n|\,d\mathbf{v}d\mathbf{r} + \int_{|\mathbf{r}-\mathbf{r}_0|>\epsilon}\int_{\mathbf{v}} |\psi_0-\psi_n|\,d\mathbf{v}d\mathbf{r}.$$

By (A1), the first integral on the right side of this equation is $O(\epsilon)$. By the continuity of the various kernels in G, the second integral can be made $O(\epsilon)$ by requiring n to be sufficiently large. Thus $\|\psi_0-\psi_n\|\to 0$, proving that L is compact if condition (i) of Sec. 2 holds.

If condition (j) of Sec. 2 holds, then from any infinite sequence in Ψ we select an infinite subsequence $\psi_n \equiv \psi_{\mathbf{r}_n{}',\mathbf{v}_n{}'}$ such that $(\mathbf{r}_n{}', \mathbf{v}_n{}')\to(\mathbf{r}_0,\mathbf{v}_0)\in\bar{D}\times V$; for some k, $\mathbf{r}_n{}'\in D_K$, all n; and $v_n{}'$ is either a decreasing or an increasing sequence. Then, as a simple modification of the above proof for condition (i) will show, the sequence ψ_n is a Cauchy sequence because of the conditions imposed on the kernels and cross sections. Thus, the operator L is again compact. This proves the lemma for cases (i) and (j) of Sec. 2. QED

Lemma 5: For $\lambda\in\rho(T)$, $[K_1(\lambda I-T)^{-1}]^{M_1+1}$ is compact.

Proof: We write

$$\begin{aligned}[K_1(\lambda I-T)^{-1}]^{M_1+1} &= [K_c(\lambda I-T)^{-1}+K_d(\lambda I-T)^{-1}]^{M_1+1}\\ &= [K_c(\lambda I-T)^{-1}]^{M_1+1}+\cdots+K_c(\lambda I-T)^{-1}\\ &\quad\times[K_d(\lambda I-T)^{-1}]^{M_1}+[K_d(\lambda I-T)^{-1}]^{M_1+1}, \qquad \text{(A2)}\end{aligned}$$

where the dots refer to operators, all of which are products of $K_c(\lambda I-T)^{-1}$ and $K_d(\lambda I-T)^{-1}$, and all of which contain the product $K_d(\lambda I-T)^{-1}K_c$, which by Lemma 4 is compact. Also by Lemma 4, the first term on the right side of (A2) is compact. By Eq. (2.10) and the fact that $(\lambda I-T)^{-1}$ is a one-speed operator, $[K_d(\lambda I-T)^{-1}]^{M_1}=0$ so that the last two terms on the right side of (A2) are zero. Therefore, $[K_1(\lambda I-T)^{-1}]^{M_1+1}$ is a sum of compact operators, and hence is itself compact. QED

Lemma 6: $\|K\|=\sup_{\mathbf{r},v} v\Sigma_s(\mathbf{r},v)$, where $\Sigma_s(\mathbf{r},v)$ is the cross section for scattering plus fission (i.e., for collisions in which secondary neutrons are emitted).

Proof: In terms of the differential cross section $\Sigma_s(\mathbf{r},\mathbf{v}'\to\mathbf{v})$, we can write

$$(K\psi)(\mathbf{r},\mathbf{v}) = \int_{\mathbf{v}'} v'\Sigma_s(\mathbf{r},\mathbf{v}'\to\mathbf{v})\psi(\mathbf{r},\mathbf{v}')d\mathbf{v}',$$

and, by definition,

$$\Sigma_s(\mathbf{r},v') = \int_{\mathbf{v}}\Sigma_s(\mathbf{r},\mathbf{v}'\to\mathbf{v})\,d\mathbf{v}.$$

We combine these equations to obtain

$$\|K\psi\| \le \int_{\mathbf{r}}\int_{\mathbf{v}} v\Sigma_s(\mathbf{r},v)\,|\psi(\mathbf{r},\mathbf{v})|\,d\mathbf{r}d\mathbf{v}, \qquad \text{(A3)}$$

where equality holds for $\psi\ge 0$. From this equation, we get

$$\|K\| \le \sup_{\mathbf{r},v} v\Sigma_s(\mathbf{r},v). \qquad \text{(A4)}$$

However, equality holds in (A4) as can be seen by taking in (A3) a nonnegative sequence ψ_n which "converges" to the delta function at the point where $v\Sigma_s(\mathbf{r},v)$ "attains" its maximum. QED

Corollary: $\|K_0\|=\sup_{\mathbf{r},v} v\Sigma_0(\mathbf{r},v)$, where $\Sigma_0(\mathbf{r},v)$ is the cross section for low energy elastic collisions. Similarly, for fixed v, $\|K_0\|^0=\sup_{\mathbf{r}} v\Sigma_0(\mathbf{r},v)$.

ACKNOWLEDGMENTS

The authors wish to express their deep appreciation to Professors Basil Nicolaenko and R. L. Bowden for many stimulating discussions. In particular, Professor Nicolaenko suggested the use of a nilpotent operator to describe inelastic slowing down; such an operator has been included in only one previous paper.[43]

* The research for this paper was supported by the Air Force Office of Scientific Research under Grant No. AFOSR-71-2107 at New York University, and by the National Science Foundation under Grant GK-35903 at Virginia Polytechnic Institute. Part of this work was carried out while one of the authors (PFZ) was visiting the Rockefeller University and another part while he was visiting the Aspen Center for Physics.

[1] J. Lehner and G. M. Wing, Commun. Pure Appl. Math. **8**, 217 (1955); Duke Math. J. **23**, 125 (1956). See also G. M. Wing, *An Introduction to Neutron Transport Theory* (Wiley, New York, 1962).

[2] K. Jörgens, Commun. Pure Appl. Math. **11**, 219 (1958).

[3] G. Pimbley, J. Math. Mech. **8**, 837 (1959).

[4] R. Van Norton, Commun. Pure Appl. Math. **15**, 149 (1962).

[5] J. Lehner, J. Math. Mech. **11**, 173 (1962).

[6] N. Corngold, P. Michael, and W. Wollman, Nucl. Sci. Eng. **15**, 13 (1963).

[7] M. Nelkin, Physica **29**, 261 (1963).

[8] A. Leonard and T. W. Mullikan, J. Math. Phys. **5**, 399 (1964).

[9] Y. Shizuta, Prog. Theor. Phys. **32**, 489 (1964).

[10] N. Corngold, Nucl. Sci. Eng. **19**, 30 (1964).

[11] I. Kuščer and N. Corngold, Phys. Rev. **139**, A981 (1965); **140**, AB5 (1965).

[12] S. Albertoni and B. Montagnini, *Proceedings Symposium on Pulsed Neutron Research* (IAEA, Vienna, 1965), Vol. I, p. 239; J. Math. Anal. Appl. **13**, 19 (1966).

[13] R. Bednarz, *Pulsed Neutron Research* (IAEA, Vienna, 1965), Vol. I, p. 259.

[14] J. Mika, J. Math. Phys. **7**, 833, 839 (1966).

[15] S. Ukai, J. Math. Anal. Appl. **18**, 297 (1967).

[16] J. Mika, Nukleonik **9**, 303 (1967).

[17] M. Borysiewicz and J. Mika, *Proceedings Symposium on Neutron Thermalization and Reactor Spectra* (IAEA, Vienna, 1967), Vol. I, p. 45.

[18] I. Kuščer, Ref. 17, p. 3.

[19] Hans G. Kaper, J. Math. Anal. Appl. **19**, 207 (1967).

[20] J. Dorning, Nucl. Sci. Eng. **33**, 65 (1968).

[21] J. J. Duderstadt, Nucl. Sci. Eng. **33**, 119 (1968).

[22] I. Vidav, J. Math. Anal. Appl. **22**, 144 (1968).

[23] I. Kuščer and I. Vidav, J. Math. Anal. Appl. **25**, 80 (1969).

[24] N. Corngold, in *Transport Theory, SIAM-AMS Proceedings* (Am. Math. Soc., Providence, R.I., 1969), Vol. I, p. 79.

[25] J. J. Duderstadt, J. Math. Phys. **10**, 266 (1969).

[26] I. Kuščer, "Relaxation Problems in Neutron Transport and in the Kinetic Theory of Gases; a Comparison," in "Neutron Transport Theory Conference Report," ORO-3858-1 (1969), p. 380.

[27] M. Borysiewicz and J. Mika, J. Math. Anal. Appl. **26**, 461 (1969).

[28] G. Meister, Atomkernenergie **15**, 109 (1970).

[29] P. Silvennoinen and P. F. Zweifel, Nucl. Sci. Eng. **42**, 103 (1970).

[30] I. Vidav, J. Math. Anal. Appl. **30**, 264 (1970).

[31] B. Nicolaenko, "The Boltzmann Equation-Seminar 1970-71," edited by F. Alberto Grunbaüm, Courant Institute of Mathematical Sciences, New York University, New York (1970).

[32] J. Mika, J. Quant. Spect. Rad. Trans. **11**, 879 (1971).

[33] P. Nelson, J. Math. Anal. Appl. **35**, 90 (1971).

[34] R. Beauwens and J. Mika, Nukleonik **14**, 461 (1961).

[35] A. Sudaholc, J. Math. Anal. Appl. **35**, 1 (1971).

[36] R. Beauwens and J. Mika, Transport Theory Stat. Phys. **2**, 91 (1972).

[37] M. Ribaric, "Functional-Analytic Concepts and Structures of Neutron Transport Theory," Vol. I, Academia Scientiarum et Artium Slovenica, Ljubljana, Yugoslavia (1973).

[38] In Lemma 2 of Ref. 36 it is essentially claimed that if

$$\int_{0\leq|\mathbf{r}|<\infty} |f(\mathbf{r})|^2 d\mathbf{r} < \infty,$$

then

$$0 = \lim_{R\to\infty} \int_{|\mathbf{r}|=R} |f(\mathbf{r})|^2 (\mathbf{r}/|\mathbf{r}|)\, d\sigma.$$

However, if $\mathbf{r}=(r,\Theta,\phi)$ where Θ is the polar angle, then a counterexample is given by

$$f(r,\Theta,\phi)=\begin{cases}1 & n\leq r\leq n+1/n^3,\ n=1,2,\ldots;\quad \text{and } 0\leq\Theta\leq\pi/2\\ 0 & \text{otherwise.}\end{cases}$$

[39] The problem of true Bragg scattering, i.e., an elastic scattering kernel involving delta functions in angle, has never been treated, nor do we consider it in the present paper [cf. Eq. (2.11)]. We should point out, however, that our analysis does not assume that the elastic scattering rate is a continuous function of v. Then, our results are valid for the polycrystalline case discussed in Ref. 36.

[40] We adopt the notation of G. Bachman and L. Narici, *Functional Analysis* (Academic, New York, 1966). Thus, $\sigma(T)$, $P\sigma(T)$ respectively represent the spectrum and the point spectrum, and $R(T)$ represents the range of T.

[41] R. E. Edwards, *Functional Analysis* (Holt, Rinehart, and Winston, New York, 1965), pp. 679–681.

[42] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators* (Amer. Math. Soc., Providence, R.I., 1969), p. 21.

[43] E.W. Larsen and J.B. Keller, J. Math. Phys. 15, 75 (1974).

[44] K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison-Wesley, Reading, Mass., 1967).

[45] Our definition of λ^* differs from the usual definition given in the literature, where one finds λ^* defined as our $\hat{\lambda}(v_0)$. However, λ^* is also used in the literature to describe the edge of the "continuum" spectrum, and this is the meaning which λ^* will have in this paper. (See Theorem 2.)

[46] N. Dunford and J. T. Schwartz, *Linear Operators*, Part I (Interscience, New York, 1967), p. 624.

[47] A. Graffstein, T. Rzeszot, and E. Warda, J. Nucl. Eng. 22, 433 (1968); N. N. Hanna and M. J. Harris, J. Nucl. Eng. 22, 587 (1968); M. T. Rainbow and A. J. Ritchie, J. Nucl. Eng. 22, 735 (1968).

[48] E. Hewett and K. Stromberg, *Real Analysis* (Springer, New York, 1965).,

[49] N. Dunford and B. J. Pettis, Trans Am. Math. Soc. 47, 323 (1940), p. 370.

Dimensional renormalization

H. J. de Vega* and F. A. Schaposnik*

Departamento de Física, Facultad de Ciencias Exactas. Universidad de la Plata, C. C. 67-La Plata, Argentina

(Received 14 December 1972)

The analytic structure of a Feynman amplitude as a function of the number of space-time dimension (ν) is studied. A renormalization prescription by using ν as an analytic regularization parameter is given. It is shown its equivalence with BPH treatment for any graph in quantum field theory.

1. INTRODUCTION

Recently, the number of dimensions (ν) of the space has been proposed as an analytic regularization parameter.[1,2]

In particular, attention has been given to its applications to quantum electrodynamics[3] and gauge theories,[2] where the advantages of dimensional regularization over other methods became clear: For example, gauge invariance is explicitly maintained for any ν.

Another good feature of dimensional regularization is the fact that it simplifies considerably the calculations of divergent Feynman graphs; particularly the method is well suited to isolate their divergent parts. That isolation involves simply the calculation of the residue of a meromorphic function of ν at its pole in $\nu = 4$.

Because of these reasons, it seems to be important to give a renormalization program based on analytic continuation in the space-time dimension ν. In this context, the program has to include

(i) A renormalization prescription for an arbitrary Feynman graph. This means a set of rules for obtaining a finite expression, called renormalized amplitude, from a divergent Feynman integral.

(ii) A proof in which it is explicity shown that this renormalization belongs to the class defined by Bogoliubov-Parasiuk-Hepp.

The aim of this note is to give an additive renormalization prescription by using a single regularization parameter, the number of space-time dimensions, and to discuss its properties and its advantages.

Section 2 is devoted to the definition of the regularized amplitude and to state its analytical structure as a function of ν. In Sec. 3 the renormalization prescription for an arbitrary graph is stated and a proof is given of the equivalence between this method and the so-called BPH one.

2. ANALYTIC STRUCTURE OF A FEYNMAN AMPLITUDE AS A FUNCTION OF ν

We call $G(v_1, v_2, \ldots, v_n; \mathcal{L})$ a Feynman graph with vertices $v_1, v_2, \ldots, v_n$ and a set of lines $\mathcal{L}$. The corresponding amplitude is the formal expression

$$\mathcal{T}_G(x_1, x_2, \ldots, x_n; \nu) = \prod_{l \in \mathcal{L}} \Delta_l(x_{fl} - x_{il}), \qquad (2.1)$$

where the propagator Δ_l is

$$\Delta_l(x_{fl} - x_{il}) = Z_l\left(-i \frac{\partial}{\partial x_{fl}}\right) \Delta_F(x_{fl} - x_{il}), \qquad (2.2)$$

with $Z_l(-i\,\partial/\partial x)$ a polynomial of degree r_l and $\Delta_F(x)$ the scalar Feynman propagator which can be written in an arbitrary ν-dimensional space-time as

$$\Delta_F(x) = \frac{1}{(4\pi i)^{\nu/2}} \int_0^\infty \frac{d\alpha}{\alpha^{\nu/2}} \exp\left(-i(m^2 - i0)\alpha - \frac{ix^2}{4\alpha}\right). \qquad (2.3)$$

Then, it is easy to show that the propagator $\Delta_l(x)$ can be written in the form

$$\Delta_l(x_{fl} - x_{il}) = \frac{1}{(4\pi i)^{\nu/2}} Z_l\left(\frac{1}{2}\, \bar{e}^{(l)} \frac{\bar{\partial}}{\partial s_l}\right) \cdot \int_0^\infty \frac{d\alpha_l}{\alpha_l^{\nu/2}} \times \exp\left(-i\alpha_l(m_l^2 - i0) + \frac{i}{4\alpha_l} (\bar{x} - \bar{s}_l)^t Q^{(l)} (\bar{x} - \bar{s}_l)\right)\Big|_{\bar{s}_l = \bar{0}}, \qquad (2.4)$$

where

(i) $\bar{x} \equiv (x_1, x_2, \ldots, x_n), \quad \bar{s}_l \equiv (s_1^{(l)}, s_2^{(l)}, \ldots, s_n^{(l)}),$

$x_1, x_2, \ldots, x_n, s_1^{(l)}, s_2^{(l)} \ldots, s_n^{(l)}$ are all ν-vectors.

(ii) $\bar{e}^{(l)} = (e_1^{(l)}, e_2^{(l)}, \ldots, e_n^{(l)})$, with

$$e_j^{(l)} = \begin{cases} 1, & \text{if } v_j = v_{fl}, \\ -1, & \text{if } v_j = v_{il}, \\ 0, & \text{otherwise,} \end{cases}$$

are the elements of the incidence matrix of the graph G:

(iii) $Q^{(l)} = (Q_{jr}^{(l)}), \quad Q_{jr}^{(l)} = e_j^{(l)} e_r^{(l)}.$

With the aid of Eq. (2.4), one obtains the following expression for the Fourier transform of the amplitude (2.1):

$$\begin{aligned}\tilde{\mathcal{T}}_G(\bar{p}; \nu) &= \frac{1}{(2\pi)^{\nu n}} \int e^{-i\bar{p}\cdot\bar{x}}\, \mathcal{T}_G(\bar{x}, \nu)\, d^\nu x_1 \cdots d^\nu x_n \\ &= \lambda\delta\left(\sum_{j=1}^n p_j\right) \prod_{l\in\mathcal{L}} Z_l\left(\frac{i}{2}\, \bar{e}^{(l)} \cdot \frac{\bar{\partial}}{\partial s_l}\right) \int_0^\infty \int_0^\infty \cdots \int_0^\infty \prod_{l\in\mathcal{L}} d\alpha_l\, \mathbb{C}(\alpha)^{-\nu/2} \exp\left[i\left(\bar{p} + \tfrac{1}{2} \sum_{l\in\mathcal{L}} \frac{Q^{(l)}\bar{s}_l}{\alpha_l}\right)^t\right. \\ &\quad \left.\times A_E^{-1}\left(\bar{p} + \tfrac{1}{2} \sum_{l\in\mathcal{L}} \frac{Q^{(l)}\bar{s}_l}{\alpha_l}\right) + i \sum_{l\in\mathcal{L}} \left(\frac{\bar{s}_l^t Q^{(l)} \bar{s}_l}{4\alpha_l} - (m_l^2 - i0)\alpha_l\right)\right],\end{aligned} \qquad (2.5)$$

where

$$\lambda = \frac{(i\pi)^{\nu(1-n-L)/2}}{2^{\nu L}\, i^{1-n}}, \quad \Delta_{ij} = -\sum_{l\in\mathcal{L}} \frac{Q_{ij}^{(l)}}{\alpha_l}, \quad \mathbb{C}(\alpha) = \prod_{l\in\mathcal{L}} \alpha_l \det A_E.$$

A_E is a matrix of order $n-1$ obtained from (A_{ij}) supressing the row and column k, $1 \leqslant k \leqslant m$.

After having performed explicitly the operations implied by Eq. (2.5), the rhs of it can be interpreted as an analytic continuation of the Feynman amplitude for complex ν.

The properties of $\mathbb{C}(\alpha)$ and A_E^{-1} were studied in detail in Ref. 6. Following the lines developed there it can be shown that[7]

(a) The rhs of Eq. (2.5) defines $\tilde{\mathcal{T}}_G(\bar{p};\nu)$ as an analytic function of ν for $\mathrm{Re}(\nu)$ sufficiently large and negative.

(b) One can continue $\tilde{\mathcal{T}}_G(\bar{p};\nu)$ to the entire ν plane as a meromorphic function of ν, finding that it has isolated poles for $\nu = n$ (n integer, $n \geqslant m$, m depending on the diagram).

$\tilde{\mathcal{T}}_G(\bar{p};\nu)$ can be decomposed:[6]

$$\tilde{\mathcal{T}}_G(p;\nu) = \sum_{\epsilon} \tilde{\mathcal{T}}_G(\bar{p};\nu,\epsilon), \tag{2.6}$$

the sum extended over all s-families ϵ. Here $\mathcal{T}_G(\bar{p};\nu,\epsilon)$ is such that

$$\tilde{\mathcal{T}}_G(\bar{p};\nu,\epsilon)\cdot\Bigl(\prod_{G\in\epsilon} \Gamma(\{2-\nu/2\}N(G') - [\omega(G')/2])\Bigr)^{-1} \tag{2.7}$$

is an entire function of ν. $N(G')$ is the number of independent loops of the graph G' and

$$w(G') = w(G'(v_1, v_2, \ldots, v_{n'};\mathcal{L})) = \sum_{\text{conn}} (r_l + 2) - 4(n'+4),$$

where $\sum_{\text{conn}}$ extends over all $l \in \mathcal{L}$ which connect two vertices from the set $\{v_1, v_2, \ldots, v_{n'}\}$. We will say that $G(v_n, \ldots, v_n;\mathcal{L})$ is "divergent" if $w(G) \geqslant 0$.

3. RENORMALIZATION

We define the L operation on a meromorphic function of ν, $F(p_1, \ldots, p_m;\nu)$, as the principal part of $F(\bar{p};\nu)$, Laurent's expansion around the point $\nu = 4$.

In order to state our renormalization prescription we define vertex parts in an analogous way as in the BPH method.[4] Let $\{v_1', v_2', \ldots, v_m'\} \subset \{v_1, v_2, \ldots, v_n\}$. Then[8]

$$\mathbf{X}_{\mathcal{L}}(v_1', \ldots, v_m') = \begin{cases} 1 & \text{if } m = 1, \\ 0 & \text{if } G(v_1', \ldots, v_m') \text{ is IPR}, \\ -L & [\bar{\mathbf{R}}_{\mathcal{L}}(v_1', \ldots, v_m')] \quad \text{otherwise}, \end{cases} \tag{3.1}$$

$$\bar{\mathbf{R}}_{\mathcal{L}}(v_1', \ldots, v_m') \equiv \sum_{p} \prod_{j=1}^{k(p)} \mathbf{X}_{\mathcal{L}}(v_{j_1}^p, \ldots, v_{jr(j)}^p) \prod_{\text{conn}} \Delta_l . \tag{3.2}$$

Here $\sum_p$ extends over all partitions of $\{v_1', \ldots, v_m'\}$ into $2 \leqslant k(p) \leqslant m$ sets and $\prod_{\text{conn}}$ is taken over all $l \in \mathcal{L}$ which connect different sets of the partition.

Finally we define the renormalized amplitude as

$$\mathbf{R}_{\mathcal{L}}(v_1, \ldots, v_n) = \bar{\mathbf{R}}_{\mathcal{L}}(v_1, \ldots, v_n) + \mathbf{X}_{\mathcal{L}}(v_1, \ldots, v_n). \tag{3.3}$$

The finiteness of $\mathbf{R}_{\mathcal{L}}$ for $\nu = 4$ can be easily checked.

We see that if in Eq. (3.1) we replace the L operation by the M operation defined by BPH as

$$M\Bigl[\delta\Bigl(\sum_{i=1}^{m} p_i'\Bigr)F(p_1', \ldots, p_m';\nu)\Bigr] = \delta\Bigl(\sum_{i=1} p_i'\Bigr)\, T(p_1', \ldots, p_m';\nu), \tag{3.4}$$

where $T(\bar{p};\nu)$ is the Taylor series of $F(\bar{p};\nu)$ around $p_1' = p_2' = \cdots = p_m' = 0$ up to the order $w(v_1', \ldots, v_m')$, we arrive at the conventional BPH prescription,[4] with ν as a regulator parameter:

$$\mathbf{X}_{\mathcal{L}}(v_1', \ldots, \nu_m') = \begin{cases} 1 & \text{if } m = 1, \\ 0 & \text{if } G(v_1', \ldots, \nu_m';\mathcal{L}) \text{ is IPR}, \\ -M'[\bar{R}_{\mathcal{L}}(v_1', \ldots, v_m')] & \\ \equiv -M[\bar{R}_{\mathcal{L}}(v_1', \ldots, v_m') & \\ +\hat{X}_{\mathcal{L}}(v_1', \ldots, v_m') & \text{otherwise}, \end{cases} \tag{3.5}$$

$\hat{X}_{\mathcal{L}}(v_1', \ldots, v_m')$ is in p-space a polynomial of degree $\leqslant \omega(v_1', \ldots, v_m')$ with finite coefficients in $\nu = 4$:

$$\bar{R}_{\mathcal{L}}(v_1', \ldots, v_m') = \sum_{p} \prod_{j=1}^{k(p)} X_{\mathcal{L}}(v_{j_1}^p, \ldots, v_{jr(j)}^p) \prod_{\text{conn}} \Delta_l , \tag{3.6}$$

$$R_{\mathcal{L}}(v_1, \ldots, v_n) = \bar{R}_{\mathcal{L}}(v_1, \ldots, v_n) + X_{\mathcal{L}}(v_1, \ldots, v_n). \tag{3.7}$$

Now we will show that the vertex parts defined by Eq. (3.1) are (in p-space) polynomials of degree smaller than or equal to $w(v_1', \ldots, v_m')$. We will also prove that the renormalized amplitudes defined by Eqs. (3.1)–(3.3) belong to the class defined by Bogoliubov, Parasiuk, and Hepp.

If one replaces repeatedly Eqs. (3.5) and (3.6) in Eq. (3.7) one arrives to an expression of $R_{\mathcal{L}}$ which only contains M' operations and propagators.

Every term contains a certain number of M' operations. This number can take values from zero to a maximum, $S_{\mathcal{L}}(v_1', \ldots, v_m')$, characteristic of the diagram. Our proof will be done by induction on $S_{\mathcal{L}}(v_1, \ldots, v_n)$.

(i) $S_{\mathcal{L}}(v_1, \ldots, v_n) = 1$. The most general diagram with $S_{\mathcal{L}}(v_1, \ldots, v_n) = 1$ is the one which has a certain number ($\geqslant 1$) of divergent subgraphs. These subgraphs have only a superficial divergence. In this case

$$R_{\mathcal{L}}(v_1, \ldots, v_n) = \sum_{p} \prod_{j=1}^{k(p)} \Bigl(-M'\Bigl[\prod_{l'\in\mathcal{L}_i} \Delta_{l'}\Bigr]\Bigr) \prod_{\text{conn}} \Delta_l , \tag{3.8}$$

$\sum_P$ runs over all partitions P of $\{v_1, \ldots, v_n\}$ in $1 \leqslant k(p) \leqslant n$ sets and $\mathcal{L}_j$ is the set of lines connecting $\{v_{j_1}^p, \ldots, v_{jr(j)}^p\}$ provided they make the corresponding subgraph $G(v_{j_1}^p, \ldots, v_{jr(j)}^p;\mathcal{L}_j)$ IPI and divergent.

From Sec. 2 we know that $\prod_{l\in\mathcal{L}_j} \Delta_l$ has a simple pole at $\nu = 4$; then

$$\begin{aligned} R_{\mathcal{L}}(v_{j_1}^p, \ldots, v_{jr(j)}^p) &= \prod_{l\in\mathcal{L}_j} \Delta_l - M'\Bigl[\prod_{l\in\mathcal{L}_j} \Delta_l\Bigr] \\ &= \prod_{l\in\mathcal{L}_j} \Delta_l + \sum_{i=-1}^{\infty} (\nu-4)^i f_{i+1}(p_j, \ldots, p_{jr(j)}) \\ &= \delta\Bigl(\sum_{k=1}^{r(j)} p_{j_k}\Bigr). \end{aligned} \tag{3.9}$$

The lhs is an analytic function of ν in some neighborhood of $\nu = 4$, as can be seen by modifying the proofs given by Hepp.[4] The $f_{i+1}(\bar{p})$ are polynomials in its arguments of degree $\leqslant w(v_{j_1}^p, \ldots, v_{jr(j)}^p)$. Performing the $(-L)$ operation on both sides of Eq. (3.9), we obtain

$$O = \mathbf{X}_{\mathcal{L}}(v_{j_1}^p, \ldots, v_{jr(j)}^p) - \delta\Bigl(\sum_{k=1}^{r(j)} p_{jk}\Bigr) f_0(p_{j_1}, \ldots, p_{jr(j)})/\nu - 4.$$

Choosing

$$\hat{X}_{\mathcal{L}}(v^p_{j_1},\ldots,v^p_{jr(j)}) = \mathbf{X}_{\mathcal{L}}(v^p_{j_1},\ldots,v^p_{jr(j)}) + M\Big(\prod_{l\in\mathcal{L}_j}\Delta_l\Big), \tag{3.10}$$

we see that

$$R_{\mathcal{L}}(v_1,\ldots,v_n) = \mathbf{R}_{\mathcal{L}}(v_1,\ldots,v_n). \tag{3.11}$$

(ii) We assume the validity of the induction hypothesis for all graphs $G(v_1,\ldots,v_r;\mathcal{M})$ with $S_{\mathcal{M}}(v_1,\ldots,v_r) \leqslant K$. Let $G(v_1,\ldots,v_n;\mathcal{L})$ be a graph with $S_{\mathcal{L}}(v_1,\ldots,v_n) = K+1$. The most general graph of this kind has a certain number ($\geqslant 1$) of divergent IPI subgraphs, $G(v_{q1},\ldots,v_{qr(q)};\mathcal{L})$, each one with $S_{\mathcal{L}}(v_{q_1},\ldots,v_{qr(q)}) = K+1$,

$$\bar{R}_{\mathcal{L}}(v_{q_1},\ldots,v_{qr(q)}) = \sum_p \prod_{i=1}^{k(p)} X_{\mathcal{L}}(v^p_{q_{i1}},\ldots,v^p_{qir(i)}) \prod_{\text{conn}}\Delta_l. \tag{3.12}$$

Here $\sum_p$ extends over all partitions of the set $\{v_{q1},\ldots,v_{qr(q)}\}$ in $2 \leqslant k(p) \leqslant r(q)$ subsets $\{v^p_{q\,i1},\ldots,v_{qir(i)}\}$. Any $X_{\mathcal{L}}(v^p_{q_1},\ldots,v^p_{qr(q)})$ is obtained by applying the $(-M')$ operation to (3.12). Then, any $X_{\mathcal{L}}(v^p_{qi_1},\ldots,v^p_{qir(i)})$ of the rhs of Eq. (3.12) has at most K M'-operations and the induction hypothesis applies, and so it is possible to replace those $X_{\mathcal{L}}$ by $\mathbf{X}_{\mathcal{L}}$ –it only means to add a finite renormalization:

$$\bar{R}_{\mathcal{L}}(v_{q_1},\ldots,v_{qr(q)}) = \bar{\mathbf{R}}_{\mathcal{L}}(v_{q_1},\ldots,v_{qr(q)} + X_{\mathcal{L}}(v_{q_1},\ldots,v_{qr(q)}). \tag{3.13}$$

Performing in both sides the $(-L)$ operation, we obtain

$$O = \mathbf{X}_{\mathcal{L}}(v_{q_1},\ldots,\nu_{qr(q)}) - \sum_{i=-d}^{-1}(\nu-4)^i f_{i+1}(p_{q_1},\ldots,p_{qr(q)}) \times \delta\Big(\sum_{m=1}^{r(q)} p_{q_m}\Big), \tag{3.14}$$

where we have made use of the analyticity of the lhs of Eq. (3.13), as can be seen by modifying the proofs given in Ref. 4. We see from Eq. (3.14) that the $\mathbf{X}_{\mathcal{L}}(v_{q_1},\ldots,v_{qr(q)})$ are polynomials of degree $\leqslant w(v_{q_1},\ldots,v_{qr(q)})$ in their arguments.

By choosing $\hat{X}_{\mathcal{L}}$ as in Eq. (3.10)

$$\hat{X}\;(v_{q_1},\ldots,v_{qr(q)}) = -\sum_0^{\infty}(\nu-4)^i f_{i+1}(p_{q_1},\ldots,p_{qr(q)}) \times \delta\Big(\sum_{m=1}^{r(q)} p_{q_m}\Big), \tag{3.15}$$

it is true that

$$R_{\mathcal{L}}(v_1,\ldots,v_n) = \mathbf{R}_{\mathcal{L}}(v_1,\ldots,v_n). \tag{3.16}$$

We have then showed the polynomial character of the $X_{\mathcal{L}}$ and the equality between BPH and dimensional renormalization provided one makes the election of a particular finite renormalization.

We conclude by making some remarks about the dimensional renormalization method developed in this note.

The renormalization program stated in the introduction was carried out without difficulties by using a single parameter as a regulator: the space–time dimension (ν). This approach simplifies considerably the calculations providing a general method for performing Feynman integrals in quantum field theories.

Due to its equivalence with BPH treatment (explicitly shown in this section) it follows that the renormalized amplitude can be obtained from an integration Lagrangian with suitable counterterms.

As it was pointed out in the Introduction, the application of dimensional renormalization to a wide variety of problems in quantum field theories, in particular to gauge theories[9] and to quantum electrodynamics, shows its natural advantages.

ACKNOWLEDGMENTS

We are gratefully indebted to Professor C. G. Bollini and Professor J. J. Giambiagi who suggested the idea and under whose advice this work was done.

*Fellow of the Consejo Nacional de Investigaciones Científicas y Técnicas.

[1] C. G. Bollini and J. J. Giambiagi, Phys. Lett. B 40, 566 (1972).

[2] G. 't Hooft and M. Veltman, preprint, Utrecht, 1972 (to be published in Nucl. Phys. B).

[3] C. G. Bollini and J. J. Giambiagi, preprint, Univ. La Plata, Feb. 1972 (to be published in Nuovo Cimento Lett.), J. F. Ashmore, Nuovo Cimento Lett. 4, 289 (1972).

[4] K. Hepp, Commun. Math. Phys. 2, 301 (1966).

[5] N. N. Bogoliubov and O. S. Parasiuk, Acta Math. 97, 227 (1957).

[6] E. Speer, J. Math. Phys. 9, 1404 (1968); *Generalized Feynman Amplitudes* (Princeton U.P., Princeton, N.J., 1969).

[7] From here on we refer to Ref. 6 for notation.

[8] Expressions like (3.1) and subsequents—defined in a ν-dimension space—must be understood in a formal sense for complex ν, bearing in mind the analytic continuation of Feynman amplitudes discussed in Sec. 2.

[9] As it has been stressed by 't Hooft and Veltman in Ref. 2.

Bounds for effective thermal, electrical, and magnetic properties of heterogeneous materials using high order statistical information*

Mostafa A. Elsayed†

Towne School, The University of Pennsylvania, Philadelphia, Pennsylvania

(Received 30 March 1973)

Bounds have been developed for the effective thermal, electrical, and magnetic properties (K^*) of a two phase statistically homogeneous and isotropic material. These bounds came in terms of the ratio of the properties of the two phases, volume fractions, and the constants G, G_2, G_3, M_1, and M_2. Bounds on the values of these constants are obtained. It is shown that the constants G, G_2, and G_3 are geometric parameters. They are calculated in the general case of spheroidal inclusions in terms of the axial ratio A. The constants M_1 and M_2 are packing parameters. General expressions of these constants in terms of packing information are obtained. It is found that certain combinations of the constants M_1 and M_2 for spheres and plates give exact solutions for the effective property K^*. The exact solutions were shown in some cases to be equal to Miller's upper bound and in other cases to Miller's lower bound. A self-consistent scheme for the case of spheres is carried out. The corresponding values of M_1 and M_2 were identified and used to plot the bounds. The bounds for plates are also calculated. It is found that these bounds introduce a great improvement over Miller's bounds. The small perturbation limit was considered. It is found that our bounds coincide to order η^5 and all values of v ($\eta = K_1/K_2 - 1$, K_i is the property of material i). Moreover, they include Miller's bounds to order η^3 and Hashin bounds to order η^2.

INTRODUCTION

The effective physical properties of two phase composites are determined by the volume fraction of the inclusions, shapes of inclusions, and the way the inclusions are placed within the matrix. The various possibilities are considerable. Complete knowledge of the functional $P[E_i(\mathbf{x}), k_{lm}(\mathbf{x})]$, where $P[E_i(\mathbf{x}), k_{lm}(\mathbf{x})]dE_1(\mathbf{x})\ldots dE_3(\mathbf{x})dk_{11}(\mathbf{x})\ldots dk_{33}(\mathbf{x})$ is defined as the probability of the realization of the particular joint field $[E_i(\mathbf{x}), k_{lm}(\mathbf{x})]$, is required to determine the effective property of the material.

Although, in this paper, we will be speaking of the "effective thermal conductivity," the mathematical formulation given applies to the physical subjects of electrical conduction, electrostatics, and magnetostatics. In Appendix A, the physical interpretation of the various quantities defined in each of these subjects is given.[1]

Although complete statistical information of the inclusion and matrix geometry is required for an exact prediction of an effective property, it can be shown that rigorous and exact bounds can be applied to the effective property, which require only a limited amount of statistical information. These bounds are obtained by first expressing the desired effective property in terms of the functional of a variational principle (i.e., a maximum or minimum principle) that governs the phenomenon of interest. Upon substituting an allowable trial function in the variational principle functional, one achieves the desired bound. In this way, bounds have been developed by Hashin and Shtrikman[2] and Hashin[3] which give bounds for thermal and elastic effective constants for statistically homogeneous and isotropic materials. These bounds are expressed in terms of the volume fraction, which is the simplest statistical information one can obtain for a material. The bounds were useful to predict the conductivity of the material when the ratio of inclusion to matrix constants is not too large or when the volume fraction of the inclusions (v_1) is very small. Subsequently, Beran and Molyneux[4,5] developed bounds for the effective thermal conductivity k^* of a statistically homogeneous and isotropic material. For a two-phase material, these bounds were in terms of the volume fraction of the inclusion, v_1, and a three-point correlation function of the form

$$R_3 = \langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')\rangle,$$

where the brackets $\langle\ \rangle$ indicate ensemble averaging. The method was to derive the upper bound from the principle of minimum potential energy and the lower bound from the principle of minimum complementary energy. The trial solution used was a perturbation function ($|k_1 - k_m| \ll k^*$, k_1 being the inclusion conductivity, k_m the matrix conductivity) of the diffusion equation modified by undetermined multipliers. The function R_3 was difficult to use in practical applications, although Corson[6] has measured this function for Pb-Al and Pb-Fe mixtures. To circumvent this practical difficulty, Miller[7] introduced a random cell model of wide applicability and was able to derive the following bounds:

$$k_l \leqslant k^* \leqslant k_U,$$

where

$$\frac{k_U}{(k_1k_M)^{1/2}} = \frac{1}{\alpha^{1/2}}\left(1 + v_1(\alpha - 1) - \frac{\frac{1}{3}v_1v_M(\alpha-1)^2}{1 + (\alpha - 1)[v_1 + 3(v_M^2G_1 - v_1^2G_M)]}\right), \tag{1}$$

and

$$\frac{k_l}{(k_1k_M)^{1/2}} = \frac{\alpha^{1/2}}{[\alpha - v_1(\alpha - 1)] - \frac{4}{3}(1-\alpha)^2v_1v_M/1 + \alpha + 3(\alpha - 1)(v_1^2G_M - v_M^2G_1)}. \tag{2}$$

Here, the subscript 1 indicates the inclusion and the subcript M the matrix, and $\alpha = k_1/k_M$; $k_1 \geqslant k_M$, G_1, and G_M are geometrical parameters which lie between $\frac{1}{9}$ and $\frac{1}{3}$. When $G_1 = \frac{1}{9}$, the inclusion geometry is spherical and when $G_1 = \frac{1}{3}$, the inclusion geometry is plate like.

The results have been extended to thermal and elastic properties of fiber reinforced materials by Beran and Silnutzer[8] and Silnutzer.[9] The bounds in these later cases include two constants representing geometric information. Each of these constants lies between $\frac{1}{4}$ and $\frac{1}{2}$. The value $\frac{1}{4}$ corresponds to a circular shape and the value $\frac{1}{2}$ to a parallel lamella shape.

The shape information in Miller's bounds reduced the spread of Hashin bounds by at least a factor of 2 for all volume fractions. At low volume fractions (below about 10%) it yielded bounds that were very close, even for inclusion to matrix conductivity ratios as high as 100. When α is large and $v_1 > 10\%$, we find that the bounds for k^* are not very restrictive and cannot be used to predict the value of k^*. What is needed is some information in the bounds to indicate how the inclusions are packed within the matrix. This packing information can be introduced by extending the trial functions used in deriving the bounds to include higher orders. This will lead to correlation functions of orders higher than 3 appearing in the bounds. These–as we shall see–are the functions that include packing information. The new bounds will be more restrictive and could be used for the prediction of k^* for higher values of α and v_1.

In Sec. I of this paper, we develop the bounds. The upper bound is derived from the principle of minimum potential energy and the lower bound from the principle of minimum complementary energy. The bounds come in terms of the volume fraction, the conductivity of each of the two phases and the constants G, G_2, G_3, M_1, and M_2. The small perturbation case ($\eta = k_1/k_2 - 1 \ll 1$) is presented in Sec. II. It is shown that the bounds coincide to order η^5 while Miller's bounds coincide to order η^3 and Hashin bounds to order η^2. In Sec. III we obtain bounds on the values of the constants G, G_2, G_3, M_1, and M_2. These bounds are obtained by comparing our bounds to Miller's bounds and/or imposing the requirement that the bounds are positive and finite. The significance of the constants is discussed in Sec. IV. We show that the constants G, G_2, and G_3 represent shape information and their values are obtained for the general case of a spheroid. On the other hand, it is shown that the constants M_1 and M_2 represent packing information. In order to motivate a choice for possible values of packing parameters, we introduce the self-consistent scheme in Sec. V. We use these values in the case of spherical inclusions to compare our bounds to Miller's bounds for spheres.

It is also shown that introducing the packing information in the bounds enables us to recommend one shape of inclusions over the other in order to obtain higher conductivities for the same volume fractions.

I. DEVELOPMENT OF THE BOUNDS FOR THE EFFECTIVE PROPERTIES

A. Variational principles

The bounds will be derived from the following variational principles.[4]

(a) The integral

$$U = \frac{1}{2V} \int_V kE_iE_i d\mathbf{x} \tag{3}$$

(E_i being the ith component of the temperature gradient vector or electric field, etc.) subject to the subsidiary condition

$$\delta_{ijK} \frac{\partial E_K}{\partial x_j} = 0 \tag{4}$$

is stationary for

$$\frac{\partial}{\partial x_j} kE_j = 0 \tag{5}$$

Here, k is the thermal conductivity, and V is the volume of the medium under consideration. We suppose that V is enclosed by a surface S. The curl condition implies that $\mathbf{E} = \partial\phi/\partial xi$, where ϕ is a scalar field (temperature, or electric potential, etc.). The variational principle requires the boundary condition; $\delta\phi(S) = 0$, where $\delta\phi(S)$ is the variation of ϕ on S.

(b) The integral

$$U = \frac{1}{2V} \int_V \frac{DiDi}{k} d\mathbf{x}, \tag{6}$$

being the heat flux (or electric displacement kE_i) subject to the subsidiary condition

$$\frac{\partial}{\partial x_j} D_j = 0 \tag{7}$$

is stationary for

$$\delta_{ijk} \frac{\partial}{\partial x_j} \left(\frac{D_k}{k}\right) = 0. \tag{8}$$

The divergence condition implies $D_i = \delta_{ijk}(\partial/\partial x_j)A_k$, and as a boundary condition, we impose the requirement that the variation of $\mathbf{A}$ on S, $\delta\mathbf{A}(S)$ is zero.

U is the total energy of the field, and it may be shown that U is minimum U_0, when the variation δU is set equal to zero.

The effective conductivity can be defined for a statistically homogeneous and isotropic medium, by

$$U_0 = \tfrac{1}{2}k^*\overline{E}_i\overline{E}_i = \tfrac{1}{2}\overline{D}_i\overline{D}_i/k^*, \tag{9}$$

where

$$\overline{E}_i = \frac{1}{V} \int_V E_i(\mathbf{x}) d\mathbf{x} \tag{10}$$

is the volume average of E_i. An upper bound on k^* is derived by the use of principle (a) by introducing an allowable trial function for E_i [i.e., a function that satisfies Eq. (4)and the boundary conditions] into the right-hand side of Eq. (3) and equating this to the first of the two expressions for U_0 that is given by Eq. (9). We note that the specified boundary conditions uniquely determine $\overline{E}_i$. A lower bound on k^* is derived by the use of principle (b) by introducing an allowable trial function for $\overline{D}_i$ [i.e., a function that satisfies Eq. (7) and the boundary conditions] into the right-hand side of Eq. (6) and equating this to the second of the two expressions for U that is given by Eq. (9).

For an explicitly statistical interpretation of the problem use is made of an ergodic hypothesis[10-12] and volume averaging is interpreted as ensemble averaging.

B. Upper bound

Our choice of a trial function to use in the first variational principle is motivated by a perturbation solution of Eq. (5) in powers of $k'(\mathbf{x})$ where

$$k(\mathbf{x}) = \bar{k}(\mathbf{x}) + k'(\mathbf{x}). \tag{11}$$

Accordingly we shall express E_i in a series of the form

$$E_i(\mathbf{x}) = \sum_{n=0}^{\infty} \overset{(n)}{E}_i(\mathbf{x}) \tag{12}$$

where $\overset{(n)}{E}_i$ is of order $(k'/\bar{k})^n\bar{E}_3$. We assume that the material is subjected to a constant temperature gradient (or electric field) in the 3 direction, i.e.,

$$\bar{E}_i = \bar{E}_3\delta_{i3}. \tag{13}$$

Convergence is assumed if $\overline{k'^n}/\bar{k}^n < 1$ for all n (Ref. 13). The bar on the top is used to indicate ensemble averaging. In Eq. (12), $E_i^{(0)}$ is a constant.

Substituting the form of E_i, Eq. (12) in Eqs. (5) and (4), we have

$$\bar{k}\frac{\partial}{\partial x_i}\sum_{n=1}^{\infty}\overset{(n)}{E}_i + \overset{(0)}{E}_i\frac{\partial}{\partial x_i}k'(\mathbf{x}) + \frac{\partial}{\partial x_i}\left(k'(\mathbf{x})\sum_{n=1}^{\infty}\overset{(n)}{E}_i\right) = 0 \tag{14}$$

and

$$\delta_{ijk}\frac{\partial}{\partial x_j}\left(\sum_{n=1}^{\infty}\overset{(n)}{E}_k\right) = 0. \tag{15}$$

In Eqs. (14) and (15) terms of the first order ($n = 1$) give

$$\bar{k}\frac{\partial}{\partial x_i}\overset{(1)}{E}_i + \overset{(0)}{E}_i\frac{\partial}{\partial x_i}k'(\mathbf{x}) = 0 \tag{16}$$

and

$$\delta_{ijk}\frac{\partial}{\partial x_j}\overset{(1)}{E}_k = 0. \tag{17}$$

The solution of Eq. (16) subject to the condition in Eq. (17), using the free space Green's function ($1/r$) is

$$\overset{(1)}{E}_i(\mathbf{x}) = \frac{-\overset{(0)}{E}_j}{4\pi\bar{k}}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}k'(\mathbf{x}')\frac{r_i}{r^3}d\mathbf{x}', \tag{18}$$

where $\mathbf{r} = \mathbf{x} - \mathbf{x}'$. Considering terms of the second order in Eqs. (14) and (15), and solving the resulting equations, we find

$$\overset{(2)}{E}_i(\mathbf{x}) = \frac{-1}{4\pi\bar{k}}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}k'(\mathbf{x}')\overset{(1)}{E}_j(\mathbf{x}')\frac{r_i}{r^3}d\mathbf{x}'. \tag{19}$$

Using the expression for $\overset{(1)}{E}_i$ given by Eq. (18) we have

$$\overset{(2)}{E}_i(\mathbf{x}) = \frac{\overset{(0)}{E}_k}{(4\pi\bar{k})^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}\left(\int_{\mathbf{x}''}k'(\mathbf{x}')k'(\mathbf{x}'')\frac{(x'_j - x''_j)}{|\mathbf{x}' - \mathbf{x}''|^3}dx''\right) \times \frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3}d\mathbf{x}'. \tag{20}$$

Ensemble-averaging both sides of Eqs. (18) and (20), we find that $E_i^{\overline{(1)}} = E_i^{\overline{(2)}} = 0$. Similarly, we can—in general—show that

$$\overset{(\bar{n})}{E}_i = 0, \quad n \geqslant 1. \tag{21}$$

Imposing the condition given by Eq. (13) on Eq. (12), we find $E_i^{(0)} = \bar{E}_3\delta_{i3}$, and hence in Eq. (12) we find

$$E_i = \bar{E}_3\delta_{i3} + \sum_{n=1}^{\infty}\overset{(n)}{E}_i. \tag{22}$$

In order to obtain the upper bound, we introduce the trial function

$${}_i(\mathbf{x}) = \bar{E}_3\delta_{i3} + \lambda_1\overset{(1)}{E}_i(\mathbf{x}) + \lambda_2\overset{(2)}{E}_i(\mathbf{x}) + \cdots + \lambda_N\overset{(N)}{E}_i(\mathbf{x}). \tag{23}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_N$ are modifying multipliers.

Here ${}_NE_i(\mathbf{x})$ is equal to $E_i(\mathbf{x})$ if the $\lambda_k = 1$, $k = 1, 2, \ldots$, and $N \to \infty$. Here we emphasize that the trial function presented above satisfies the conditions required for an allowable trial function exactly. There is no approximation in the bounds and their validity does not rest on any convergence requirements.

From the above mentioned variational principles and Eqs. (3) and (9) we may write

$$k^*\bar{E}_3^2 \leqslant \langle kE_iE_i\rangle, \tag{24}$$

where $E_i(\mathbf{x})$ is any trial function. Here we shall consider the form given in Eq. (23). For $N = 2$, Eq. (24) gives

$$k^*\bar{E}_3^2 \leqslant \bar{k}\bar{E}_3^2 + \bar{k}\lambda_1^2\langle\overset{(1)}{E}_i\overset{(1)}{E}_i\rangle + 2\lambda_1\bar{E}_3\langle k'\overset{(1)}{E}_3\rangle + \lambda_1^2\langle k'\overset{(1)}{E}_i\overset{(1)}{E}_i\rangle + 2\bar{E}_3\lambda_2\langle k'\overset{(2)}{E}_3\rangle + 2\bar{k}\lambda_1\lambda_2\langle\overset{(1)}{E}_i\overset{(2)}{E}_i\rangle + 2\lambda_1\lambda_2\langle k'\overset{(1)}{E}_i\overset{(2)}{E}_i\rangle + \bar{k}\lambda_2^2\langle\overset{(2)}{E}_i\overset{(2)}{E}_i\rangle + \lambda_2^2\langle k'\overset{(2)}{E}_i\overset{(2)}{E}_i\rangle. \tag{25}$$

Equations (18) and (20) are next substituted in Eq. (25). We note that several of the integrations required by the resulting expressions can be either fully or partially accomplished (see, for example, Ref. 14). After an extensive amount of manipulation we obtain the following inequality on k^*

$$k^* \leqslant \bar{k} + (\tfrac{1}{3}\lambda_1^2 - \tfrac{2}{3}\lambda_1)\langle k'^2\rangle/\bar{k} + \lambda_1^2I + 2\lambda_2J + 2\bar{k}\lambda_1\lambda_2K + 2\lambda_1\lambda_2L + \bar{k}\lambda_2^2M + \lambda_2^2T, \tag{26}$$

where

$$I = \frac{1}{\bar{E}_3^2}\langle k'\overset{(1)}{E}_i\overset{(1)}{E}_i\rangle = \frac{1}{(4\pi\bar{k})^2}\int_{\mathbf{x}'}\int_{\mathbf{x}''}\frac{\partial^2}{\partial x'_3\partial x''_3}\langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')\rangle \times \frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3}\,\frac{(x_i - x''_i)}{|\mathbf{x} - \mathbf{x}''|^3}d\mathbf{x}'d\mathbf{x}'', \tag{27}$$

$$J = \frac{1}{\bar{E}_3}\langle k'\overset{(2)}{E}_3\rangle = \frac{1}{(4\pi\bar{k})^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}\left(\int_{\mathbf{x}''}\frac{\partial}{\partial x''_3}\langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')\rangle \times \frac{(x'_j - x''_j)}{|\mathbf{x}' - \mathbf{x}''|^3}d\mathbf{x}''\right)\frac{(x_3 - x'_3)}{|\mathbf{x} - \mathbf{x}'|^3}d\mathbf{x}', \tag{28}$$

$$K = \frac{1}{\bar{E}_3^2}\langle\overset{(1)}{E}_i\overset{(2)}{E}_i\rangle = \frac{-1}{(4\pi\bar{k})^3}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_3}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x'_j}\left(\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_3}\langle k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')\rangle \times \frac{(x''_j - x'''_j)}{|\mathbf{x}'' - \mathbf{x}'''|^3}d\mathbf{x}'''\right)\frac{(x'_i - x''_i)}{|\mathbf{x}' - \mathbf{x}''|^3}d\mathbf{x}''\right]\frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3}dx', \tag{29}$$

$$L = \frac{1}{\bar{E}_3^2}\langle k'\overset{(1)}{E}_i\overset{(2)}{E}_i\rangle = \frac{-1}{(4\pi\bar{k})^3}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_3}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x''_j} \times \left(\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_3}\langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')\rangle\frac{(x''_j - x'''_j)}{|\mathbf{x}'' - \mathbf{x}'''|^3}d\mathbf{x}'''\right) \times \frac{(x'_i - x''_i)}{|\mathbf{x}' - \mathbf{x}''|^3}d\mathbf{x}''\right]\frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3}d\mathbf{x}', \tag{30}$$

$$M = \frac{1}{\bar{E}_3^2}\langle\overset{(2)}{E}_i\overset{(2)}{E}_i\rangle = \frac{1}{(4\pi\bar{k})^4}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}\left\{\int_{\mathbf{x}''}\frac{\partial}{\partial x''_3}\right.$$

$$\times \left[\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_m} \left(\int_{\mathbf{x}''''} \frac{\partial}{\partial x''''_3} \langle k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')k'(\mathbf{x}'''')\rangle \right.\right.$$
$$\left.\left. \times \frac{(x'''_m - x''''_m)}{|\mathbf{x}''' - \mathbf{x}''''|^3} d\mathbf{x}''''\right) \frac{(x''_i - x'''_i)}{|\mathbf{x}'' - \mathbf{x}'''|^3} d\mathbf{x}'''\right]$$
$$\left. \times \frac{(x'_j - x''_j)}{|\mathbf{x}' - \mathbf{x}''|^3} d\mathbf{x}''\right\} \frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3} d\mathbf{x}', \qquad (31)$$

and

$$T = \frac{1}{\bar{E}_3^2}\langle k'\overset{(2)}{E}_i\overset{(2)}{E}_i\rangle = \frac{1}{(4\pi\bar{k})^4} \int_{\mathbf{x}'} \frac{\partial}{\partial x'_j} \left\{\int_{\mathbf{x}''} \frac{\partial}{\partial x''_3}\right.$$
$$\times \left[\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_m} \left(\int_{\mathbf{x}''''} \frac{\partial}{\partial x''''_3} \langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')k'(\mathbf{x}'''')\rangle \right.\right.$$
$$\left.\left.\left. \times \frac{(x'''_m - x''''_m)}{|\mathbf{x}''' - \mathbf{x}''''|^3} d\mathbf{x}''''\right) \frac{(x''_i - x'''_i)}{|\mathbf{x}'' - \mathbf{x}'''|^3} d\mathbf{x}'''\right] \frac{(x'_j - x''_j)}{|\mathbf{x}' - \mathbf{x}''|^3} d\mathbf{x}''\right\}$$
$$\times \frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3} d\mathbf{x}'. \qquad (32)$$

The upper bound is given by the rhs of Eq. (26). The values of λ_1 and λ_2 are chosen so as to provide the best upper bound. That is, to minimize the right-hand side of Eq. (26). The correct values for λ_1 and λ_2 are

$$\lambda_1 = \frac{\frac{1}{3}(\overline{k'^2}/\bar{k})(\bar{k}M + T) + J(\bar{k}K + L)}{[\frac{1}{3}(\overline{k'^2}/\bar{k}) + I](\bar{k}M + T) - (\bar{k}K + L)^2} \qquad (33)$$

and

$$\lambda_2 = \frac{-[\frac{1}{3}(\overline{k'^2}/\bar{k}) + I]J - \frac{1}{3}\overline{k'^2}/\bar{k}\,(\bar{k}K + L)}{(\frac{1}{3}\overline{k'^2}/\bar{k} + I)(\bar{k}M + T) - (\bar{k}K + L)^2}. \qquad (34)$$

The expressions for I, J, K, L, M, and T show them to include correlation functions of third-, fourth-, and fifth-orders. The information contained in these correlation functions is enormous and would be difficult to obtain (see the work of Corson[6] involving measurements of third-order correlation functions). Thus, as a general expression, one must expect that Eq. (26) will have limited practical applicability. If, however, one limits consideration to a class of materials termed a two-phase cell material, then a simplified expression can be achieved that promises to be of a great practical utility. As we shall see for a two-phase cell material, the difficulty of determining the higher-order correlation functions required by the bounds can be expressed in terms of a limited number of parameters. Further, we shall see that these parameters can be associated with simple geometrical concepts. That is, they amount to shape factors, or clustering factors.

The concept of a two-phase cell material was introduced by Miller,[7] who applied it to Beran and Molyneux bounds. Miller showed that, for his model, the integrals of the three-point correlation function required by the bounds could be expressed in terms of two constant parameters. These parameters were subsequently identified as shape factors. The bounds so obtained were seen to represent a considerable improvement over the Hashin–Shtrikman bounds, which have been shown to be the best possible bounds that one can obtain that incorporates only volume fraction information. Further, Miller discussed the applicability of the two-phase cell material to real two-phase materials. For example, he showed that a Poisson material[15,16] is a special case of the cell material. In addition, he showed that by varying the values of the shape factors required by his bounds he could sweep out most of the space between the Hashin–Shtrikman bounds. From this, one can conclude that the two-phase cell material covers a broad class of physically important two-phase materials.

In the present paper we introduce a symmetric two-phase cell model and simplify the general expressions we obtain as bounds. The extension to the more general asymmetric model is straightforward but requires a greatly increased amount of algebraic manipulations. This, and the fact that the results for the asymmetric model reduce to those of the symmetric model for low concentrations, motivates this initial restriction to the symmetric model. It is planned to carry out the required extensions in a subsequent effort.

Briefly, in the symmetric two-phase cell material, the space is divided into a large number of closed surfaces. These closed regions are called cells. The subdivision of the space is arbitrary except for fulfilling the following requirements:

(1) Space is completely covered by cells;

(2) cells are distributed in a manner such that the material is statistically homogeneous and isotropic;

(3) the material property k of a cell is statistically independent of the material property of any other cell;

(4) the conditional probabilities of n points in and m points not being in the same cell of a particular material, given that one point is in a cell of that material, are the same for each material.

The bounds obtained by Miller [Eqs. (1) and (2)], included shape information (through G_1 and G_M). As we mentioned before, these bounds offer a substantial improvement over the Hashin–Shtrikman bounds. There is an improvement of 50% at $\alpha = 10$, for $v = \frac{1}{2}$, and greater improvement at the other values of v. By "improvement" we mean the reduction in width of the new bounds as compared to the Hashin–Shtrikman bounds.

From the expressions for the correlation functions (Appendix B), and Eq. (27) we write

$$I = k_2\eta^3[v(1-v)(1-2v)/(1+v\eta)^2]G, \qquad (35)$$

where

$$G = \frac{1}{(4\pi)^2} \int_{\mathbf{x}'}\int_{\mathbf{x}''} \frac{\partial^2 g(\mathbf{x},\mathbf{x}',\mathbf{x}'')}{\partial x'_3\,\partial x''_3} \frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3} \frac{(x_i - x''_i)}{|\mathbf{x}-\mathbf{x}''|^3} d\mathbf{x}'d\mathbf{x}''. \qquad (36)$$

Here $g_n = g_n(\mathbf{x},\mathbf{x}',\mathbf{x}'')$ is the conditional probability that all three points–picked at random–are in the same cell of material property k_n, given that one of the points is in a cell with material property k_n. For a symmetric cell material, $g_1 = g_2 = g$.

Similarly

$$J = k_2\eta^3[v(1-v)(1-2v)/(1+v\eta)^2]\overline{G}, \qquad (37)$$

where

$$\overline{G} = \frac{1}{(4\pi)^2} \int_{\mathbf{x}'} \frac{\partial}{\partial x'_j} \left(\int_{\mathbf{x}''} \frac{\partial}{\partial x''_3} g(\mathbf{x},\mathbf{x}',\mathbf{x}'')\right.$$
$$\left. \times \frac{(x'_j - x''_j)}{|\mathbf{x}'-\mathbf{x}''|^3} d\mathbf{x}''\right) \frac{(x_3 - x'_3)}{|\mathbf{x}-\mathbf{x}'|^3} d\mathbf{x}', \qquad (38)$$

and

$$K = -\eta^3[v(1-v)(1-2v)/(1+v\eta)^3]\overline{\overline{G}}, \tag{39}$$

where

$$\overline{\overline{G}} = \frac{1}{(4\pi)^3}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_3}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x''_j}\left(\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_3}g(\mathbf{x}',\mathbf{x}'',\mathbf{x}''') \times \frac{(x''_j - x'''_j)}{|\mathbf{x}''-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right)\frac{(x'_i - x''_i)}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]\frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'. \tag{40}$$

In Appendix C we show that $G = \overline{G} = \overline{\overline{G}}$. We also have

$$L = [-k_2 v(1-v)\eta^4/(1+v\eta)^3]\{[(1-v)^3 + v^3]G_2 + v(1-v)M_1\}, \tag{41}$$

where

$$G_2 = \frac{1}{(4\pi)^3}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_3}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x''_j}\left(\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_3}\mathbf{g}(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''') \times \frac{(x''_j - x'''_j)}{|\mathbf{x}''-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right)\frac{(x'_i - x''_i)}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]\frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' \tag{42}$$

and

$$M_1 = \frac{1}{(4\pi)^3}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_3}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x''_j}\left(\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_3}Q(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''') \times \frac{(x''_j - x'''_j)}{|\mathbf{x}''-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right)\frac{(x'_i - x''_i)}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]\frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'. \tag{43}$$

The function $\mathbf{g}$ is defined in a similar way to g except it refers to four points instead of three.

Here $Q_n = Q_n(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''')$ is the conditional probability that any two points are in one cell and the other two points are in another cell given that one point ($\mathbf{x}'''$) is in a cell with material property k_n. For a symmetric cell material $Q_1 = Q_2 = Q$.

Similarly

$$M = [v(1-v)\eta^4/(1+v\eta)^4]\{[(1-v)^3 + v^3]\overline{G}_2 + v(1-v)\overline{M}_1\}, \tag{44}$$

where

$$\overline{G}_2 = \frac{1}{(4\pi)^4}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}\left\{\int_{\mathbf{x}''}\frac{\partial}{\partial x''_3}\left[\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_m}\left(\int_{\mathbf{x}''''}\frac{\partial}{\partial x''''_3} \times \mathbf{g}(\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''')\frac{(x'''_m - x''''_m)}{|\mathbf{x}'''-\mathbf{x}''''|^3}\,d\mathbf{x}''''\right)\frac{(x''_i - x'''_i)}{|\mathbf{x}''-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right] \times \frac{(x'_j - x''_j)}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right\}\frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' \tag{45}$$

and

$$\overline{M}_1 = \frac{1}{(4\pi)^4}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}\left\{\int_{\mathbf{x}''}\frac{\partial}{\partial x''_3}\left[\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_m}\left(\int_{\mathbf{x}''''}\frac{\partial}{\partial x''''_3} \times Q(\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''')\frac{(x'''_m - x''''_m)}{|\mathbf{x}'''-\mathbf{x}''''|^3}\,d\mathbf{x}''''\right)\frac{(x''_i - x'''_i)}{|\mathbf{x}''-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right] \times \frac{(x'_j - x''_j)}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right\}\frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'. \tag{46}$$

Finally,

$$T = [k_2 v(1-v)\eta^5/(1+v\eta)^4]\{[(1-v)^4 - v^4]G_3 + v(1-v)(1-2v)M_2\}, \tag{47}$$

where

$$G_3 = \frac{1}{(4\pi)^4}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}\left\{\int_{\mathbf{x}''}\frac{\partial}{\partial x''_3}\left[\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_m}\left(\int_{\mathbf{x}''''}\frac{\partial}{\partial x''''_3} \times \mathrm{g}(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''')\frac{(x'''_m - x''''_m)}{|\mathbf{x}'''-\mathbf{x}''''|^3}\,d\mathbf{x}''''\right) \times \frac{(x''_i - x'''_i)}{|\mathbf{x}''-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right]\frac{(x'_j - x''_j)}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right\}\frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' \tag{48}$$

and

$$M_2 = \frac{1}{(4\pi)^4}\int_{\mathbf{x}'}\frac{\partial}{\partial x'_j}\left\{\int_{\mathbf{x}''}\frac{\partial}{\partial x''_3}\left[\int_{\mathbf{x}'''}\frac{\partial}{\partial x'''_m} \times \left(\int_{\mathbf{x}''''}\frac{\partial}{\partial x''''_3}\mathbf{Q}(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''')\frac{(x'''_m - x''''_m)}{|\mathbf{x}'''-\mathbf{x}''''|^3}\,d\mathbf{x}''''\right) \times \frac{(x''_i - x'''_i)}{|\mathbf{x}''-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right]\frac{(x'_j - x''_j)}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\quad\frac{(x_i - x'_i)}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'. \tag{49}$$

Here, g is defined in a similar way to g except it refers to five points instead of three. The function $\mathbf{Q}$ is also defined in a similar way to Q except it refers to three points in the same cell and two points in another cell. We show in Appendix C that $G_2 = \overline{G}_2$ and $M_1 = \overline{M}_1$.

C. Lower bound

To derive the lower bound, we proceed in a similar fashion. Our motivation of a trial function to use in the second variational principle is motivated by a perturbation solution of Eq. (8) in powers of $k'(\mathbf{x})$.

Accordingly, we might begin by expressing

$$D_i(\mathbf{x}) = \sum_{p=0}^{\infty}\overset{(p)}{D}_i(\mathbf{x}). \tag{50}$$

Here, $\overset{(p)}{D}_i$ is the order $(k'^p/\bar{k}^p)\overline{D}_3$, where

$$\overline{D}_3 = k^*\overline{E}_3. \tag{51}$$

Equation (51) is an alternate definition of k^* consistent with the definition given by Eq. (9).[10] As in Eq. (13) the condition

$$\overline{D}_i = \overline{D}_3\delta_{i3} \tag{52}$$

must be satisfied. Substituting D_i given by Eq. (50) into Eqs. (7) and (8), separating terms of the first and second order, we obtain differential equations in $D_i^{(1)}$ and $D_i^{(2)}$. Solving these equations and ensemble averaging, we find that $\overline{D_i^{(1)}} = 0$ while the solution for $D_i^{(2)}$ leads to the result $\overline{D_i^{(2)}} \neq 0$. This was not encountered in the case of the upper bound. For this reason we find it convenient to write in place of Eq. (50)

$$D_i(\mathbf{x}) = \overset{(0)}{D}_i + \sum_{p=1}^{\infty}(\overset{(p)}{D}_i + \overset{(p)}{\overline{D}}_i), \tag{53}$$

where $D_i^{(0)}$ is a constant and $\overline{D_i^{(p)}} = 0$. Here $(D_i^{(p)} + \overline{D}_i^{(p)})$ is of the order $(k'^p/\bar{k}^p)\overline{D}_3$. The condition given by Eq. (52) should still be satisfied. Equations (50) and (52) give

$$\overline{D}_3\delta_{i3} = \overset{(0)}{D}_i + \sum_{p=1}^{\infty}\overset{(p)}{\overline{D}}_i. \tag{54}$$

In this case, terms of the first order in the differential equations lead to

$$\frac{\partial}{\partial x_i}(\overset{(1)}{D}_i + \overset{(1)}{\overline{D}}_i) = 0 \tag{55}$$

and

$$\delta_{ijk}\frac{\partial}{\partial x_j}\left(-\overset{(0)}{D}_k\frac{k'}{\bar{k}^2} + \frac{1}{\bar{k}}(\overset{(1)}{D}_k + \overset{(1)}{\overline{D}}_k)\right) = 0; \tag{56}$$

solving these equations, we obtain

$$\left(\overset{(1)}{D}_i + \overset{(1)}{\overline{D}}_i\right) = \overset{(0)}{D}_i\frac{k'}{\bar{k}^2} - \frac{\overset{(0)}{D}_j}{4\pi\bar{k}}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}k'(\mathbf{x}')\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}'. \tag{57}$$

Ensemble-averaging Eq. (57), we obtain

$$\overset{(1)}{\overline{D}}_i = 0; \tag{58}$$

therefore,

$$\overset{(1)}{D}_i = \overset{(0)}{D}_i\frac{k'}{\bar{k}} - \frac{\overset{(0)}{D}_j}{4\pi\bar{k}}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}k'(\mathbf{x}')\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}'. \tag{59}$$

For terms of order $(k'^2/\bar{k}^2)\overline{D}_3$, the solution of the resulting differential equations is

$$(\overset{(2)}{D}_i + \overset{(2)}{\overline{D}}_i) = -\overset{(0)}{D}_i\frac{k'^2}{\bar{k}^2} + \frac{k'}{\bar{k}}(\overset{(1)}{D}_i + \overset{(1)}{\overline{D}}_i) + \frac{1}{4\pi\bar{k}}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'} \times\left(\overset{(0)}{D}_j\frac{k'^2(\mathbf{x})}{\bar{k}} - k'(\mathbf{x}')\left(\overset{(1)}{D}_j(\mathbf{x}') + \overset{(1)}{\overline{D}}_j(\mathbf{x}')\right)\right)\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}'. \tag{60}$$

Ensemble averaging the above equation, realizing that $\overline{D}_i^{(1)} = 0$ and $\partial/\partial x_j'\langle k'^2(\mathbf{x}')\rangle = 0$, yields

$$\overset{(2)}{\overline{D}}_i = -\overset{(0)}{D}_i\frac{\langle k'^2\rangle}{\bar{k}^2} + \frac{1}{\bar{k}}\langle k'\overset{(1)}{D}_i\rangle - \frac{1}{4\pi\bar{k}}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'} \times\langle k'(\mathbf{x}')\overset{(1)}{D}_j(\mathbf{x}')\rangle\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}'. \tag{61}$$

Using Eq. (59), we can write

$$\overset{(2)}{\overline{D}}_i = -\frac{\overset{(0)}{D}_j}{4\pi\bar{k}^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}\langle k'(\mathbf{x})k'(\mathbf{x}')\rangle\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}' + \frac{\overset{(0)}{D}_k}{(4\pi\bar{k})^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x_k''}\langle k'(\mathbf{x}')k'(\mathbf{x}'')\rangle \times\frac{(x_j' - x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}d\mathbf{x}''\right]\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}', \tag{62}$$

which simplifies to

$$\overset{(2)}{\overline{D}}_i = -\overset{(0)}{D}_i\langle k'^2\rangle/3\bar{k}^2. \tag{63}$$

Using this relation and Eq. (59) in Eq. (60), we have after simplification

$$\overset{(2)}{D}_i = \frac{\overset{(0)}{D}_i\langle k'^2\rangle}{3\bar{k}^2} - \frac{D_j^{(0)}}{4\pi\bar{k}^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}k'(\mathbf{x})k'(\mathbf{x}')\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}' + \frac{\overset{(0)}{D}_k}{(4\pi\bar{k})^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x_k''}k'(\mathbf{x}')k'(\mathbf{x}'') \times\frac{(x_j' - x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}d\mathbf{x}''\right]\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}d\mathbf{x}'. \tag{64}$$

Following the previous analysis, we can in general show that

$$\overset{(p)}{\overline{D}}_i = (-)^{(p-1)}(\overset{(0)}{D}_i/3^{(p-1)}\bar{k}^p)\langle k'^p\rangle, \quad p \geq 2. \tag{65}$$

Referring to Eq. (54) and substituting for $\overline{D}_i^{(p)}$, the expression in Eq. (65), we obtain

$$\overline{D}_3\delta_{i3} = \overset{(0)}{D}_i\left(1 + \sum_{p=1}^{\infty}(-)^{(p-1)}\frac{\langle k'^p\rangle}{3^{(p-1)}\bar{k}^p}\right). \tag{66}$$

Since the quantity between brackets in the rhs of the above equation is not equal to zero, we can write

$$\overset{(0)}{D}_i = \overset{(0)}{D}_3\delta_{i3}; \tag{67}$$

Eqs. (53) and (54) give

$$D_i(\mathbf{x}) = \overline{D}_3\delta_{i3} + \sum_{p=1}^{\infty}\overset{(p)}{D}_i(\mathbf{x}). \tag{68}$$

In order to obtain the lower bound, we introduce the trial function

$${}_ND_i(\mathbf{x}) = \overline{D}_3\delta_{i3} + \mu_1\overset{(1)}{D}_i(\mathbf{x}) + \cdots + \mu_N\overset{(N)}{D}_i(\mathbf{x}), \tag{69}$$

where $\mu_1, \mu_2, \ldots, \mu_N$ are modifying multipliers. Here ${}_ND_i(\mathbf{x})$ is equal to $D_i(\mathbf{x})$ if the $\mu_k = 1$ and $N \to \infty$. Using Eqs. (6), (9), and (52), we can write

$$\overline{D}_3^2/k^* \leq \langle D_iD_i/k\rangle, \tag{70}$$

where $D_i(\mathbf{x})$ is any trial function. We shall consider $D_i(\mathbf{x})$ as given by Eq. (69) for $N = 2$. Equation (70) yields

$$\overline{D}_3^2/k^* \leq \langle 1/k\rangle\overline{D}_3^2 + 2\mu_1\overline{D}_3\langle\overset{(1)}{D}_3(\mathbf{x})/k(\mathbf{x})\rangle + 2\mu_2\overline{D}_3\langle\overset{(2)}{D}_3(\mathbf{x})/k(\mathbf{x})\rangle + \mu_1^2\langle\overset{(1)}{D}_i(\mathbf{x})\overset{(1)}{D}_i(\mathbf{x})/k(\mathbf{x})\rangle + 2\mu_1\mu_2\langle\overset{(1)}{D}_i(\mathbf{x})\overset{(2)}{D}_i(\mathbf{x})/k(\mathbf{x})\rangle + \mu_2^2\langle\overset{(2)}{D}_i(\mathbf{x})\overset{(2)}{D}_i(\mathbf{x})/k(\mathbf{x})\rangle. \tag{71}$$

Using the expression for $\overset{(1)}{D}_i(\mathbf{x})$ –Eq. (59)–together with Eq. (67) yields, after performing the resulting integration,

$$2\overline{D}_3\langle\overset{(1)}{D}_3(\mathbf{x})/k(\mathbf{x})\rangle = \tfrac{4}{3}(\overline{D}_3\overset{(0)}{D}_3/\bar{k})\langle k'/k\rangle. \tag{72}$$

Similarly we may show that[4]

$$\langle\overset{(1)}{D}_i\overset{(1)}{D}_i/k\rangle = \overset{(0)^2}{D}_3[(1/3\bar{k}^2)\langle k'^2/k\rangle + Q], \tag{73}$$

where

$$Q = \frac{1}{16\pi^2}\frac{1}{\bar{k}^2}\int_{\mathbf{x}'}\int_{\mathbf{x}''}\frac{\partial^2}{\partial x_3'\partial x_3''}\langle k'(\mathbf{x}')k'(\mathbf{x}'')/k(\mathbf{x})\rangle \times\frac{(x_i - x_i')}{|\mathbf{x}-\mathbf{x}'|^3}\frac{(x_i - x_i'')}{|\mathbf{x}-\mathbf{x}''|^3}d\mathbf{x}'d\mathbf{x}'',$$

and we can rewrite Eq. (71) as

$$k^* \geq \overline{D}_3^2/\{\langle 1/k\rangle\overline{D}_3^2 + \tfrac{4}{3}\mu_1(\overline{D}_3\overset{(0)}{D}_3/\bar{k})\langle k'/k\rangle + \mu_1^2\overset{(0)^2}{D}_3 \times[1/3\bar{k}^2)\langle k'^2/k\rangle + Q] + 2\mu_2\overline{D}_3^2R + 2\mu_1\mu_2\overline{D}_3^2W + \mu_2^2\overline{D}_3^2Z\}. \tag{74}$$

The rhs of this equation represents the lower bound. In Eq. (74) we have

$$R = \frac{1}{\overline{D}_3}\left\langle\frac{\overset{(2)}{D}_3(\mathbf{x})}{k(\mathbf{x})}\right\rangle = \frac{1}{3}\frac{\overset{(0)}{D}_3}{\overline{D}_3}\left\langle\frac{1}{k}\right\rangle\frac{\langle k'^2\rangle}{\bar{k}^2} - \frac{\overset{(0)}{D}_3}{4\pi\bar{k}^2\overline{D}_3}$$
$$\times \int_{\mathbf{x}'}\frac{\partial}{\partial x_3'}\left\langle\frac{k'(\mathbf{x})k'(\mathbf{x}')}{k(\mathbf{x})}\right\rangle\frac{(x_3 - x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' + \frac{\overset{(0)}{D}_3}{(4\pi\bar{k})^2\overline{D}_3}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}$$
$$\times \left[\int_{\mathbf{x}''}\frac{\partial}{\partial x_3''}\left\langle\frac{k'(\mathbf{x}')k'(\mathbf{x}'')}{k(\mathbf{x})}\right\rangle\frac{(x_j'-x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]\frac{(x_3-x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}', \tag{75}$$

$$W = \frac{1}{\overline{D}_3^2}\left\langle\frac{\overset{(1)}{D}_i(\mathbf{x})\overset{(2)}{D}_i(\mathbf{x})}{k(\mathbf{x})}\right\rangle = \frac{1}{3}\frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2}\left\langle\frac{k'}{k}\right\rangle\frac{\langle k'^2\rangle}{\bar{k}^3} - \frac{1}{3}\frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2}\frac{\langle k'^2\rangle}{\bar{k}^2}$$
$$\times \frac{1}{4\pi\bar{k}}\int_{\mathbf{x}'}\frac{\partial}{\partial x_3'}\left\langle\frac{k'(\mathbf{x}')}{k(\mathbf{x})}\right\rangle\frac{(x_3-x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' - \frac{D_3^{(0)^2}}{\overline{D}_3^2 4\pi\bar{k}^3}$$
$$\times \int_{\mathbf{x}'}\frac{\partial}{\partial x_3'}\left\langle\frac{k'^2(\mathbf{x})k'(\mathbf{x}')}{k(\mathbf{x})}\right\rangle\frac{(x_3-x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' + \frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2 16\pi^2\bar{k}^3}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}$$
$$\times \left[\int_{\mathbf{x}'}\frac{\partial}{\partial x_3''}\left\langle\frac{k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')}{k(\mathbf{x})}\right\rangle\frac{(x_j'-x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]\frac{(x_3-x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'$$
$$+ \frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2 16\pi^2\bar{k}^3}\int_{\mathbf{x}'}\int_{\mathbf{x}''}\frac{\partial^2}{\partial x_3'\partial x_3''}\left\langle\frac{k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')}{k(\mathbf{x})}\right\rangle\frac{(x_i-x_i')}{|\mathbf{x}-\mathbf{x}''|^3}$$
$$\times \frac{(x_i-x_i''')}{|\mathbf{x}-\mathbf{x}''|^3}\,d\mathbf{x}'d\mathbf{x}'' - \frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2(4\pi\bar{k})^3}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x_3''}\left(\int_{\mathbf{x}'''}\frac{\partial}{\partial x_3'''}\right.\right.$$
$$\left.\left.\times \left\langle\frac{k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')}{k(\mathbf{x})}\right\rangle\frac{(x_i-x_i''')}{|\mathbf{x}-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right)\frac{(x_j'-x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]$$
$$\times \frac{(x_i-x_i')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}', \tag{76}$$

and

$$Z = \frac{1}{\overline{D}_3^2}\left\langle\frac{\overset{(2)}{D}_i(\mathbf{x})\overset{(2)}{D}_i(\mathbf{x})}{k(\mathbf{x})}\right\rangle = \frac{1}{9}\frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2}\left[\frac{\langle k'^2\rangle}{\bar{k}^2}\right]^2\left\langle\frac{1}{k}\right\rangle + \frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2(4\pi\bar{k})^2}$$
$$\times \int_{\mathbf{x}'}\int_{\mathbf{x}''}\frac{\partial^2}{\partial x_3'\partial x_3''}\left\langle\frac{k'^2(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')}{k(\mathbf{x})}\right\rangle\frac{(x_i-x_i')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'\frac{(x_i-x_i'')}{|\mathbf{x}-\mathbf{x}''|^3}\,d\mathbf{x}''$$
$$+ \frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2}\frac{1}{(16\pi^2\bar{k})^2}\int_{\mathbf{x}'}\int_{\mathbf{x}'''}\frac{\partial^2}{\partial x_j'\partial x_m'''}\left(\int_{\mathbf{x}''}\int_{\mathbf{x}''''}\frac{\partial^2}{\partial x_3''\partial x_3''''}\right.$$
$$\left.\times\left\langle\frac{k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')k'(\mathbf{x}'''')}{k(\mathbf{x})}\right\rangle\frac{(x_j'-x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\frac{(x_m'''-x_m'''')}{|\mathbf{x}'''-\mathbf{x}''''|^3}\,d\mathbf{x}''''\right)$$
$$\times \frac{(x_i-x_i')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'\frac{(x_i-x_i''')}{|\mathbf{x}-\mathbf{x}'''|^3}\,d\mathbf{x}''' - \frac{2}{3}\frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2}\frac{\langle k'^2\rangle}{\bar{k}^2}\frac{1}{4\pi\bar{k}^2}$$
$$\times \int_{\mathbf{x}'}\frac{\partial}{\partial x_3'}\left\langle\frac{k'(\mathbf{x})k'(\mathbf{x}')}{k(\mathbf{x})}\right\rangle\frac{(x_3-x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' + \frac{2}{3}\frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2}$$
$$\times \frac{\langle k'^2\rangle}{\bar{k}^2}\frac{1}{16\pi^2\bar{k}^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}\left(\int_{\mathbf{x}''}\frac{\partial}{\partial x_3''}\left\langle\frac{k'(\mathbf{x}')k'(\mathbf{x}'')}{k(\mathbf{x})}\right\rangle\right.$$
$$\left.\times \frac{(x_j'-x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right)\frac{(x_3-x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}' - 2\frac{\overset{(0)^2}{D}_3}{\overline{D}_3^2}\frac{1}{(4\pi)^3\bar{k}^4}\int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}$$
$$\times \left[\int_{\mathbf{x}''}\frac{\partial}{\partial x_3''}\left(\int_{\mathbf{x}'''}\frac{\partial}{\partial x_3'''}\left\langle\frac{k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')}{k(\mathbf{x})}\right\rangle\right.\right.$$
$$\left.\left.\times \frac{(x_i-x_i''')}{|\mathbf{x}-\mathbf{x}'''|^3}\,d\mathbf{x}'''\right)\frac{(x_j'-x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]\frac{(x_i-x_i')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}'. \tag{77}$$

Using the expressions for the correlation functions in Appendix B together with the relations between the constants in Appendix C, we can write

$$R = \frac{1}{3}\frac{\overset{(0)}{D}_3}{\overline{D}_3}\left\langle\frac{1}{k}\right\rangle\frac{\langle k'^2\rangle}{\bar{k}^2} - \frac{1}{3}\frac{\overset{(0)}{D}_3}{\overline{D}_3}\frac{1}{\bar{k}^2}\left\langle\frac{k'^2}{k}\right\rangle + \frac{\overset{(0)}{D}_3}{\overline{D}_3\bar{k}^2}\frac{k_1'^2}{k_1}$$
$$\times \frac{v(2v-1)\eta G}{(1-v)} + \frac{\overset{(0)}{D}_3}{\overline{D}_3}\frac{1}{\bar{k}^2}\frac{k_1'^3}{k_1}\frac{v}{(1-v)}(1+\eta-v\eta)\frac{1}{16\pi^2}$$
$$\times \int_{\mathbf{x}'}\frac{\partial}{\partial x_j'}\left[\int_{\mathbf{x}''}\frac{\partial}{\partial x_3''}f(\mathbf{x}',\mathbf{x}'')\frac{(x_j'-x_j'')}{|\mathbf{x}'-\mathbf{x}''|^3}\,d\mathbf{x}''\right]\frac{(x_3-x_3')}{|\mathbf{x}-\mathbf{x}'|^3}\,d\mathbf{x}',$$

where $f_n(\mathbf{x}',\mathbf{x}'')$ is the conditional probability that two points $(\mathbf{x}',\mathbf{x}'')$ are in the same cell with material property k_n, given that one point is in a cell with material property k_n. For symmetric cell materials $f_1 = f_2 = f$.

The last integral in this equation is shown in Appendix D to be zero, and we can write

$$R = [\overset{(0)}{D}_3(1+\eta)^{-1}/k_2\overline{D}_3(1+v\eta)^2]v(1-v)(1-2v)\eta^3(\tfrac{1}{3}-G), \tag{78}$$

where $\eta = \alpha - 1 = k_1/k_2 - 1$.

Similarly after a great deal of manipulations we can write

$$W = \frac{\overset{(0)^2}{D}_3(1+\eta)^{-1}}{k_2\overline{D}_3^2(1+v\eta)^3}v(1-v)\eta^3\{(-\tfrac{1}{3}+G)+\eta(G_2-G)$$
$$+ v[(\tfrac{2}{3}-2G)+\eta(-\tfrac{5}{9}+5G-3G_2+M_1)]$$
$$+ v^2[\eta(\tfrac{8}{9}-6G+3G_2-M_1)]\} \tag{79}$$

and

$$Z = \frac{\overset{(0)^2}{D}_3(1+\eta)^{-1}}{k_2\overline{D}_3^2(1+v\eta)^4}v(1-v)\eta^4\{(G-G_2)+\eta(G_2-G_3)$$
$$+ v[(\tfrac{2}{9}-4G+3G_2-M_1)+\eta(\tfrac{1}{9}+\tfrac{4}{3}G-6G_2+4G_3$$
$$+ M_1-M_2)] + v^2[(-\tfrac{2}{9}+4G-3G_2+M_1)+\eta(-\tfrac{1}{9}$$
$$- 5G+12G_2-4M_1-6G_3+3M_2)] + v^3[\eta(\tfrac{14}{3}G$$
$$- 9G_2+4G_3+3M_1-2M_2)]\}. \tag{80}$$

The multipliers μ_1 and μ_2 in Eq. (74) are chosen to give the best lower bound. It is found that

$$\mu_1 = \frac{D_3}{\overset{(0)}{D}_3}\frac{-\tfrac{2}{3}(Z/\bar{k})\langle k'/k\rangle + (\overline{D}^3/\overset{(0)}{D}_3)WR}{Z[(1/3\bar{k}^2)\langle k'^2/k\rangle + Q] - W^2(\overline{D}_3/\overset{(0)}{D}_3)^2} \tag{81}$$

and

$$\mu_2 = \frac{-R[(1/3\bar{k}^2)\langle k'^2/k\rangle + Q] + \tfrac{2}{3}(\overline{D}_3/\overset{(0)}{D}_3)(W/\bar{k})\langle k'/k\rangle}{Z[(1/3\bar{k}^2)\langle k'^2/k\rangle + Q] - W^2(\overline{D}_3/\overset{(0)}{D}_3)^2}. \tag{82}$$

II. SMALL PERTURBATION CASE

The bounds are expanded for small perturbations ($\eta = \alpha - 1 \ll 1$). It is found that the upper and lower bounds coincide to order η^5 and take the form:

$$\begin{aligned}k^*/k^2 = {} & 1 + v\eta - \tfrac{1}{3}v\eta^2 + Gv\eta^3 - G_2v\eta^4 + G_3v\eta^5 + \cdots \\ & + \tfrac{1}{3}v^2\eta^2 + (\tfrac{1}{3} - 3G)v^2\eta^3 + (4G_2 - 2G - M_1)v^2\eta^4 \\ & + (3G_2 - 5G_3 + M_2)v^2\eta^5 + \cdots \\ & + (-\tfrac{1}{3} + 2G)v^3\eta^3 + (-\tfrac{1}{3} + 6G - 6G_2 + 2M_1)v^3\eta^4 \\ & + (3G - 12G_2 + 3M_1 + 10G_3 - 4M_2)v^3\eta^5 + \cdots \\ & + (\tfrac{1}{3} - 4G + 3G_2 - M_1)v^4\eta^4 + (\tfrac{1}{3} - 9G + 18G_2 \\ & - 6M_1 - 10G_3 + 5M_2)v^4\eta^5 + \cdots \\ & + (-\tfrac{1}{3} + 6G - 9G_2 + 3M_1 + 4G_3 - 2M_2)v^5\eta^5 + \cdots \end{aligned} \tag{83}$$

(to order η^5 and all powers of v).

The perturbation solution was calculated and found to coincide to the bounds to order η^5 as given by the above equation.

If Miller's bounds–Eqs. (1) and (2)–are expanded for small η, we obtain

$$\begin{aligned}k_{UM}/k_2 = {} & 1 + v\eta - \tfrac{1}{3}v\eta^2 + Gv\eta^3 - 3G^2v\eta^4 + 9G^3v\eta^5 + \cdots \\ & + \tfrac{1}{3}v^2\eta^2 + (\tfrac{1}{3} - 3G)v^2\eta^3 + (15G^2 - 2G)v^2\eta^4 \\ & + (9G^2 - 63G^3)v^2\eta^5 + \cdots \\ & + (-\tfrac{1}{3} + 2G)v^3\eta^3 + \cdots \end{aligned} \tag{84}$$

and

$$\begin{aligned}k_{LM}/k_2 = {} & 1 + v\eta - \tfrac{1}{3}v\eta^2 + Gv\eta^3 + (\tfrac{3}{2}G^2 - 2G + \tfrac{1}{6})v\eta^4 \\ & + \tfrac{1}{4}(9G^3 - 15G^2 + 11G - 1)v\eta^5 + \cdots \\ & + \tfrac{1}{3}v^2\eta^2 + (\tfrac{1}{3} - 3G)v^2\eta^3 + (-\tfrac{19}{18} + 8G - \tfrac{15}{2}G^2)v^2\eta^4 \\ & + (\tfrac{19}{12} - \tfrac{155}{12}G + \tfrac{87}{4}G^2 - \tfrac{63}{4}G^3)v^2\eta^5 + \cdots \\ & + (-\tfrac{1}{3} + 2G)v^3\eta^3 + \cdots . \end{aligned} \tag{85}$$

Similarly Hashin bounds for small perturbations give

$$\begin{aligned}k_{UH}/k_2 = {} & 1 + v\eta - \tfrac{1}{3}v\eta^2 + \tfrac{1}{3}v\eta^3 + \cdots \\ & + \tfrac{1}{3}v^2\eta^2 - \tfrac{4}{9}v^2\eta^3 + \cdots - \tfrac{2}{3}v^3\eta^3 + \cdots \end{aligned} \tag{86}$$

and

$$\begin{aligned}k_{LH}/k_2 = {} & 1 + v\eta - \tfrac{1}{3}v\eta^2 + \tfrac{1}{9}v\eta^3 + \cdots + \tfrac{1}{3}v^2\eta^2 \\ & - \tfrac{2}{9}v^2\eta^3 + \cdots + \tfrac{1}{9}v^3\eta^3 + \cdots . \end{aligned} \tag{87}$$

We can see that Hashin bounds provide an exact solution for the conductivity k^*, for small perturbations up to order η^2, while Miller bounds give an exact solution up to order η^3 since the bounds coincide to these orders. We also see that our bounds give the exact solution to order η^5 and all values of v. The solution provided by our bounds includes both Miller and Hashin solutions. We also notice that the information included in Hashin bounds are volume fraction information (v). In case of Miller bounds, they include–beside the volume fraction information–the number G. It was shown by Miller[7] that this number has a geometric significance. It was bounded and found to lie between $\frac{1}{9}$ and $\frac{1}{3}$. The value $\frac{1}{9}$ was shown to belong to spherical inclusions and the value $\frac{1}{3}$ to platelike inclusions. On the other hand, our bounds include the constants G, G_2, G_3, M_1, and M_2. Here G is the same number appearing in Miller bounds. In this paper, we will find general bounds on the rest of the constants, and show their physical significance.

The small perturbation expansion of our bounds shows an improvement over Miller bounds since our bounds coincide to order η^5 while his coincide to order η^3 only. A numerical evaluation of this improvement will be given later for the special cases of spherical and plate-shaped inclusions.

III. BOUNDS ON THE CONSTANTS G, G_2, G_3, M_1, AND M_2

In this section we present bounds on the constants G, G_2, G_3, M_1, and M_2. These bounds give the range of existence of each of these constants. The bounds on the constant G were obtained by Miller. He found that $\frac{1}{9} \leqslant G \leqslant \frac{1}{3}$. Miller used the fact that his upper and lower bounds are positive and finite to obtain these bounds. Here we obtained the same bounds on G in an alternative way. Miller's upper and lower bounds must satisfy the relation $k_{UM} \geqslant k_{LM}$. Going to the small perturbation limit and taking terms up to order $v\eta^4$ leads to the required bounds on G. The bounds on G_2, G_3, and M_2 are obtained by going to the perturbation limit and applying the condition that our bounds lie within Miller's bounds, or

$$k_{UM} \geqslant k_U \quad \text{and} \quad k_{LM} \leqslant k_L.$$

This can be easily proven using the Cauchy Schwartz inequality.[17]

In the case of $v = 0.5$, the bounds are independent of G, G_3, and M_2. This is also expected since in this case we have a symmetric random medium. For such a medium the odd-order moments could be related to the lower even order moments.[10] The constants G, G_3, and M_2 are derived from the third and fifth order moments which could be reduced to the second and fourth order moments expressed in terms of the constants G_2 and M_1. This case was considered, and the condition that the bounds must be positive is used to obtain bounds on M_1 in terms of G_2. Because of the great deal of algebra involved in these calculations, we only list here the bounds on these constants:

$$\tfrac{1}{9} \leqslant G \leqslant \tfrac{1}{3}, \tag{88}$$

$$3G^2 \leqslant G_2 \leqslant -(\tfrac{3}{2}G^2 - 2G + \tfrac{1}{6}), \tag{89}$$

$$G_2^2/G \leqslant G_3 \leqslant G_2 - (G_2 - G)^2/(\tfrac{1}{3} - G), \tag{90}$$

$$-G_2 \leqslant M_1 \leqslant \tfrac{2}{9} - G_2, \tag{91}$$

$$\begin{aligned} & \frac{(1-6G)}{27G^2}\,\frac{(GG_3 - G^2)^2}{(\frac{1}{3}G_2 - G^2)^2} - \frac{(\frac{1}{3}G_3 - GG_2)^2}{(\frac{1}{3}G_2 - G^2)^2}(\tfrac{1}{3} - 2G) + 4G \\ & \quad \times \frac{(GG_3 - G_2^2)}{(\frac{1}{3}G_2 - G^2)} + (4G_3 - G_2) + 2(G + M_1 - 3G_2) \\ & \quad \times \frac{(\frac{1}{3}G_3 - GG_2)}{(\frac{1}{3}G_2 - G_2)} \geqslant M_2, \end{aligned} \tag{92a}$$

$$\begin{aligned} & \frac{-8G}{(1-3G)^2}\,\frac{(\frac{1}{3}G_2 - \frac{1}{3}G_3 + GG_2 + GG_3 - G_2^2 - G^2)}{(-\frac{1}{9} + \frac{4}{3}G - \frac{2}{3}G_2 - G^2)^2} + 2G \\ & \quad \times \frac{(G_2 - \frac{2}{3}G_3 - GG_2 - \frac{1}{3}G + G^2)^2}{(-\frac{1}{9} + \frac{4}{3}G - \frac{2}{3}G_2 - G^2)^2} - 4(\tfrac{1}{3} - G) \\ & \quad \times \frac{(\frac{1}{3}G_2 - \frac{1}{3}G_3 + GG_2 + GG_3 - G_2^2 - G^2)}{(-\frac{1}{9} + \frac{4}{3}G - \frac{2}{3}G_2 - G^2)} \\ & \quad + 2(-\tfrac{5}{9} + 5G - 3G_2 + M_1) \\ & \quad \times \frac{(G_2 - \frac{2}{3}G_3 - GG_3 - \frac{1}{3}G + G^2)}{(-\frac{1}{9} + \frac{4}{3}G - \frac{2}{3}G_2 - G^2)} + (\tfrac{1}{9} + \tfrac{4}{3}G - 6G_2 \\ & \quad + 4G_3 + M_1) \leqslant M_2. \end{aligned} \tag{92b}$$

Taking all the possible values of G in Eq. (89), we find

$$\tfrac{1}{27} \leqslant G_2 \leqslant \tfrac{1}{3}. \tag{93}$$

Similarly considering all possible values of G and G_2 in Eq. (90), we find

$$\tfrac{1}{81} \leqslant G_3 \leqslant \tfrac{1}{3}. \tag{94}$$

IV. SIGNIFICANCE OF THE CONSTANTS

A. The constants G, G_2, and G_3

In order to show the nature of these constants, we present the solution of the single body problem. By this we mean the solution for the effective conductivity of a single body embedded in an infinite medium of another material. This problem has been solved in the literature for different geometries. Here we shall outline the nature of the problem. Let k_1 and k_2 be the thermal conductivities of the body and the medium respectively. It can be shown that[18]

$$k^* = k_2 + (k_1 - k_2)v_1 f_1, \tag{95}$$

where $f_1 = \overline{E}_1/\overline{E}$ is the ratio between the temperature gradient averaged over the volume of the inclusion ($\overline{E}_1$) and the applied temperature gradient ($\overline{E}$). As we see, k^* may be calculated if f_1 is known.

The special case of a spheroid was considered and f_1 was calculated.[19] The spheroid was placed in a constant temperature gradient (or electric field). After f_1 was substituted into Eq. (95) it was found that

$$k^* = k_2 + v_1 \sum_{i=1}^{3} \frac{(k_1 - k_2)\cos^2\alpha_i}{1 + A_i(k_1/k^* - 1)}, \tag{96}$$

where α_i are the angles made by the spheroid axes and the applied field. We assume randomly oriented spheroids ($\cos^2\alpha_i = \frac{1}{3}$, $i = 1, 2, 3$). Here A_i depend on the axial ratios of the spheroid, where

$$A_2 = A_3 = A \quad \text{and} \quad A_1 = 1 - 2A \tag{97}$$

For

$A = 0$ the spheroid reduces to a plate,
$A = \frac{1}{3}$ the spheroid reduces to a phere,
$A = \frac{1}{2}$ the spheroid reduces to a needle.

In Eq. (96) the spheroid is assumed to be placed in a homogeneous medium of conductivity k^*. If we assume the medium to have a conductivity k_2, we place $k^* = k_2$ in the rhs of Eq. (96). If we use Eq. (97) in Eq. (96) and expand the rhs for small perturbations writing v_1 as v, we find

$$\begin{aligned} k^*/k_2 &= 1 + v\eta - \tfrac{1}{3}v\eta^2 + \tfrac{1}{3}v\eta^3[2A^2 + (1 - 2A)^2] \\ &\quad - \tfrac{1}{3}v\eta^4[2A^3 + (1 - 2A)^3] + \tfrac{1}{3}v\eta^5[2A^4 \\ &\quad + (1 - 2A)^4] + \cdots . \end{aligned} \tag{98}$$

Since our bounds give the exact solution to order η^5; we compare terms of orders $v\eta^3$, $v\eta^4$, and $v\eta^5$ in Eqs. (83) and (98) to obtain values for G, G_2, and G_3 in the general case of a spheroid. We find

$$\begin{aligned} G &= \tfrac{1}{3}[2A^2 + (1 - 2A)^2], \\ G_2 &= \tfrac{1}{3}[2A^3 + (1 - 2A)^3], \end{aligned} \tag{99}$$

and

$$G_3 = \tfrac{1}{3}[2A^4 + (1 - 2A)^4].$$

For spheres, $A = \frac{1}{3}$, Eq. (99) gives

$$G = \tfrac{1}{9}, \quad G_2 = \tfrac{1}{27}, \quad \text{and} \quad G_3 = \tfrac{1}{81}\cdot \tag{100}$$

For plates $A = 0$, and we have

$$G = \tfrac{1}{3}, \quad G_2 = \tfrac{1}{3}, \quad \text{and} \quad G_3 = \tfrac{1}{3}. \tag{101}$$

Similarly for needles ($A = \frac{1}{2}$) we obtain

$$G = \tfrac{1}{6}, \quad G_2 = \tfrac{1}{12}, \quad \text{and} \quad G_3 = \tfrac{1}{24}. \tag{102}$$

As we see from the above, the knowledge of the geometry is enough to obtain an exact solution for small concentrations (neglecting the v^2 terms, or the effect of the particles on each other). We also see that we can identify the constants G, G_2, and G_3 for any geometry. Equation (99) gives the constants for a spheroid. Therefore the constants G, G_2, and G_3 are geometric constants. From the bounds on G, G_2, and G_3 [Eqs. (88), (93) and (94)] and Eqs. (100) and (101), we see that the lower bounds, on these constants correspond to a spherical geometry while the upper bounds correspond to a platelike geometry.

In general, we need an infinite number of these constants in order to define the geometry completely. In some cases, only a few number of these constants is required. For example, in the cases of spheres and plates, the constant G is enough to specify the geometry. For $G = \frac{1}{9}$ (sphere), the bounds on G_2 [Eq. (89)] give

$$\tfrac{1}{27} \leqslant G_2 \leqslant \tfrac{1}{27} \quad \text{or} \quad G_2 = \tfrac{1}{27},$$

which is the value for spheres, given by Eq. (100). For $G = \frac{1}{9}$ and $G_2 = \frac{1}{27}$, Eq. (90) gives

$$\tfrac{1}{81} \leqslant G_3 \leqslant \tfrac{1}{81} \quad \text{or} \quad G_3 = \tfrac{1}{81},$$

which is again the value for spheres [Eq. (100)]. In other words, the specification of the value $G = \frac{1}{9}$ leads to the specification of the values of G_2, G_3—and so on. Similar results can be obtained for plates.

For needles, the value $G = \frac{1}{6}$ [Eq. (102)] leads to $\frac{1}{12} \leqslant G_2 \leqslant \frac{1}{8}$. We see here that for a needle geometry the value $G = \frac{1}{6}$ is not enough to specify G_2. If we pick the value $G = \frac{1}{12}$ for a needle [Eq. (102)] and substitute this value in Eq. (90) we find

$$\tfrac{1}{24} \leqslant G_3 \leqslant \tfrac{1}{24} \quad \text{or} \quad G_3 = \tfrac{1}{24},$$

which is the value for a needle [Eq. (102)]. So for needles, we expect that the constants G and G_2 to specify the geometry.

In general, using higher orders in the trial functions [Eqs. (23) and (69)] introduces an infinite set of parameters appearing in the terms of order v that define the shape of the individual cells. Miller[7] showed that G is independent of the size of the cells. It can be shown exactly in the same way that G_2 and G_3 are also independent of the absolute cell size.

We notice that the bounds on G_2 [Eq. (89)] are functions of G, and the bounds on G_3 [Eq. (90) are functions of G and G_2. That is why, in general, for each value of G, there is a range in which G_2 exists, and for specific values of G and G_2 there is a range of existence of G_3 and so on.

B. The constants M_1 and M_2

If terms of order higher than v are required in the solution for k^*, more refined geometrical information than shape information is required. This more detailed information is packing information. By packing, we mean how the individual cells fit together. The first packing parameters that appear in the first order packing terms (v^2 terms) are M_1 and M_2 [see Eq. (83)]. Referring to the integrals defining the geometric parameters G, G_2 and G_3 [Eqs. (36), (42), and (48)], we find that they include probability functions (g, g, and g) based on one cell. In the integrals representing M_1 and M_2 [Eqs. (43) and (49)] the probabilities Q and $\mathbf{Q}$ are based on two cells. In other words M_1 and M_2 include information about the effect of the neighboring cell or "the first order packing information."

In the special cases of spheres and plates, the general bounds on M_1 and M_2 [Eqs. (91) and (92)] reduce to

$$-\tfrac{1}{27} \leqslant M_1 \leqslant \tfrac{5}{27}, \qquad -\tfrac{2}{81} \leqslant M_2 \leqslant \tfrac{28}{81} \tag{103}$$

for spheres and

$$-\tfrac{1}{3} \leqslant M_1 \leqslant -\tfrac{1}{9}, \qquad -\tfrac{2}{3} \leqslant M_2 \leqslant -\tfrac{2}{9} \tag{104}$$

for plates.

Including higher orders in the trial functions [Eqs. (23) and (69)] will introduce more packing parameters. In general, we must have an infinite number of constants to define the packing completely. Sometimes certain combinations of some of these constants are enough to define the packing completely. In the case of spheres, it is found that if we take the combination $M_1 = -\frac{1}{27}$ and $M_2 = -\frac{2}{81}$, our bounds coincide for all v and η giving an exact value for k^* equal to Miller's upper bound. Moreover, the combination $M_1 = \frac{5}{27}$ and $M_2 = \frac{28}{81}$ gives bounds that coincide to Miller's lower bound. Similar results are obtained for plates. The values $M = -\frac{1}{3}$ and $M_2 = -\frac{2}{3}$ lead to Miller's upper bound, while the values $M_1 = -\frac{1}{9}$ and $M_2 = -\frac{2}{9}$ lead to Miller's lower bound.

V. SELF-CONSISTENT SCHEME

An effective constant for the material may be determined by a consistency argument through the so called "self-consistent approximation." This approach has been developed in elastic problems by Budiansky[20] and Hill.[21] In this approach, a cell of conductivity k_1 is placed in a homogeneous matrix of conductivity k^*. The effective conductivity of such a model for spheres was found to be

$$k^* = k_2 + \tfrac{1}{3} v[(k_1 - k_2)/k_1](2k_1 + k^*). \tag{105}$$

Here we propose to use Eq. (105) in an iterative procedure. First we assume that $k^* = k_2$ in the rhs of Eq. (105) as a first approximation. This gives

$$k^*/k_2 = (1 + 3v\eta/(3+\eta)), \qquad \eta = k_1/k_2 - 1. \tag{106}$$

Equation (106) is identical to Eq. (96) for randomly oriented spheres ($\cos^2\alpha_i = \frac{1}{3}$, $i = 1, 2, 3$, $A = \frac{1}{3}$) and gives the exact solution up to terms of order v. In this case, spheres are assumed to be very far apart such that each could be considered in a homogeneous medium of property k_2 (small concentration solution). As a second approximation we assume that k^* in the rhs of Eq. (105) is given by Eq. (106) and calculate k^* and so on. This introduces the effect of the first sphere on the second one through its effect on the property of the medium around it. Repeating this procedure until we recover all the terms of order $\eta^n, n \leqslant 5$, we obtain the following result, neglecting terms of under η^6 and higher:

$$\begin{aligned} k^*/k_2 = 1 &+ v\eta - \tfrac{1}{3}v\eta^2 + \tfrac{1}{9}v\eta^3 - \tfrac{1}{27}v\eta^4 + \tfrac{1}{81}v\eta^5 + \cdots \\ &+ \tfrac{1}{3}v^2\eta^2 - \tfrac{1}{9}v^2\eta^4 + \tfrac{8}{81}v^2\eta^5 + \cdots \\ &- \tfrac{1}{9}v^3\eta^3 + \tfrac{5}{27}v^3\eta^4 - \tfrac{2}{27}v^3\eta^5 + \cdots \\ &- \tfrac{1}{27}v^4\eta^4 - \tfrac{8}{81}v^4\eta^5 + \cdots + \tfrac{5}{81}v^5\eta^5 + \cdots . \end{aligned} \tag{107}$$

Equation (107) gives the exact solution given by the self-consistent scheme (s.c.s.) in the case of spheres up to terms of order η^5. We compare this solution to the exact solution—up to order η^5—obtained before from the coincident part of our bounds [Eq. (83)]. We find that substituting the values $M_1 = \frac{1}{27}$ and $M_2 = \frac{4}{81}$ in Eq. (83) gives us the s.c.s. solution [Eq. (107)]. Therefore these values of M_1 and M_2 can be identified with the s.c.s. solution for spheres. We see that these values lie between the bounds for M_1 and M_2 given by Eq. (103).

A similar procedure was tried for the case of plates. It was found that we cannot identify any values for the constants M_1 and M_2 in our solution [Eq. (83)] which give us the s.c.s. solution for plates. The reason for that is in the nature of the iteration procedure. In Eq. (96) we specify $\cos^2\alpha_i = \frac{1}{3}$ for randomly oriented spheroids (see Ref. 18) from which spheres and plates are special cases. In other words the direction of the applied field makes equal angles with the axis of the spheroid. When we iterate, we effectively place a new spheroid in a homogeneous medium with an effective property equal to that of the first spheroid placed in a medium of property k_2. The equation used in the second iteration is the same as the original equation used in the first iteration. It follows that the second ellipsoid is also placed with its axis making equal angles with the applied field and this process is repeated. We eventually end up by a set of spheroids with corresponding axis aligned. In case of spheres this is always satisfied without violation of the randomness and the isotropy assumed in deriving the bounds. In the case of plates, the result obtained by the s.c.s. violates the above conditions. Therefore, if the volume fraction is high (such that v^2 terms become important), the s.c.s. leads to a different result from the one obtained through the coincident part of our bounds.

Hashin[22] proposed a modification to the s.c.s. In the new scheme, we place a composite sphere with a core of conductivity k_1 and a shell of conductivity k_2 in a medium of conductivity k^*. The ratio of the radii of the inner and outer spheres, ρ, can be varied as we wish. This kind of scheme cannot be used in conjunction with Miller's or our bounds. We cannot identify a value for the geometric or packing parameters for this scheme. The reason is that it violates the assumption of independence of the cells in Miller's model. The model we used of a single sphere of a homogeneous material in a medium of conductivity k^* is consistent with Miller's model assumption.

VI. IMPROVEMENT OVER MILLER'S BOUNDS

As we saw before, for small perturbations, Miller's bounds give the exact solution to order η^3 while our bounds give the solution to order η^5. In the special cases of spheres and plates we saw that if the lower bounds on M_1 and M_2 are combined, we obtain bounds that coincide giving an exact solution equal to Miller's

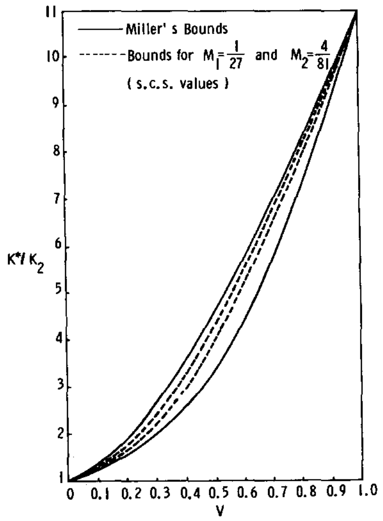


FIG. 1. $\eta = 10$.

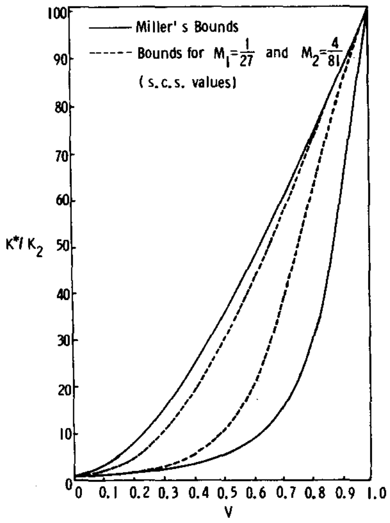


FIG. 2. $\eta = 100$.

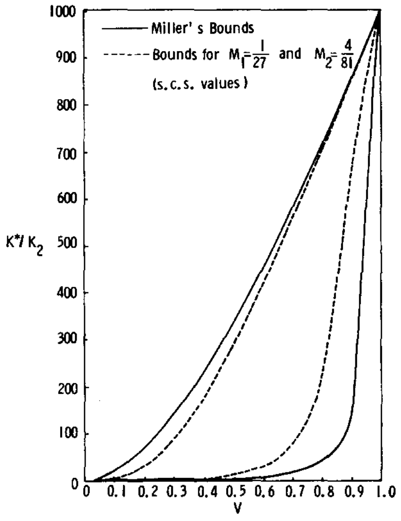


FIG. 3. $\eta = 1000$.

FIGS. 1-3. Bounds on effective conductivity (spheres).

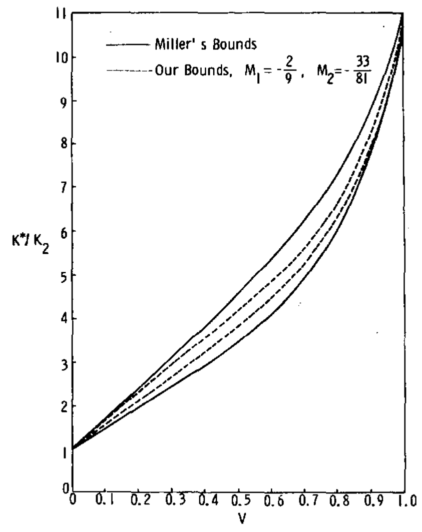


FIG. 4. Bounds for plates, $\eta = 10$.

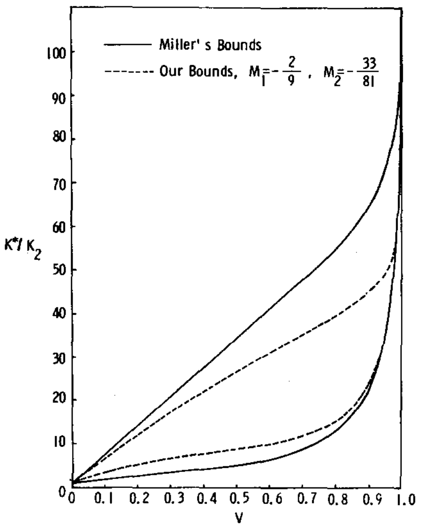


FIG. 5. Bounds for plates, $\eta = 100$.

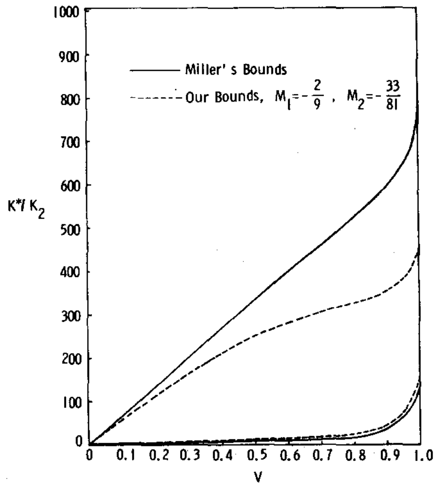


FIG. 6. Bounds for plates, $\eta = 1000$.

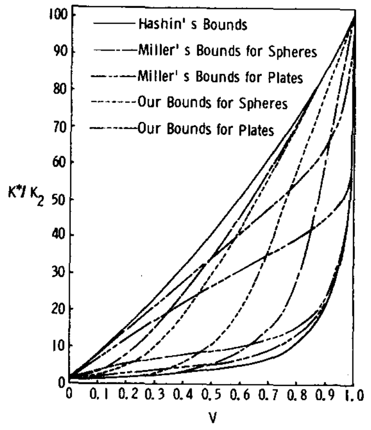


FIG. 7. Comparison of the bounds for $\eta = 100$.

TABLE I. Bounds for thermal conductivity.

klM/k_2	kUM/k_2	kl/k_2	kU/k_2	v	G	M_1	M_2	Geometry of inclusions
1.2	1.6	—	—	0.05	$\frac{1}{9}$	—	—	Spheres
2.4	3.3	—	—	0.05	$\frac{1}{3}$	—	—	Plates
1.3	3.0	1.4	1.7	0.1	$\frac{1}{9}$	$\frac{1}{27}$	$\frac{4}{81}$	Spheres
2.7	7.7	4.6	7.4	0.1	$\frac{1}{3}$	$-\frac{2}{9}$	$-\frac{33}{81}$	Plates

upper bound. If the upper bounds are combined, we obtain an exact solution equal to Miller's lower bound. Aside from these cases in which we have a 100% improvement over Miller's bounds, we considered two other cases. First we plot our bounds and Miller's for spheres and for $M_1 = \frac{1}{27}$ and $M_2 = \frac{4}{81}$. These are the values of M_1 and M_2 corresponding to the s.c.s. We used the values of $\eta = 10, 100,$ and 1000, and the bounds together with Miller's bounds are plotted in Figs. 1, 2, and 3. It is found that for $\eta = 10$, the spread of the bounds is at the most 28% of the spread of Miller's bounds. For large η, the spread increases. For $\eta = 1000$, it reaches 80%; however, for volume fractions less than 20% it is within 42%. For plates we took the values $M_1 = -\frac{2}{9}$ and $M_2 = -\frac{33}{81}$. The reason for choosing these values will be clear later. For $\eta = 10$, the spread is at the most 29% of Miller's bounds. For $\eta = 1000$ it reaches 88%. The bounds are plotted in Figs. 4, 5, and 6. For comparison, these bounds are plotted together with Hashin bounds for $\eta = 100$ (Fig. 7).

In order to show the importance of the introduction of packing information, we give the following example. Let us suppose we have a low conductivity thermal storage material that we wish to make more highly conductive by putting in inclusions where $v = 0{\cdot}05$ or $0{\cdot}1$ and for which $\eta = 100$. In Table I, we present the values of Miller's and our bounds for $G = \frac{1}{9}$ and $\frac{1}{3}$. Here k_l/k_2 and k_U/k_2 are the lower and upper bounds divided by the matrix conductivity.

It is clear from Table I, that, for both 0·05 and 0·1, platelike inclusions are superior to spherical inclusions. However (and this is often overlooked in design), the relative advantage of the shape of inclusion may be lost if the inclusions are not properly packed. For example, when $v = 0{\cdot}05$, the platelike inclusions are always better than spherical inclusions and not as much care need be taken in the packing. When $v = 0{\cdot}1$, the lower bound of the platelike inclusions (2.7) is lower than the upper bound of the spherical inclusions (3.0). It is thus necessary to identify the packing parameters which assume that if platelike inclusions are used, a value near (7.7) will be obtained. We see that the choice of $M_1 = -\frac{2}{9}$ and $M_2 = \frac{-33}{81}$ for plates gives the bounds (4.6) and (7.4) for $v = 0{\cdot}1$. We also see that the lower bound (4.6) is even higher than Miller's upper bound for spheres (3.0). This will assure a higher value of the conductivity if plates are used over spheres for $v = 0{\cdot}1$.

ACKNOWLEDGMENTS

The author would like to express his appreciation to Professor Mark Beran of the University of Pennsylvania for his encouragement and guidance and to Professor John McCoy of the Catholic University of America for reviewing this paper and for his interest in this work. This work was supported by the Institute of Direct Energy Conversion, University of Pennsylvania, Philadelphia, Pennsylvania.

APPENDIX A: DIFFERENT INTERPRETATIONS OF THE VARIOUS QUANTITIES

Physical subject	ϕ	$\mathbf{E} = \nabla\phi$	D	k_{ij}
Thermal conduction	Temperature	Temperature gradient	Heat flux	Thermal conductivities
Electrical conduction	Electric potential	Electric field intensity	Current density	Electric conductivities
Electrostatics	Electric potential	Electric field intensity	Electric induction Electric displacement	Dielectric constants, Permittivities
Magnetostatics	Magnetic potential	Magnetic field intensity	Magnetic induction	Magnetic permeabilities

APPENDIX B: EVALUATION OF THE CORRELATION FUNCTIONS

Here we shall write the expressions for the correlation functions appearing in the bounds, using Miller's symmetric cell model. Because of the gread deal of algebra involved, we shall present only the final results and definitions. For the details of calculations see Ref. 14.

1. The correlation functions $\langle\kappa'(\mathbf{x})\,\kappa'(\mathbf{x}')\kappa'(\mathbf{x}'')\rangle$ and $\langle\kappa(\mathbf{x}')\kappa'(\mathbf{x}'')/\kappa(\mathbf{x})\rangle$

These functions have been derived by Miller.[7] They were found to take the form

$$\langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')\rangle = k_1'^3\,[v(1-2v)/(1-v)^2]g \qquad \text{(B1)}$$

and

$$\langle k'(\mathbf{x}')k'(\mathbf{x}'')/k(\mathbf{x})\rangle = [k_1'^2 v/k_1(1-v)]$$
$$\times\,[\eta(2v-1)g + (1+\eta-v\eta)f(\mathbf{x}',\mathbf{x}'')], \qquad \text{(B2)}$$

where $g_n = g_n(\mathbf{x},\mathbf{x}',\mathbf{x}'')$ is the conditional probability that all three points–thrown at random–are in the same cell of material property k_n, given that one of the points is in a cell with material property k_n, and $f_n = f_n(\mathbf{x}',\mathbf{x}'')$ is the conditional probability that two points $(\mathbf{x}',\mathbf{x}'')$ are in the same cell with material property k_n, given that one point is in a cell with material property k_n. For symmetric cell materials $g_1 = g_2 = g$ and $f_1 = f_2 = f$.

In our bounds we also have the following correlation functions

$$\left\langle\frac{k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')}{k(\mathbf{x})}\right\rangle = \left[\frac{k_1'^3}{k_1}v - \frac{k_1'^3}{k_1}v^2 - \frac{k_1'^2k_2'}{k_2}v(1-v) + \frac{k_2'^3(1-v)}{k_2} - \frac{k_1'k_2'^2}{k_1}v(1-v) - \frac{k_2'^3}{k_2}(1-v)^2\right] \times g(\mathbf{x},\mathbf{x}',\mathbf{x}'') + \left[\frac{k_1'^3}{k_1}v^2 + \frac{k_1'^2k_2'}{k_2}v(1-v) + \frac{k_1'k_2'^2}{k_1}v(1-v) + \frac{k_2'^3}{k_2}(1-v)^2\right] f(\mathbf{x}',\mathbf{x}''). \tag{B3}$$

and

$$\left\langle\frac{k'^2(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')}{k(\mathbf{x})}\right\rangle = k_1'^4\left[\frac{v}{k_1} + \frac{v^4}{k_2(1-v)^3} - \frac{v^2}{k_1} - \frac{v^3}{k_2(1-v)} - \frac{v^3}{k_1(1-v)} - \frac{v^4}{k_2(1-v)^2}\right] g(\mathbf{x},\mathbf{x}',\mathbf{x}'') + k_1'^4\left[\frac{v^2}{k_1} + \frac{v^3}{k_2(1-v)} + \frac{v^3}{k_1(1-v)} + \frac{v^4}{k_2(1-v)^2}\right] \times f(\mathbf{x}',\mathbf{x}''). \tag{B4}$$

2. Four-point correlation functions of the form $\langle\kappa'(\mathbf{x})\kappa'(\mathbf{x}')\kappa'(\mathbf{x}'')\kappa'(\mathbf{x}''')\rangle$ and $\langle\kappa'(\mathbf{x}')\kappa'(\mathbf{x}'')\kappa'(\mathbf{x}''')/\kappa(\mathbf{x})\rangle$

After a great deal of algebra,[14] we find

$$\langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')\rangle = k_1'^4\,[v + v^4/(1-v)^3]\mathbf{g} + k_1'^4[v^2 + 2v^3/(1-v) + v^4/(1-v)^2]Q, \tag{B5}$$

and

$$\left\langle\frac{k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')}{k(\mathbf{x})}\right\rangle = \left(\frac{k_1'^3}{k_1}v - \frac{k_1'^3}{k_2}\frac{v^3}{(1-v)^2} - \frac{k_1'^3v^2}{k_1} - \frac{k_1'^3v(1-v)}{k_2} + \frac{k_1'^3v^3}{k_2(1-v)} + \frac{k_1'^3v^4}{k_1(1-v)^2}\right)\mathbf{g}(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''') + \left(\frac{k_1'^3v^2}{k_1} + \frac{k_1'^3v(1-v)}{k_2} - \frac{k_1'^3}{k_2}\frac{v^3}{(1-v)} - \frac{k_1'^3v^4}{k_1(1-v)^2}\right) \times g(\mathbf{x}',\mathbf{x}'',\mathbf{x}''') + \left(\frac{k_1'^3}{k_1}v^2 - \frac{k_1'^3v^3}{k_2(1-v)} + \frac{k_1'^3v^3}{k_1(1-v)} - \frac{k_1'^3v^2}{k_2}\right) \times Q(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}'''), \tag{B6}$$

$\mathbf{g}_n = \mathbf{g}_n(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''')$ is the conditional probability that four points are in the same cell given that one point is in a cell with material property k_n, and $Q_n = Q_n(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''')$ is the conditional probability that any two points are in one cell and the other two points are in another cell given that one point ($\mathbf{x}'''$) is in a cell with material property k_n.

For a symmetric cell material

$$\mathbf{g}_1 = \mathbf{g}_2 = \mathbf{g} \quad \text{and} \quad Q_1 = Q_2 = Q.$$

We also have the correlation function

$$\left\langle\frac{k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')}{k(\mathbf{x})}\right\rangle = \left(\frac{k_1'^4v}{k_1} - \frac{k_1'^3k_2'v^3}{k_2(1-v)^2} - \frac{k_1'^4v^2}{k_1} - \frac{k_1'^3k_2'}{k_2}v(1-v) + \frac{k_1'^3k_2'}{k_2}\frac{v^3}{(1-v)} + \frac{k_1'^4v^4}{k_1(1-v)^2}\right)\mathbf{g} + \left(\frac{k_1'^4v^2}{k_1} + \frac{k_1'^3k_2'}{k_2}v(1-v) - \frac{k_1'^3k_2'}{k_2}\frac{v^3}{(1-v)^2} - \frac{k_1'^4v^4}{k_1(1-v)^2}\right) \times g(\mathbf{x}',\mathbf{x}'',\mathbf{x}''') + \left(\frac{k_1'^4}{k_1}v^2 - \frac{k_1'^3k_2'}{k_2} + \frac{v^3}{(1-v)} + \frac{k_1'^4v^3}{k_1(1-v)} - \frac{k_1'^3k_2'v^2}{k_2}\right)Q. \tag{B7}$$

3. Five-point correlation functions of the form $\langle\kappa'(\mathbf{x})\,\kappa'(\mathbf{x}')\kappa'(\mathbf{x}'')\kappa'(\mathbf{x}''')\kappa'(\mathbf{x}'''')\rangle$ and $\langle\kappa'(\mathbf{x}')\kappa'(\mathbf{x}'')\kappa'(\mathbf{x}''')\kappa'(\mathbf{x}'''')/\kappa(\mathbf{x})\rangle$

We have

$$\langle k'(\mathbf{x})k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')k'(\mathbf{x}'''')\rangle = k_1'^5\left(v - \frac{v^5}{(1-v)^4}\right)\mathsf{g} + k_1'^5\left(v^2 + \frac{v^3}{1-v} - \frac{v^4}{(1-v)^2} - \frac{v^5}{(1-v)^3}\right)\mathbf{Q}, \tag{B8}$$

and

$$\left\langle\frac{k'(\mathbf{x}')k'(\mathbf{x}'')k'(\mathbf{x}''')k'(\mathbf{x}'''')}{k(\mathbf{x})}\right\rangle = \frac{k_1'^4}{k_1}v(1-v)\left(\frac{v^4}{(1-v)^4} - 1\right)\eta\;\mathsf{g}(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''') + \frac{k_1'^4}{k_1}v\left(1 + \eta - v\eta + \frac{v^3}{(1-v)^3} + \frac{\eta v^3}{(1-v)^2}\right) \times \mathbf{g}(\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''') + \frac{k_1'^4}{k_2}\eta\frac{v^2}{(1+\eta)}\left(\frac{v^2}{(1-v)^2} - 1\right)Q(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''') + \frac{k_1'^4}{k_1}\frac{v^2}{(1-v)^2}(1 + \eta - v\eta)Q(\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}''''), \tag{B9}$$

$\mathsf{g}_n = \mathsf{g}_n(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''')$ is the conditional probability that five points are in the same cell given that one point is in a cell with material property k_n, and $\mathbf{Q}_n = \mathbf{Q}_n(\mathbf{x},\mathbf{x}',\mathbf{x}'',\mathbf{x}''',\mathbf{x}'''')$ is the conditional probability that any three points are in one cell and the other two points are in another cell given that one point ($\mathbf{x}''''$) is in a cell with material property k_n.

For symmetric cell materials

$$\mathsf{g}_1 = \mathsf{g}_2 = \mathsf{g} \quad \text{and} \quad \mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{Q}.$$

APPENDIX C: RELATION BETWEEN THE CONSTANTS

As we saw in Sec. I of the paper, we obtained the upper and lower bounds in terms of the constants G, $\overline{G}$, G_2, $\overline{G}_2$, G_3, M_1, $\overline{M}_1$, and M_2. In the following we shall find the relations that exist between these constants.

In the expression for $\overline{G}$ [Eq. (38)], integrating by parts twice over x_3' and x_j', respectively, we have

$$\overline{G} = \frac{1}{(4\pi)^2}\int_{\mathbf{x}'}\frac{\partial}{\partial x_3'}\left(\int_{\mathbf{x}''}\frac{\partial}{\partial x_3''}g(\mathbf{x},\mathbf{x}',\mathbf{x}'')\frac{(x_j' - x_j'')}{|\mathbf{x}' - \mathbf{x}''|^3}\,d\mathbf{x}''\right) \times \frac{(x_j - x_j')}{|\mathbf{x} - \mathbf{x}'|^3}\,d\mathbf{x}'.$$

Changing the variables from $\mathbf{x}, \mathbf{x}', \mathbf{x}''$ to $\mathbf{0}, \mathbf{r}, \mathbf{s}$ where $\mathbf{x} = \mathbf{0}, \mathbf{r} = \mathbf{x}' - \mathbf{x}''$, and $\mathbf{s} = \mathbf{x} - \mathbf{x}'$, we have therefore $\partial/\partial x'_3 = \partial/\partial r_3 - \partial/\partial s_3$, $\partial/\partial x''_3 = -\partial/\partial r_3$ and

$$\begin{aligned}\overline{G} &= -\frac{1}{(4\pi)^2} \int_{\mathbf{s}} \left(\frac{\partial}{\partial r_3} - \frac{\partial}{\partial s_3}\right) \left(\int_{\mathbf{r}} \frac{\partial}{\partial r_3} g(\mathbf{0}, \mathbf{r}, \mathbf{s}) \frac{r_j}{r^3}\, d\mathbf{r}\right) \frac{s_j}{s^3}\, d\mathbf{s} \\ &= -\frac{1}{(4\pi)^2} \int_{\mathbf{s}} \frac{\partial}{\partial r_3} \left(\int_{\mathbf{r}} \frac{\partial}{\partial r_3} g(\mathbf{0}, \mathbf{r}, \mathbf{s}) \frac{r_j}{r^3} d\mathbf{r}\right) \frac{s_j}{s^3}\, d\mathbf{s} \\ &\quad + \frac{1}{(4\pi)^2} \int_{\mathbf{r}} \int_{\mathbf{s}} \frac{\partial}{\partial r_3} \frac{\partial}{\partial s_3} g(\mathbf{0}, \mathbf{r}, \mathbf{s}) \frac{r_j}{r^3} \frac{s_j}{s^3}\, d\mathbf{r} d\mathbf{s}. \qquad \text{(C1)}\end{aligned}$$

The first integral is zero; since the integral between brackets is a function of $\mathbf{0}$ and $\mathbf{s}$. In the expression for G [Eq. (36)] changing the variables to $\mathbf{0}, \mathbf{t}$, and $\mathbf{w}$ where $\mathbf{x} = \mathbf{0}, \mathbf{t} = \mathbf{x} - \mathbf{x}'$, and $\mathbf{w} = \mathbf{x} - \mathbf{x}''$, we obtain

$$G = \frac{1}{(4\pi)^2} \int_{\mathbf{t}} \int_{\mathbf{w}} \frac{\partial}{\partial t_3} \frac{\partial}{\partial w_3} g(\mathbf{0}, \mathbf{t}, \mathbf{w}) \frac{t_j}{t^3} \frac{w_j}{w^3}\, d\mathbf{t} d\mathbf{w}, \qquad \text{(C2)}$$

which is equal to $\overline{G}$ [see Eq. (C1)].

Therefore $G = \overline{G}$.

Similarly in Eq. (38), if we let $\mathbf{x} = \mathbf{x}', \mathbf{x}' = \mathbf{x}''$, and $\mathbf{x}'' = \mathbf{x}'''$, we obtain

$$\begin{aligned}\overline{G} = \frac{1}{(4\pi)^2} \int_{\mathbf{x}''} \frac{\partial}{\partial x''_j} \left(\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_3} g(\mathbf{x}', \mathbf{x}'', \mathbf{x}''') \frac{(x''_j - x'''_j)}{|\mathbf{x}'' - \mathbf{x}'''|^3}\, d\mathbf{x}'''\right) \\ \times \frac{(x'_3 - x''_3)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}''. \qquad \text{(C3)}\end{aligned}$$

Because of isotropy we can write $\overline{\overline{G}}$ [Eq. (40)] as

$$\begin{aligned}\overline{\overline{G}} &= \frac{1}{3(4\pi)^3} \int_{\mathbf{x}'} \frac{\partial}{\partial x'_3} \left[\int_{\mathbf{x}''} \frac{\partial}{\partial x''_j} \left(\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_3} g(\mathbf{x}', \mathbf{x}'', \mathbf{x}''')\right.\right. \\ &\quad \left.\left. \times \frac{(x''_j - x'''_j)}{|\mathbf{x}'' - \mathbf{x}'''|^3}\, d\mathbf{x}'''\right) \frac{(x'_3 - x''_3)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}''\right] \frac{(x_3 - x'_3)}{|\mathbf{x} - \mathbf{x}'|^3}\, d\mathbf{x}' \\ &= \frac{1}{(4\pi)^3} \int_{\mathbf{x}'} \frac{\partial}{\partial x'_i} \left[\int_{\mathbf{x}''} \frac{\partial}{\partial x''_j} \left(\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_3} g(\mathbf{x}', \mathbf{x}'', \mathbf{x}''')\right.\right. \\ &\quad \left.\left. \times \frac{(x''_j - x'''_j)}{|\mathbf{x}'' - \mathbf{x}'''|^3}\, d\mathbf{x}'''\right) \frac{(x'_3 - x''_3)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}''\right] \frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3}\, d\mathbf{x}'.\end{aligned}$$

In the first integration, integrating by parts over x_i, we obtain

$$\begin{aligned}\overline{\overline{G}} &= -\frac{1}{(4\pi)^3} \int_{\mathbf{x}'} \left[\int_{\mathbf{x}''} \frac{\partial}{\partial x''_j} \left(\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_3} g(\mathbf{x}', \mathbf{x}'', \mathbf{x}''')\right.\right. \\ &\quad \left.\left. \times \frac{(x''_j - x'''_j)}{|\mathbf{x}'' - \mathbf{x}'''|^3}\, d\mathbf{x}'''\right) \frac{(x'_3 - x''_3)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}''\right] \frac{\partial}{\partial x'_i} \frac{(x_i - x'_i)}{|\mathbf{x} - \mathbf{x}'|^3}\, d\mathbf{x}'.\end{aligned}$$

Realizing that $(x_i - x'_i)/|\mathbf{x} - \mathbf{x}'|^3 = (\partial/\partial x'_i)(1/|\mathbf{x} - \mathbf{x}'|)$ and that $1/|\mathbf{x} - \mathbf{x}'|$ is the free space Green's function for Laplace's operator $\partial^2/\partial x'^2_i$, we can write

$$\begin{aligned}\overline{\overline{G}} = \frac{1}{(4\pi)^2} \int_{\mathbf{x}''} \frac{\partial}{\partial x''_j} \left(\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_3} g(\mathbf{x}', \mathbf{x}'', \mathbf{x}''')\right. \\ \left. \times \frac{(x''_j - x'''_j)}{|\mathbf{x}'' - \mathbf{x}'''|^3}\, d\mathbf{x}'''\right) \frac{(x'_3 - x''_3)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}'', \qquad \text{(C4)}\end{aligned}$$

which is equal to $\overline{G}$ [see Eq. (C3)].

Therefore

$$G = \overline{G} = \overline{\overline{G}}. \qquad \text{(C5)}$$

Also in Eq. (42) writing $\mathbf{x} = \mathbf{x}'$, $\mathbf{x}' = \mathbf{x}''$, $\mathbf{x}'' = \mathbf{x}'''$, we have

$$\begin{aligned}G_2 &= \frac{1}{(4\pi)^3} \int_{\mathbf{x}''} \frac{\partial}{\partial x''_3} \left[\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_j} \left(\int_{\mathbf{x}''''} \frac{\partial}{\partial x''''_3} \mathrm{g}(\mathbf{x}', \mathbf{x}'', \mathbf{x}''', \mathbf{x}'''')\right.\right. \\ &\quad \left.\left. \times \frac{(x'''_j - x''''_j)}{|\mathbf{x}''' - \mathbf{x}''''|^3}\, d\mathbf{x}''''\right) \frac{(x''_i - x'''_i)}{|\mathbf{x}'' - \mathbf{x}'''|^3}\, d\mathbf{x}'''\right] \frac{(x'_i - x''_i)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}''. \qquad \text{(C6)}\end{aligned}$$

Again because of isotropy we can write Eq. (45) as

$$\begin{aligned}\overline{G}_2 &= \frac{1}{(4\pi)^4} \int_{\mathbf{x}'} \frac{\partial}{\partial x'_j} \Big\} \int_{\mathbf{x}''} \frac{\partial}{\partial x''_3} \left[\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_m} \left(\int_{\mathbf{x}''''} \frac{\partial}{\partial x''''_3}\right.\right. \\ &\quad \left.\left. \times \mathrm{g}(\mathbf{x}', \mathbf{x}'', \mathbf{x}''', \mathbf{x}'''') \frac{(x'''_m - x''''_m)}{|\mathbf{x}''' - \mathbf{x}''''|^3}\, d\mathbf{x}''''\right) \frac{(x''_i - x'''_i)}{|\mathbf{x}'' - x''''|^3}\, d\mathbf{x}'''\right] \\ &\quad \times \frac{(x'_i - x''_i)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}'' \Big\{ \frac{(x_j - x'_j)}{|\mathbf{x} - \mathbf{x}'|^3}\, d\mathbf{x}'.\end{aligned}$$

Similarly, integrating by parts over x'_j we find

$$\begin{aligned}\overline{G}_2 &= \frac{1}{(4\pi)^3} \int_{\mathbf{x}''} \frac{\partial}{\partial x''_3} \left[\int_{\mathbf{x}'''} \frac{\partial}{\partial x'''_m} \left(\int_{\mathbf{x}''''} \frac{\partial}{\partial x''''_3} \mathrm{g}(\mathbf{x}', \mathbf{x}'', \mathbf{x}''', \mathbf{x}'''')\right.\right. \\ &\quad \left.\left. \times \frac{(x'''_m - x''''_m)}{|\mathbf{x}''' - \mathbf{x}''''|^3}\, d\mathbf{x}''''\right) \frac{(x''_i - x'''_i)}{|\mathbf{x}'' - \mathbf{x}'''|^3}\, d\mathbf{x}'''\right] \\ &\quad \times \frac{(x'_i - x''_i)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}'',\end{aligned}$$

which is equal to G_2 [see Eq. (C6)].

Therefore $G_2 = \overline{G}_2$

Similarly we can show that

$$M_1 = \overline{M}_1.$$

APPENDIX D

In this appendix we will prove that the integral

$$t = \int_{\mathbf{x}'} \frac{\partial}{\partial x'_j} \left(\int_{\mathbf{x}''} \frac{\partial}{\partial x''_3} f(\mathbf{x}', \mathbf{x}'') \frac{(x'_j - x''_j)}{|\mathbf{x}' - \mathbf{x}''|^3}\, d\mathbf{x}''\right) \frac{(x_3 - x'_3)}{|\mathbf{x} - \mathbf{x}'|^3}\, d\mathbf{x}' \qquad \text{(D1)}$$

is equal to zero.

Let $\mathbf{x} = 0$, $\mathbf{r} = \mathbf{x} - \mathbf{x}'$, and $\mathbf{s} = \mathbf{x}' - \mathbf{x}''$. Changing the variables from $\mathbf{x}'$ and $\mathbf{x}''$ to $\mathbf{r}$ and $\mathbf{s}$, we can write

$$\frac{\partial}{\partial x_j} = -\frac{\partial}{\partial r_j} + \frac{\partial}{\partial s_j}$$

and

$$\frac{\partial}{\partial x''_3} = -\frac{\partial}{\partial s_3};$$

therefore $f(\mathbf{x}', \mathbf{x}'') = f[-\mathbf{r}, -(\mathbf{r} + \mathbf{s})]$ and for a homogeneous and isotropic material, $f[-\mathbf{r}, -(\mathbf{r} + \mathbf{s})] = f(-\mathbf{s}) = f(s)$. Equation (D1) gives

$$t = \int_{\mathbf{r}} \int_{\mathbf{s}} \frac{\partial^2 f(s)}{\partial r_j \partial s_3} \frac{s_j}{s_3}\, d\mathbf{s} \frac{r_3}{r^3}\, d\mathbf{r}$$

$$-\int_{\mathbf{r}} [\frac{\partial}{\partial s_j} \int_{\mathbf{s}} \frac{\partial f(s)}{\partial s_3} \frac{s_j}{s^3} d\mathbf{s}] \frac{r_3}{r^3} d\mathbf{r}.$$

The first integral is obviously zero. The second integral goes to zero by integrating over $\mathbf{r}$. Therefore $t = 0$.

* Extracted from a dissertation submitted in partial fulfillment of Doctor of Philosophy Degree, University of Pennsylvania, Philadelphia, Pennsylvania.

† Present address: The Catholic University of America, Washington, D.C.

[1] Z. Hashin, "Theory of Fiber Reinforced Materials", Contract NAS1-8818, NASA (University of Pennsylvania, 1970).

[2] Z. Hashin and S. Shtrikman, J. Mech. Phys. Solids **11**, 127 (1963).

[3] Z. Hashin, J. Mech. Phys. Solids **13**, 119 (1965).

[4] M. Beran, Nuovo Cimento **38**, 771 (1965).

[5] M. Beran and J. Molyneaux, Quart. Appl. Math. **24**, 107 (1966).

[6] P. Corson, PhD thesis (University of Pennsylvania, 1971).

[7] M. Miller, J. Math. Phys. **10**, 1988 (1969).

[8] M. Beran and N. Silnutzer, J. Comp. Mater. **5**, 246 (1971).

[9] N. Silnutzer, PhD thesis (University of Pennsylvania, 1972).

[10] M. Beran, *Statistical Continuum Theories* (Interscience, New York, 1968).

[11] R. Hoffman, Proc. Symp. Appl. Math. **16**, 117 (1964).

[12] J. Kampe De Feriet, Proc. Symp. Appl. Math. **13**, 165 (1962).

[13] M. Beran and J. Molyneux, Nuovo Cimento **30**, 1406 (1963).

[14] M. Elsayed, PhD thesis (University of Pennsylvania, 1972).

[15] H. L. Frisch, Trans. Soc. Rheal. **9**, 293 (1965).

[16] E. N. Gilbert, Ann. Math. Stat. **33**, 958 (1962).

[17] The proof was supplied by the reviewer.

[18] Reynolds and Hough, Proc. Phys. Soc. London **70**, 769 (1957).

[19] J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941).

[20] B. Budiansky, J. Mech. Phys. Solids **13**, 223 (1965).

[21] R. Hill, J. Mech. Phys. Solids **13**, 213 (1965).

[22] Z. Hashin, "Theory of Composite Materials", Mechanics of Composite Materials, Proc. Fifth Symp. Naval Structural Mech., Philadelphia, 1970, p.201.

Representations of multidimensional symmetries in networks

W. G. Harter

Instituto de Fisica, Universidade Estadual de Campinas, C.P. 1170 Campinas S.P., Brazil

(Received 29 June 1973)

Physical systems that have resonances corresponding to representations of multidimension symmetry groups can be constructed from electric circuit elements. Examples involving symmetries of two four-dimensional polytopes are shown. Also a group theoretical analysis of linear constraints is described.

I. INTRODUCTION

Characters of four-dimensional cubic symmetry were calculated on computer by Birman and Chen, [1] who speculated that these representations and symmetries might possibly be associated with accidental degeneracies in some crystal lattice vibration frequencies. In the following we demonstrate some interesting realizations of these representations and degeneracies in the vibrations of certain electrical networks, and suggest what other sorts of symmetries can be visualized and treated similarly.

In fact, the vector representations given by Birman were found by Littlewood over thirty years before,[2] but it was not made clear by Littlewood whether he knew or cared about higher point symmetries, since his main concern was the study of the permutation group and its subgroups. Nevertheless, his methods are general enough to produce the characters of any one of such higher symmetries. (Higher point symmetries are effectively catalogued by the existing regular polytopes as listed in Appendix C.)

The treatment of complex symmetric networks contained below is a straightforward extension of the usual group projection techniques,[3,4] except that one must take account of the Kirchhoff current conservation constraints. A group theoretical method for deriving the constraint effects is described in Sec. II in connection with an example that can also be treated conventionally.

In Sec. III the results of the group analysis are displayed in the form of current-flow illustrations for the elementary resonances on oscillating networks having the connectivity of the four-dimensional cube and tetrahedron. The correspondence of the high degeneracies found in each case with higher symmetry representations is demonstrated.

It is apparently incorrect to claim that such analyses fill a need in circuit engineering since probably no laboratory has considered such network configurations. It is better that we simply offer the examples as interesting diversions, and the methods as solutions awaiting a problem.

II. GROUP THEORETICAL ANALYSIS OF KIRCHHOFF CONSTRAINTS

The coordinates that describe the state of internal currents in an electric network must be chosen to be independent. The twelve coordinates $\{j_1 \ldots j_{12}\}$ indicated in Fig. 1 (a) are too many, since the number of degrees of freedom of this network is seven. In general, Kirchhoff current conservation constraints reduce the number of independent coordinates of a b-branched closed network to $b-n+1$, where n is the number of nodes or junctions.[5]

For networks that are planar like the example in Fig. 1, it is possible to choose exactly the right number of mesh loops to be the independent coordinates as is done in Fig. 1, but it is sometimes not convenient to define mesh loops for planar or especially, nonplanar networks. However if the network possesses some topological symmetry a group theoretical coordinate definition may be more convenient in either case. For example, to define coordinates for the network in Fig. 1 the irreducible representations (IR) of the cubic-octahedral group O_h characterized by Table I, are employed. Each group operator is labeled by its effect on the Cartesian coordinates $(xyz)[(y\bar{x}\bar{z})$: $x \rightarrow y, y \rightarrow -x, z \rightarrow -z]$, and, in turn, their effect on any of the 12 (nonconservative) current states $|j_n\rangle$ is obtained by inspection as shown in Eq. (1) ($|j_n\rangle$ is the state in which unit current is flowing in branch n)

$$(y\bar{x}\bar{z})|j_1\rangle = -|j_8\rangle \cdots (y\bar{x}\bar{z})|j_{12}\rangle = -|j_9\rangle. \tag{1}$$

From these are obtained orthonormal vectors of Eq. (2) that transform irreducibly as per Eq. (3):

$$|\Psi_j{}^\beta\rangle = \sum_n^{12} |j_n\rangle\langle j_n|\Psi_j{}^\beta\rangle, \tag{2}$$

$$\langle\Psi_i^\alpha|(g)|\Psi_j^\beta\rangle = \delta^{\alpha\beta}\mathfrak{D}_{ij}^{(\alpha)}(g). \tag{3}$$

(The standard procedures that accomplish this are sketched in Appendix B.)

The coefficients $\langle j_n|\Psi_j^\beta\rangle$ define currents in the diagrams of Fig. 2 and it is seen there that some IR bases conserve currents while others do not. The seven conservative bases may replace the seven mesh loops of Fig. 1(b). In fact relations like Eq. (4) are obtained by inspection of Fig. 2:

$$|l_1\rangle = \frac{3|\Psi^{A_{2g}}\rangle}{2\sqrt{3}} - \frac{(|\Psi_1^{T_{1g}}\rangle + |\Psi_2^{T_{1g}}\rangle + |\Psi_3^{T_{1g}}\rangle}{2} + \frac{2(|\Psi_1^{T_{2g}}\rangle + |\Psi_2^{T_{2g}}\rangle + |\Psi_3^{T_{2g}}\rangle}{2\sqrt{2}}. \tag{4}$$

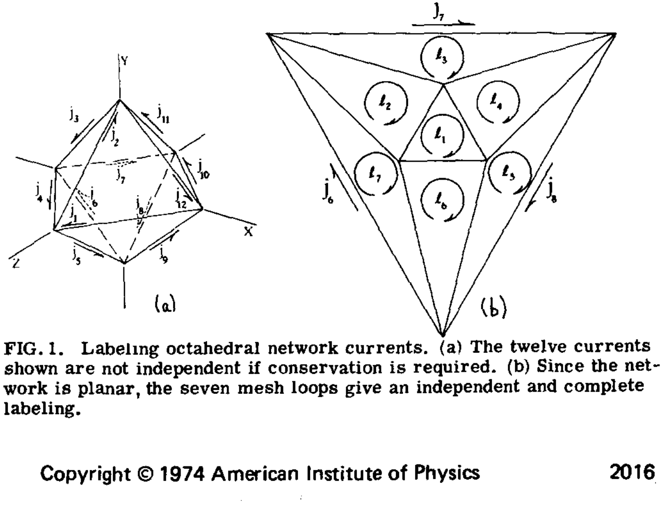


FIG. 1. Labeling octahedral network currents. (a) The twelve currents shown are not independent if conservation is required. (b) Since the network is planar, the seven mesh loops give an independent and complete labeling.

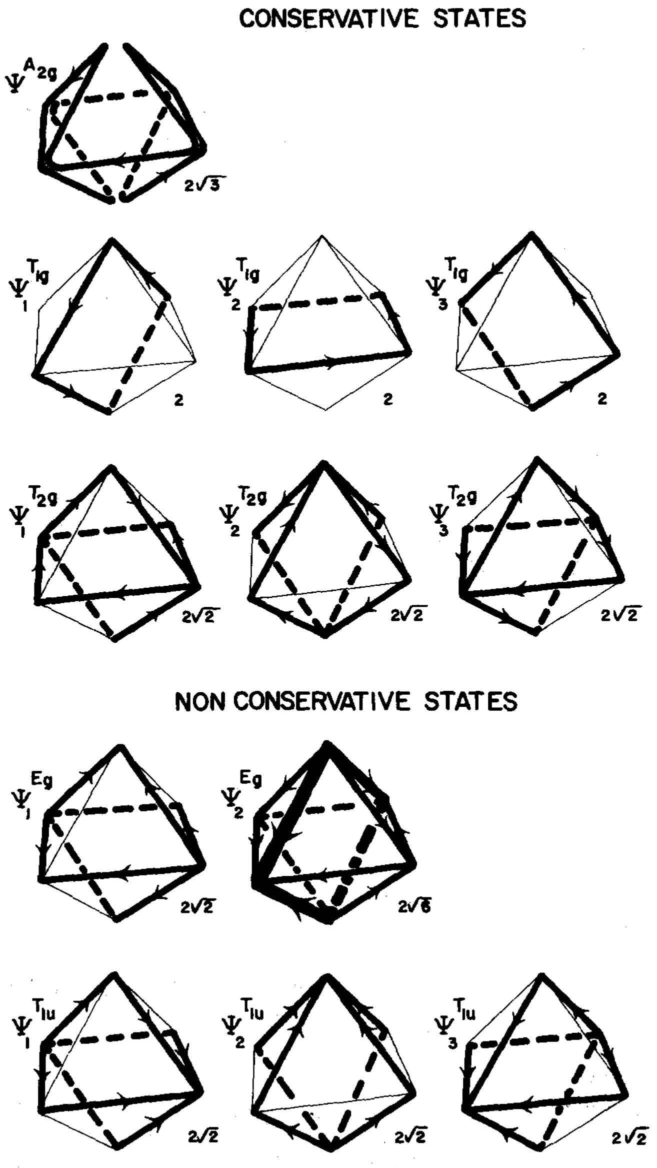


FIG. 2. Independent conservative and nonconservative states for octahedral network. Directed arrows in all configurations except $\Psi_2^{E_g}$ represent unit current flow. Thicker arrow in the latter represents twice unit flow. Normalization denominators are shown under each figure.

A nonconservative current state will in general require all 12 states in Fig. 2 as, for example, does $|j_1\rangle$ in Eq. (5):

$$|j_1\rangle = \frac{|\Psi^{A_{2g}}\rangle}{2\sqrt{3}} - \frac{|\Psi_2^{T_{1g}}\rangle}{2} + \frac{|\Psi_1^{T_{2g}}\rangle + |\Psi_3^{T_{2g}}\rangle}{2\sqrt{2}} + \frac{|\Psi_1^{E_g}\rangle}{2\sqrt{2}} + \frac{|\Psi_2^{E_g}\rangle}{2\sqrt{6}} - \frac{(|\Psi_1^{T_{1u}}\rangle - |\Psi^{T_{1u}}\rangle)}{2\sqrt{2}}. \tag{5}$$

The question of which sets are conservative can be deduced abstractly in such a way that one can see the general results of applying symmetric linear constraints like the Kirchhoff current conservation relations.

The current conservation constraints for the octahedral network are linear in j_n and can be written in rectangular matrix form of Eq. (6), where zeroing of $c_n (n = 1, \ldots, 6)$ implies conservation:

$$\boxed{\qquad K \qquad} \begin{vmatrix} j_1 \\ j_2 \\ \cdot \\ \cdot \\ \cdot \\ j_{12} \end{vmatrix} = \begin{vmatrix} c_1 \\ c_2 \\ \cdot \\ c_6 \end{vmatrix}. \tag{6}$$

These constraints presumably have the topological symmetry of the network and this is expressed by a generalized commutation relation

$$K \cdot J(g) = C(g) \cdot K. \tag{7}$$

In the above, $J(g)$ is a 12×12 matrix that represents transformation of branch currents by group operation g of O_h following Eq. (1). $C(g)$ is an analogous 6×6 matrix that represents transformation of vertices. Both J and C are reducible, the former into $A_{2g} + T_{1g} + T_{2u} + E_g + T_{1u}$ (see Fig. 2) and the latter into $A_{1g} + E_g + T_{1u}$, and these are indicated in Eq. (8):

$$\begin{aligned} V^{-1} J V &= A_{2g} + T_{1g} + T_{2u} + E_g + T_{1u}, \\ U^{-1} C U &= A_{1g} + E_g + T_{1u}. \end{aligned} \tag{8}$$

The columns of V are the previously mentioned [Eq.

TABLE I. Character table of three-dimensional cubic octahedral symmetry group O_h. Polynomials corresponding to IR's of O_h are given.

	(xyz)	(yzx) $(y\bar{z}\bar{x})$ $(\bar{y}z\bar{x})$ $(\bar{y}\bar{z}x)$ (zxy) $(\bar{z}x\bar{y})$ $(\bar{z}\bar{x}y)$ $(z\bar{x}\bar{y})$	$(x\bar{y}\bar{z})$ $(\bar{x}y\bar{z})$ $(\bar{x}\bar{y}z)$	$(xz\bar{y})$ $(\bar{z}yx)$ $(y\bar{x}z)$ $(x\bar{z}y)$ $(zy\bar{x})$ $(\bar{y}xz)$	$(z\bar{y}x)$ $(\bar{z}\bar{y}\bar{x})$ $(yx\bar{z})$ $(\bar{y}\bar{x}\bar{z})$ $(\bar{x}\bar{z}\bar{y})$ $(\bar{x}zy)$	$(\bar{x}\bar{y}\bar{z})$	$(\bar{y}\bar{z}\bar{x})$ $(\bar{y}zx)$ $(y\bar{z}x)$ $(yz\bar{x})$ $(\bar{z}\bar{x}\bar{y})$ $(z\bar{x}y)$ $(zx\bar{y})$ $(\bar{z}xy)$	$(\bar{x}yz)$ $(x\bar{y}z)$ $(xy\bar{z})$	$(\bar{x}\bar{z}y)$ $(z\bar{y}\bar{x})$ $(\bar{y}x\bar{z})$ $(\bar{x}z\bar{y})$ $(\bar{z}\bar{y}x)$ $(y\bar{x}\bar{z})$	$(\bar{z}y\bar{x})$ (zyx) $(\bar{y}\bar{x}z)$ (yxz) (xzy) $(x\bar{z}\bar{y})$	
A_{1g}	1	1	1	1	1	1	1	1	1	1	$x^2 + y^2 + z^2$
A_{2g}	1	1	1	−1	−1	1	1	1	−1	−1	$x^4(y^2 - z^2) + y^4(z^2 - x^2) + z^4(x^2 - y^2)$
E_g	2	−1	2	0	0	2	−1	2	0	0	$2z^2 - x^2 - y^2, \sqrt{3}(x^2 - y^2)$
T_{1g}	3	0	−1	1	−1	3	0	−1	1	−1	$y^3z - z^3y, z^3x - x^2z, x^3y - y^3x$
T_{2g}	3	0	−1	−1	1	3	0	−1	−1	1	yz, xz, xy
A_{1u}	1	1	1	1	1	−1	−1	−1	−1	−1	$x(y^3z - z^3y) + y(z^3x - x^3y) + z(x^3y - y^3x)$
A_{2u}	1	1	1	−1	−1	−1	−1	−1	1	1	xyz
E_u	2	−1	2	0	0	−2	1	−2	0	0	$\sqrt{3}xyz(x^2 - y^2), xyz(2z^2 - x^2 - y^2)$
T_{1u}	3	0	−1	1	−1	−3	0	1	−1	1	x, y, z
T_{2u}	3	0	−1	−1	1	−3	0	1	1	−1	$x(y^2 - z^2), y(z^2 - x^2), z(x^2 - y^2)$
	identity	± 120° rotation class	180° rotation class	± 90° rotation class	180° rotation class	inversion	± 120° rotation inverstion class	mirror reflection class	± 90° rotation inversion class	mirror reflection class	
			0								

(2)] current vectors $|\Psi_j^{(\alpha)}\rangle$. Now, rewritting Eq. (7) as Eq. (9) below,

$$U^{-1}K\ V\ V^{-1}\ J(g)V = U^{-1}\ C(g)\ U\ U^{-1}\ K\ V, \tag{9}$$

and applying Schur's lemmas[6] to selected block submatrices of the matrix $U^{-1}\ K\ V$, one proves that the latter must have the form:

$$U^{-1}\ K\ V = \begin{array}{ccccccc|ccccc|l}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \left.\right] A_{1g} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \alpha & \cdot & \cdot & \cdot & \cdot & \left.\right] E_g \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \alpha & \cdot & \cdot & \cdot & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \beta & \cdot & \cdot & \left.\right] T_{1u} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \beta & \cdot & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \beta & \\
\hline
A_{2g} & \multicolumn{3}{c}{T_{1g}} & \multicolumn{3}{c}{T_{2g}} & \multicolumn{2}{c}{E_g} & \multicolumn{3}{c}{T_{1u}} &
\end{array}\ . \tag{10}$$

Finally one obtains the following:

$$K\ V = \boxed{\begin{array}{c|c} \text{submatrix} & \text{submatrix} \\ \text{guaranteed} & \text{not guaranteed} \\ \text{zero} & \text{zero} \end{array}} \tag{11}$$

which, if compared with Eq. (6), is seen to state explicitly that the first seven current states (Fig. 2) are con servative while the remainder may not be.

For the preceding analysis of constraints to be valuable, one only needs to know some topological symmetry of the network in question, which in turn presumably corresponds to the symmetry of the constraints. If in addition the equation of motion for transient currents in the network has this same symmetry, then the conservative IR bases (Fig. 2) will be the normal modes or elementary resonances of the network. In this latter case we can say that the physical symmetry is the same as the topological symmetry. When the physical symmetry is lower than the topological symmetry some mixing of the conservative states may be necessary to produce the resonant modes.

Also mixing will be necessary for repeated equivalent IR's should they appear in columns or rows of relations like Eq. (8). The procedures for dealing with these occurrences are straightforward.

III. EXAMPLES OF NONPLANAR NETWORKS

The cubic configuration shown in Fig. 3 has 17 conservative degrees of freedom, but it is not immediately clear how 17 independent loops could be drawn into the 32 branches. However, the IR of O_h, which correspond to conservative states, are easily found (Fig. 4). The IR's T_{1g} and E_u both appear twice, and so one is at first free to pick arbitrary orthogonal combinations of the pair of T_{1g} – IR's and similarly for the E_u.

A most interesting application of this involves finding the frequencies of normal vibration of a linear tank circuit constructed upon this network.

The equations of motion are a coupled set of 32 differential equations, the first of which is given by

$$\frac{d^2 i_1}{dt^2} = ai_1 + bi_2 - ci_3 + bi_4 + bi_5 - ci_6 + bi_7 - bI_1 + bI_7 + cj_1. \tag{12}$$

The coefficients a, b, etc. are assumed to be constants dependent upon impedance values of branches and arranged so that the *physical* symmetry is O_h.

If $b = b'$ and $c = c'$ the currents drawn in Fig. 4 are in precisely the right proportion to decouple the 32

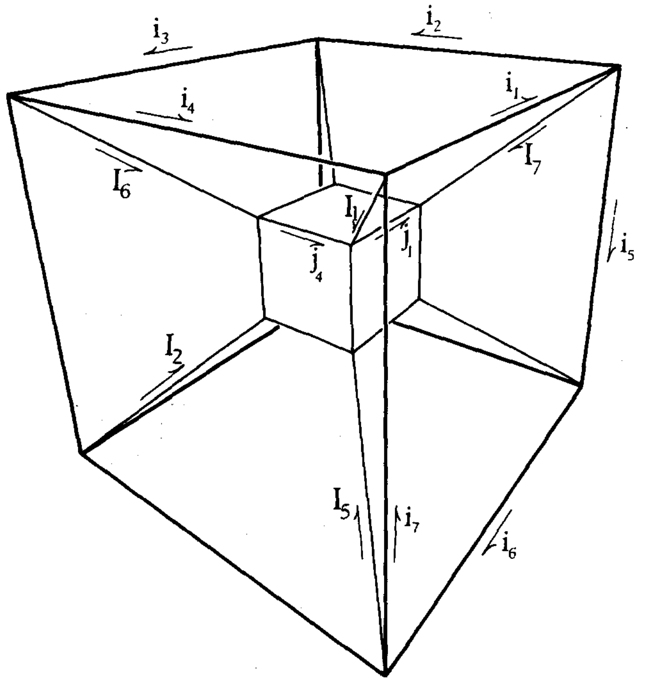


FIG. 3. Nonplanar network having cubic symmetry. The 32 currents shown are not independent. Furthermore, the mesh loop procedure successful in Fig. 1 cannot be applied here.

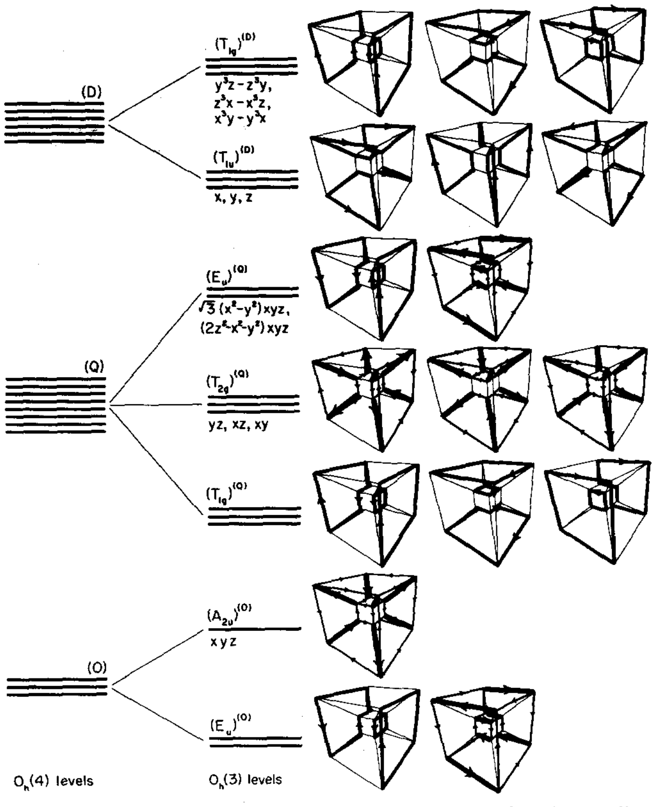


FIG. 4. Independent conservative states of cubic network and level diagrams for modes under three- and four-dimensional cubic symmetry. As shown in the text, the higher degeneracies can be traced to certain IR's of the higher symmetry.

equations of motion. The resulting equation for each of the first six modes (the D modes) is

$$\frac{d^2D}{dt^2} = (a + 2b + c)D. \tag{13}$$

The next eight modes (labeled Q modes) all satisfy another equation:

$$\frac{d^2Q}{dt^2} = (a + 2b - c)Q. \tag{14}$$

The remaining three modes have still another equation.

$$\frac{d^2O}{dt^2} = (a + 2b - 3c)O. \tag{15}$$

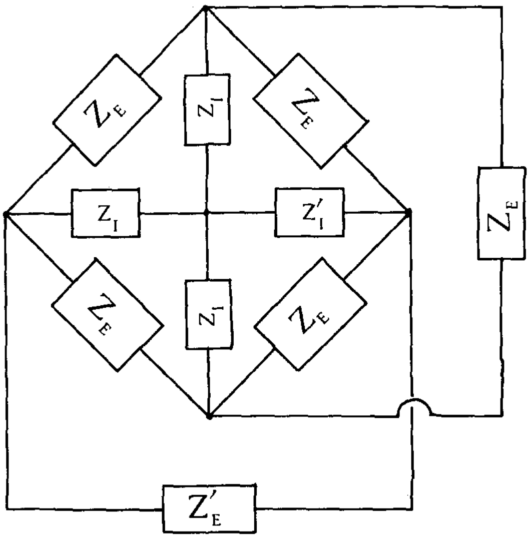


FIG. 5. Impedance bridge. Each impedance block Z_n represents series effective capacitance and inductance C_n and L_n, respectively. This is the simplest example of a nonplanar network.

The degeneracies in frequency of $(T_{1g})^{(D)}$ and $(T_{1u})^{(D)}$ at $\omega^D = \sqrt{a} + 2b + c$, of $(E_u)^{(Q)}$, $(T_{2g})_{(Q)}$, and $(T_{1g})^{(Q)}$ at $\omega^Q = \sqrt{a} + 2b - c$, and of $(A_{2u})^{(O)}$ and $(E_u)^{(O)}$ at $\omega^{(O)} = \sqrt{a} + 2b - 3c$ might seem unexpected (accidental) but one can prove that they correspond to an IR of the four-dimensional cubic octahedral group $O_h(4)$.

This is accomplished shortly after one realizes that the order of the group $O_h^{(4)}$ must be 384 (Appendix C), for there exists a subgroup of S_8 of order 384 which Littlewood has found, along with a great number of other groups that he has listed.[2] With a bit of patience one may finally sort and identify the characters and classes of Littlewood's group with this group of higher cubic symmetry. The result is tabulated below (Table II) and comparison with Table I verifies the degeneracies and splittings shown in the level diagrams of Fig. 4.

As a final example, consider the simplest nonplanar network: the well known Wheatstone bridge in Fig. 5.

The topological symmetry is S_5, which happens to be isomorphic to the four-dimensional tetrahedral symmetry. The physical symmetry of the bridge depends, of course, upon the values of the impedances, and examples varying from S_5 to S_1 are shown in Fig. 6. The IR of S_n are labeled by Young tableaux.

ACKNOWLEDGMENTS

I would like to thank John E. Heighway, Jan Smit, and Martin Gundersen for interesting conversations concerning these objects of thought.

APPENDIX A: CONSTRUCTING IR FROM POLYNOMIALS

Polynomials in $x^\alpha y^\beta z^\gamma$ that form the bases of IR of O_h are given in Table I. A norm is defined for these bases as follows:

$$\langle x^\alpha y^\beta z^\gamma | x^{\alpha'} y^{\beta'} z^{\gamma'} \rangle = \delta_{\alpha\alpha'}\delta_{\beta\beta'}\delta_{\gamma\gamma'}. \tag{A1}$$

TABLE II. The character table of four-dimensional "cubic-octahedral" symmetry $O_h(4)$. The four-dimensional cube has eight volumes that will be permuted one into the other by the rotations of $O_h(4)$, hence the latter is isomorphic to a subgroup of S_8. Littlewood's procedure is used to derive the characters.

		identity	± 120° rotations	180 rotations	± 90° rotations	180° rotations	inversion	± 120° rotation inversions	mirror reflections	± 90° rotation inversions	mirror reflections	classes of $O_h(3)$ found in $O_h(4)$ classes										$O_h(3)$ content of $O_h(4)$ IR
IR of $O_h(4)$	A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	A_{1g}
	B	1	1	1	−1	−1	−1	−1	−1	1	1	1	1	1	−1	−1	−1	−1	1	1	1	A_{2u}
	C	6	0	−2	−2	2	0	0	0	0	0	6	0	2	−2	0	0	0	−2	0	0	$T_{2u} + T_{2g}$
	D	6	0	−2	2	−2	0	0	0	0	0	6	0	2	2	0	0	0	−2	0	0	$T_{1u} + T_{1g}$
	E	4	1	0	2	0	−2	1	2	0	2	−4	−1	0	−2	0	0	−1	0	0	−2	$A_{1g} + T_{1u}$
	F	4	1	0	−2	0	2	−1	−2	0	2	−4	−1	0	2	0	0	1	0	0	−2	$A_{2u} + T_{2g}$
	G	4	1	0	−2	0	−2	1	2	0	−2	−4	−1	0	2	0	0	−1	0	0	2	$A_{2g} + T_{2u}$
	H	4	1	0	2	0	2	−1	−2	0	−2	−4	−1	0	−2	0	0	1	0	0	2	$A_{1u} + T_{1g}$
	I	6	0	−2	0	0	0	0	0	−2	2	6	0	−2	0	0	0	0	2	0	2	$T_{2g} + T_{1u}$
	J	6	0	−2	0	0	0	0	0	2	−2	6	0	−2	0	0	0	0	2	0	−2	$T_{1g} + T_{2u}$
	K	2	−1	2	0	0	2	−1	2	0	0	2	−1	2	0	0	2	−1	2	0	0	E_g
	L	2	−1	2	0	0	−2	1	−2	0	0	2	−1	2	0	0	−2	1	2	0	0	E_u
	M	3	0	3	1	1	−3	0	−3	−1	−1	3	0	−1	1	−1	1	0	−1	1	−1	$A_{1u} + E_u$
	N	3	0	3	−1	−1	3	0	3	−1	−1	3	0	−1	−1	1	−1	0	−1	1	−1	$A_{2g} + E_g$
	O	3	0	3	−1	−1	−3	0	−3	1	1	3	0	−1	−1	1	1	0	−1	−1	1	$A_{2u} + E_u$
	P	3	0	3	1	1	3	0	3	1	1	3	0	−1	1	−1	−1	0	−1	−1	1	$A_{1g} + E_g$
	Q	8	−1	0	0	0	4	1	−4	0	0	−8	1	0	0	0	0	−1	0	0	0	$T_{1g} + T_{2g} + E_u$
	R	8	−1	0	0	0	−4	−1	4	0	0	−8	1	0	0	0	0	1	0	0	0	$T_{1u} + T_{2u} + E_g$
	S	1	1	1	−1	−1	1	1	1	−1	−1	1	1	1	−1	−1	1	1	1	−1	−1	A_{2g}
	T	1	1	1	1	1	−1	−1	−1	−1	−1	1	1	1	1	1	−1	−1	1	−1	−1	A_{1u}
Cycle structure of S_8 class		1^8	$1^2,3^2$	$1^4,2^2$	$1^4,4$	$1^2,2^3$	$1^2,2^3$	$1,6$	$1^6,2$	$1^2,2,4$	$1^4,2^2$	2^4	$2,6$	4^2	$2^2,4$	8	$2^2,4$	$2,3^2$	2^4	4^2	2^4	
Order of $O_h(4)$ class		1	32	6	12	24	4	32	4	24	12	1	32	12	12	48	24	32	12	48	12	

From this the IR matrix components $\mathfrak{D}$ follow immediately. For example, $|{}^{E_g}_{1} \equiv (|x^2\rangle - |y^2\rangle/\sqrt{2}$, and $|{}^{E_g}_{2} - (2|z^2\rangle - |x^2\rangle - |y^2\rangle)/\sqrt{6}$ are bases of IR E_g according to Table I, and of unit norm according to (A1). Then for group element (yzx) we have the following:

$$\mathfrak{D}^{E_s}_{11}(yzx) = \left\langle {}^{E_g}_{1} \right| (yzx) \left| {}^{E_g}_{1} \right\rangle = \left\langle {}^{E_g}_{1} \right| (yzx) \frac{|x^2\rangle - |y^2\rangle}{\sqrt{2}}$$

$$= \left\langle {}^{E_g}_{1} \right| \frac{|y^2\rangle - |z^2\rangle}{\sqrt{2}} = -\tfrac{1}{2}$$

$$\mathfrak{D}^{E_g}_{12}(yzx) = \left\langle {}^{E_g}_{1} \right| (yzx) \left| {}^{E_g}_{2} \right\rangle = \left\langle {}^{E_g}_{1} \right| \frac{2|x^2\rangle - |y^2\rangle - |z^2\rangle}{\sqrt{6}}$$

$$= \sqrt{3}/2$$

$$\mathfrak{D}^{E_g}_{21}(yzx) = -\sqrt{3}/2$$

$$\mathfrak{D}^{E_g}_{22}(yzx) = -\tfrac{1}{2}.$$

APPENDIX B: CONSTRUCTING IR CURRENTS

States $|\Psi_1^{(\alpha)}\rangle, |\Psi_2^{(\alpha)}\rangle, \ldots, |\Psi_l^{(\alpha)}(\alpha)\rangle$ that obey Eq. (3), and thereby comprise a normalized basis of IR (α), are found by applying projection operators $P_{1m}^{(\alpha)}, P_{2m}^{(\alpha)}, \ldots, P_{l^{(\alpha)}(\alpha)_m}$ defined by (B1):

$$P_{lm}^{(\alpha)} = \left(\frac{l^{(\alpha)}}{\text{number of group operators}} \right) \sum_{\substack{\text{group}\\ \text{operators}\\ g}} \mathfrak{D}_{lm}^{(\alpha)}(g)^* g \tag{B1}$$

to state vector like $|j_1\rangle$ as in (B2). (It will be assumed that vectors like $g|j_1\rangle$ span the entire space in question, which in this first case is the 12-branch octahedral network. If not, other state vectors, like $|I_1\rangle$ and $|i_1\rangle$ in the case of Fig. 3, are picked, and the process to be described here is repeated for each.)

$$P_{lm}^{(\alpha)} |j_1\rangle = N_m^{(\alpha)} |\Psi_l^{(\alpha)}\rangle. \tag{B2}$$

In (B2) the scalar $N_m^{(\alpha)}$ is either a normalization constant or zero, and is determined quickly by (B3):

$$N_m^{(\alpha)} = \langle j_1 | P_{mm}^{(\alpha)} | j_1 \rangle. \tag{B3}$$

For those in which $N_m^{(\alpha)} \neq 0$, exactly $l^{(\alpha)}$ orthonormal states $|\Psi_l^{(\alpha)}\rangle$ $(l = 1, 2, \ldots, l^{(\alpha)})$ are constructed according to (B2). Those m for which $N_m^{(\alpha)} = 0$ give nothing.

For example, the operators (B4) with $(\alpha) = T_{1g}$

$$\begin{aligned}
P_{12}^{T_{1g}} = \tfrac{1}{16}\{&(zxy) + (\bar{z}x\bar{y}) - (\bar{z}\bar{x}y) - (z\bar{x}\bar{y}) - (y\bar{x}z) + (\bar{y}xz)\\
&+ (yx\bar{z}) - (\bar{x}\bar{y}\bar{z}) + (\bar{z}\bar{x}\bar{y}) + (z\bar{x}y) - (zx\bar{y}) - (\bar{z}xy)\\
&- (\bar{y}x\bar{z}) + (y\bar{x}\bar{z}) + (\bar{y}\bar{x}z) - (xyz)\},\\
P_{22}^{T_{1g}} = \tfrac{1}{16}\{&(xyz) - (x\bar{y}\bar{z}) + (\bar{x}y\bar{z}) - (\bar{x}\bar{y}z) + (\bar{z}yx) + (zy\bar{x})\\
&- (z\bar{y}x) - {}'(\bar{z}\bar{y}\bar{x}) + (\bar{x}\bar{y}\bar{z}) - (\dot{x}yz) + (x\bar{y}z) - (xy\bar{z})\\
&- (z\bar{y}\bar{x}) + (\bar{z}\bar{y}x) - (\bar{z}y\bar{x}) - (zyx)\}\\
P_{32}^{T_{1g}} = \tfrac{1}{16}\{&(yzx) - (y\bar{z}\bar{x}) + (\bar{y}z\bar{x}) - (\bar{y}\bar{z}x) + (xzy) - (x\bar{z}y)\\
&- (\bar{x}\bar{z}\bar{y}) + (\bar{x}zy) + (\bar{y}\bar{z}\bar{x}) - (\bar{y}zx) + (y\bar{z}x) - (yz\bar{x})\\
&+ (\bar{x}\bar{z}\bar{y}) - (\bar{x}z\bar{y}) - (xzy) + (x\bar{z}\bar{y})\}
\end{aligned} \tag{B4}$$

will give three states when applied to $|j_1\rangle$ since $N_2^{T_{1g}}$ is nonzero,

$$\begin{aligned}
&\langle j_1 | P_{22} | j_1 \rangle\\
&\quad = \langle j_1 | \tfrac{1}{16} \{(xyz) \cdots - (z\bar{y}x) \cdots + (x\bar{y}z) \cdots - (zyx)\} | j_1 \rangle\\
&\quad = \tfrac{1}{4} = N_2^{T_{1g}}.
\end{aligned}$$

The resulting orthonormal states are given in (B5), and

$$\begin{aligned}
|\Psi_1^{T_{1g}}\rangle &= \frac{1}{\sqrt{N_2^{T_{1g}}}} P_{12}^{T_{1g}} |j_1\rangle = \frac{1}{2}(|j_2\rangle - |j_5\rangle + |j_8\rangle - |j_{11}\rangle),\\
|\Psi_2^{T_{1g}}\rangle &= \frac{}{\sqrt{N_2^{T_{1g}}}} P_{22}^{T_{1g}} |j_1\rangle = \frac{1}{2}(|j_1\rangle - |j_4\rangle + |j_7\rangle - |j_{10}\rangle),\\
|\Psi_3^{T_{1g}}\rangle &= \frac{1}{\sqrt{N_2^{T_{1g}}}} P_{32}^{T_{1g}} |j_1\rangle = \frac{1}{2}(-|j_3\rangle + |j_6\rangle - |j_9\rangle + |j_{12}\rangle)
\end{aligned} \tag{B5}$$

drawn in Fig. 1.

The number $f^{(\alpha)}$ of independent multiplets $\{|\Psi_1^{(\alpha)}\rangle \ldots |\Psi_l^{(\alpha)}(\alpha)\rangle\} \cdots \{|\Psi_1^{(\alpha)} \cdots |\Psi_l^{(\alpha)}(\alpha)\rangle\}^{f^{(\alpha)}}$ is given by the standard frequency formula[7]:

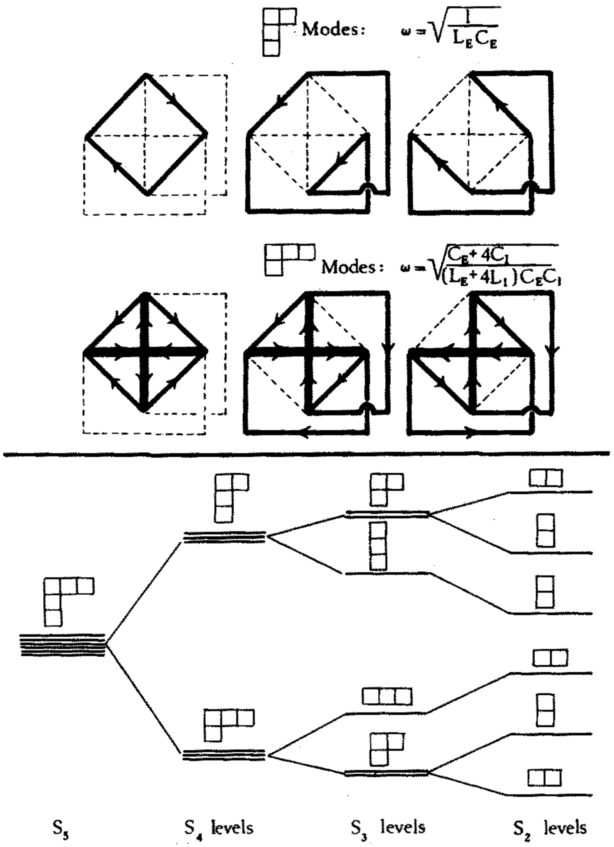

Modes: $\omega = \sqrt{\frac{1}{L_E C_E}}$
Modes: $\omega = \sqrt{\frac{C_E + 4C_I}{(L_E + 4L_I)C_E C_I}}$
S_5
S_4 levels
S_3 levels
S_2 levels

FIG. 6. Mode and frequency level diagram for various physical symmetries of bridge. Values of impedances are $S_5 \ldots Z_E = Z_E' = Z_I = Z_I'$, $S_4 \ldots Z_E' = Z_E \neq Z_I = Z_I'$, $S_3 \ldots Z_E = Z_E' \neq Z_I \neq Z_I'$, and $S_2 \ldots Z_E \neq Z_E' \neq Z_I \neq Z_I'$.

TABLE III. Characteristics of four-dimensional "regular solids."

	Vertices	Lines	Surfaces	3-Volumes	4-Volumes
"Tetrahedron"	5	10	10	5	1
"Octahedron"	8	24	32	16	1
"600 Cell"	120	720	1200	600	1
"Cube"	16	32	24	8	1
"24 Cell"	24	96	96	24	1
"Dodecahedral complex"	600	1200	0	120	1

TABLE IV. n-Dimensional "Solids."

"Tetrahedron"

n	Vertices	Lines	Surfaces	3-Volumes	4-Volumes	5-Volumes
0	1					
1	2	1				
2	3	3	1			
3	4	6	4	1		
4	5	10	10	5	1	
5	6	15	20	15	6	1
.	.	.	.	.	.	.
.	.	x	y	.	.	.
.	.	.	$x+y$	.	.	.

"Cube"

n	Vertices	Lines	Surfaces	3-Volumes	4-Volumes	5-Volumes
0	1					
1	2	1				
2	4	4	1			
3	8	12	6	1		
4	16	32	24	8	1	
5	32	80	80	40	10	1
.	.	.	.	.	.	.
.	.	x	y	.	.	.
.	.	.	$x+2y$	.	.	.

"Octahedron"

n	Vertices	Lines	Surfaces	3-Volumes	4-Volumes	5-Volumes
0	1					
1	2	1				
2	4	4	1			
3	6	12	8	1		
4	8	24	32	16	1	
5	10	40	80	80	32	1
.	.	.	.	.	.	.
.	.	x	y	.	.	.
.	.	.	$2x+y$	.	.	.

$$f^{(\alpha)} = \frac{1}{\text{number of group operators}} \sum_{\substack{\text{group}\\ \text{operators}\\ g}} \chi^{(\alpha)*}(g)\,\mathrm{Tr}\,J(g), \tag{B6}$$

where

$$\mathrm{Tr}\,J(g) = \sum_{n=1}^{12} \langle j_n | (g) | j_n \rangle.$$

In the above example $f^{T_{1g}} = 1$.

APPENDIX C: GENERAL POINT SYMMETRY

There is a correspondence between a regular (Platonic) polytope and a highest point symmetry in a give Euclidian n space. The five three-dimensional regular solids are the tetrahedron having point symmetry T_d the cube and octahedron each having point symmetry O_h, and finally the icosahedron and dodecahedron having Y_h symmetry. No three-dimensional point symmetries exist outside of these except the symmetries $R(2)$ and $R(3)$ of the cylinder and sphere, respectively, and their subgroups.

Similarly the six four-dimensional regular "solids" described in Table III correspond to high four-dimensional point symmetry.

The fourth "solid" is topologically represented in Fig. 3, and has an order 384 point symmetry corresponding to all combinations of $(\pm x_1, \pm x_2, \pm x_3, \pm x_4)$.

Beyond this there are only three n-dimensional solids for any given $n \geqslant 5$. These are recorded in the easily remembered triangle tables given in Table IV.

[1] L. C. Chen and J. L. Birman, J. Math. Phys. 12, 2454 (1971).

[2] D. E. Littlewood, *Theory of Group Characters* (Oxford, London, 1958), p.278.

[3] E. Wigner, *Group Theory and Applications* (Academic, New York, 1959).

[4] M. Hamermesh, *Group Theory and Applications to Physical Problems* (Addison-Wesley, Reading, Mass., 1964).

[5] Steven Bose, *Network Theory* (Harper & Row, New York, 1965).

[6] Reference 3, p. 75.

[7] Reference 4, p. 104.

Choquet simplexes of states in physical systems

Etang Chen*

Department of Mathematics and Institute for Fundamental Studies, The University of Rochester, Rochester, New York 14627

(Received 18 June 1974)

Two extreme types of Choquet simplexes, prime and Bauer, are shown in the states of some physical systems.

Given a C^*-algebra $\mathfrak{A}$ with identity, let $\mathcal{S}$ be the set of all states on $\mathfrak{A}$. The state space of $\mathfrak{A}$ is $\mathcal{S}$ endowed with the w^*-topology, i.e., the $\sigma(\mathfrak{A}^*, \mathfrak{A})$ topology, where $\mathfrak{A}^*$ is the dual Banach space of $\mathfrak{A}$. The state space $\mathcal{S}$ is compact. The facial structure of $\mathcal{S}$ has been extensively studied. It is well known that $\mathcal{S}$ can be a Choquet simplex (in fact a Bauer simplex) if and only if $\mathfrak{A}$ is Abelian. However, in mathematical physics, the states interested in physical systems are not the whole state space $\mathcal{S}$ but only some compact convex subset K of $\mathcal{S}$, e.g., the equilibrium states for a given temperature, the translational invariant states, and the translational invariant equilibrium states, etc. Hence, it would be interesting in studying the structure of Choquet simplex K, which may correspond to the states of some physical systems.

In a previous paper,[1] we have shown that under a certain condition the Choquet simplex K_β of KMS states, with respect to a nontrivial automorphism of $\mathfrak{A}$, for a given inverse temperature $\beta > 0$ is not a Bauer simplex. For the Choquet simplex $\mathcal{S}_G$ of the invariant states under a group G, it can be a Bauer simplex only in a very special case.[2]

We shall study the simplicial structure of K in the present paper, and show that under a certain physical assumption K can be a *prime* simplex, which is "complementary" to a Bauer simplex in some sense. On the other hand, K can be a Bauer simplex in some trivial cases. We shall also construct a nontrivial Bauer simplex of states with "lower" symmetry. Therefore, two extreme types of Choquet simplexes, prime and Bauer, can appear in the states of some physical systems.

Before formulating our main results, we give some elementary definitions and recall some known results from the theory of Choquet simplexes. We refer to Ref. 3 for more detailed information.

A *Choquet simplex* K in the state space $\mathcal{S}$ is a compact convex subset of $\mathcal{S}$ such that the associated cone $\cup_{\lambda\geqslant 0}\lambda K$ is a lattice in its own order. A convex subset F of K is called a *face* if for $x, y \in K$, $\lambda x + (1-\lambda)y \in F$ entails that $x, y \in F$. Denote by $\mathcal{E}(K)$ the extreme boundary of K, i.e., the set of all extreme points of K. A *Bauer simplex* K is a Choquet simplex with closed extreme boundary. K is said to be *prime*,[4] if $K = \text{co}(F_1 \cup F_2)$ for any two closed faces F_1 and F_2 of K, then either $K = F_1$ or $K = F_2$, where co(·) means the convex hull of a set (·).

A partially ordered vector space X is called an *antilattice* if for any pair x and y in X, the lattice infinimum $x \wedge y$ exists in X implies that either $x \wedge y = x$ or $x \wedge y = y$. By the equality $x \vee y = -(-x \wedge -y)$, if $x \vee y$ exists in an antilattice then necessarily $x \vee y = x$ or $x \vee y = y$. Hence, an antilattice X is a partially ordered vector space, where only the trivial lattice infinima and suprema exist.

We denote by $A(K)$ the Banach space of all real continuous functions on K. Then, K is a prime simplex if and only if $A(K)$ is an antilattice.[4] On the other hand, K is a Bauer simplex if and only if $A(K)$ is a lattice.[3] This gives a justification for the statement that a prime simplex is "complementary" to a Bauer simplex.

First, we shall show that K can be a prime simplex under a certain conditions, which can be implied by some physical assumptions. Then, we apply our result to the equilibrium states, i.e., KMS states, and translational-invariant equilibrium states, which are defined as follows.

Let $t \to \sigma_t$ be a representation of the additive group of real numbers into the group of $*$-automorphisms of $\mathfrak{A}$. A state $\varphi \in \mathcal{S}$ is KMS for a given $\beta > 0$, if it satisfies the KMS boundary condition for $\beta > 0$: For each $x, y \in \mathfrak{A}$, there exists a holomorphic function F in $0 < \text{Im} z < \beta$, which is continuous in $0 \leqslant \text{Im} z \leqslant \beta$ with boundary values

$$F(t) = \varphi(\sigma_t(x)y) \quad \text{and} \quad F(t + i\beta) = \varphi(y\sigma_t(x)).$$

Let K_β be the set of all KMS states with respect to σ_t for a given $\beta > 0$. If K_β is compact in the state space, then K_β is a Choquet simplex. For the details on K.M.S. states, we refer to Ref. 5 and the references given there. In the present paper, we shall always consider that K_β is a Choquet simplex.

In addition to the time evolution defined above, we are also interested in the spatial translation. In fact, we consider a more general case: Let G be a group, and $g \to \alpha_g$ a representation of G into the group of $*$-automorphisms of $\mathfrak{A}$. Denote by $\mathcal{S}_G$ the set of all G-invariant states, i.e., $\varphi \circ \alpha_g = \varphi$ for all $g \in G$. $\mathcal{S}_G$ is then a compact convex set in the state space. We shall also study the simplicial structure of $K_\beta \cap \mathcal{S}_G$, which is a Choquet simplex again if nonempty.[5] $K_\beta \cap \mathcal{S}_G$ corresponds to the translational invariant equilibrium states for $G = R^\nu$ or $\mathbb{Z}^\nu$.

Let H be a face of $\mathcal{S}$, and $H \cap K$ nonempty; then $H \cap K$ is a face of K. If F is a face of K such that $F = H \cap K$ for a face H of $\mathcal{S}$, then F is *induced* by H. We note that if K itself is a face of $\mathcal{S}$ (this is called a *facial simplex* in Ref. 1), then each face of K is induced by a face of $\mathcal{S}$.

For $\varphi \in K$, let F_φ (resp. F_φ^K) be the smallest closed face of $\mathcal{S}$ (resp. K) containing φ. Then, they have the following relation:

Lemma 1: If F_φ^K is induced by a closed face of $\mathcal{S}$, then F_φ^K is induced by F_φ.

Proof: Let $F_\varphi^K = K \cap H$ for some closed face H of $\mathcal{S}$. As $\varphi \in K$, $K \cap F_\varphi$ is nonempty; it is a closed face of K

containing φ. Hence, $F_\varphi^K \subseteq K \cap F_\varphi$. On the other hand, φ is also in H; then $F_\varphi \subseteq H$, so that $F_\varphi^K = K \cap H \supseteq K \cap F_\varphi$. Therefore $F_\varphi^K = K \cap F_\varphi$.

We note that if F_φ^K is induced by F_φ and $K \cap F_\varphi = \{\varphi\} = F_\varphi^K$, then $\varphi \in \mathcal{E}(K)$; in particular, if $F_\varphi = \{\varphi\}$, i.e., $\varphi \in \mathcal{E}(\mathcal{S})$, the set of all pure states. However, if $F_\varphi = \{\varphi\}$, then, in general, $F_\varphi \cap K = \phi$, for physical systems, e.g., $\varphi \in K_\beta$. In fact, we are interested only in the cases, where $K \cap \mathcal{E}(\mathcal{S}) = \phi$.

For a subset N in the state space $\mathcal{S}$, the w^*-closure of N is denoted by N^-. Let Π_φ be the cyclic representation of $\mathfrak{A}$ induced by $\varphi \in \mathcal{S}$, and $\mathcal{H}_\varphi$ its representation space with cyclic vector ξ_φ. Define the state $\tilde{\varphi}$ of $\Pi_\varphi(\mathfrak{A})^-$ by $\tilde{\varphi}(x) = w_{\xi_\varphi}(x)$ for $x \in \Pi_\varphi(\mathfrak{A})^-$. We note that $\tilde{\varphi}$ is faithful if and only if ξ_φ is separating for $\Pi_\varphi(\mathfrak{A})^-$.

Lemma 2: In the above notations; if $\tilde{\varphi}$ is faithful, then either $F_\varphi = \{\varphi\}$ or $F_\varphi = \mathcal{S}$.

Proof: If φ is a pure state, then $F_\varphi = \{\varphi\}$. Suppose φ is not a pure state. Let H_φ be the norm closure of the following set:

$$\{\tau \in \mathcal{S};\ \tau \leqslant \alpha\varphi \text{ for some } \alpha > 0\}.$$

Then $F_\varphi \supseteq H_\varphi$, hence $F_\varphi \supseteq H_\varphi^-$. However, from, e.g., Ref. 6,

$$H_\varphi = \{\tau \in \mathcal{S};\ \tau = \omega_\xi \circ \Pi_\varphi \text{ for } \xi \in [\Pi_\varphi(\mathfrak{A})'\xi_\varphi]\}.$$

As $\tilde{\varphi}$ is faithful, ξ_φ is a separating vector for $\Pi_\varphi(\mathfrak{A})^-$, so that $[\Pi_\varphi(\mathfrak{A})'\xi_\varphi] = \mathcal{H}_\varphi$. Hence $H_\varphi^- = \mathcal{S}$. Consequently, $F_\varphi = \mathcal{S}$.

For $\varphi \in K_\beta$, φ can never be pure, and $\tilde{\varphi}$ is faithful, since ξ_φ is separating for $\Pi_\varphi(\mathfrak{A})^-$, see, e.g., Ref. 5. Therefore, we have the following:

Corollary 3: For $\varphi \in K_\beta$, $F_\varphi = \mathcal{S}$.

We are now able to show our main result:

Proposition 4: Let K be a Choquet simplex in the state space $\mathcal{S}$ such that $K \cap \mathcal{E}(\mathcal{S}) = \phi$. Then K is a prime simplex, if there is a state φ in $\mathcal{E}(K)^- \backslash \mathcal{E}(K)$ such that $\tilde{\varphi}$ is faithful and F_φ^K is induced by a closed face of $\mathcal{S}$.

Proof: We follow a similar method given in Ref. 4. Suppose

$$K = \mathrm{co}(F_1 \cup F_2),$$

where F_1 and F_2 are two closed faces of K. By the given assumptions, φ lies in $\mathcal{E}(K)^- = \mathcal{E}(F_1)^- \cup \mathcal{E}(F_2)^- \subseteq F_1 \cup F_2$. Hence, $F_\varphi^K \subseteq F_i$ ($i = 1, 2$). From Lemmas 1, 2 and the hypothesis of the theorem, $F_\varphi^K = K$; therefore $F_i = K$ ($i = 1, 2$).

An immediate consequence is

Corollary 5: K_β is a prime simplex if there is a KMS state φ in $\mathcal{E}(K_\beta)^- \backslash \mathcal{E}(K_\beta)$ such that $F_\varphi^{K_\beta}$ is induced by a closed face of $\mathcal{S}$.

We note that the key assumption in the above corollary is $F_\varphi^{K_\beta} = K_\beta \cap F_\varphi$, which has other alternative forms. More precisely, the following statements are equivalent for $\varphi \in K_\beta$:

(i) $F_\varphi^{K_\beta} = K_\beta \cap F_\varphi$,

(ii) $F_\varphi^{K_\beta} = K_\beta$,

(iii) $F_\varphi^{K_\beta} \supseteq \mathcal{E}(K_\beta)$.

Indeed, (i)⇔(ii) follows from Corollary 3; (ii)⇒(iii) is trivial, and (iii)⇒(ii) from Krein–Milman theorem.

We note furthermore that (iii) can be obtained from some physical assumptions as follows: If we consider $\varphi \in K_\beta$ as an equilibrium state at a given inverse temperature β for a certain quantum system, then $\tau \in \mathcal{E}(K_\beta)$ can be interpreted as a pure phase of the given system. φ can be obtained from the convex combination of some pure phases τ_i; i.e., φ is a mixture of some pure phases. If $\varphi = \lambda\tau_1 + (1-\lambda)\tau_2$ for $0 < \lambda < 1$, the τ_1 and τ_2 are components of φ; i.e., φ is a mixture of two pure phases τ_1, τ_2 as components. Similarly, φ can be a mixture of n pure phases $\tau_1, \ldots, \tau_n \in \mathcal{E}(K_\beta)$. An equilibrium state φ in K_β is called a *complete mixture* if it is a mixture of *all* pure phases in $\mathcal{E}(K_\beta)$. Mathematically speaking, $\varphi \in K_\beta$ is a complete mixture if $\varphi = \sum_i \lambda_i \tau_i$ for $\tau_i \in \mathcal{E}(K_\beta)$, where $\sum \lambda_i = 1$ and $\lambda_i \neq 0$ for all i.

As φ is a complete mixture, then $\varphi \geqslant \lambda_i \tau_i$ for all $\tau_i \in \mathcal{E}(K_\beta)$, hence $(1/\lambda_i)\varphi \geqslant \tau_i$ (since $\lambda_i \neq 0$). Thus, $\tau_i \in F_\varphi^{K_\beta}$ for all $\tau_i \in \mathcal{E}(K_\beta)$, i.e., $F_\varphi^{K_\beta} \supseteq \mathcal{E}(K_\beta)$, which is (iii).

In fact, the above argument of a complete mixture holds not only for an equilibrium state in K_β, but also for any state in K; viz., if there is a complete mixture $\varphi \in K$, then $F_\varphi^K \supseteq \mathcal{E}(K)$. Therefore, using the same arguments in the proof of Proposition 4, we have the following.

Corollary 6: A Choquet simplex K in the state space $\mathcal{S}$ is a prime simplex if there is a complete mixture φ in $\mathcal{E}(K)^- \backslash \mathcal{E}(K)$.

We give another example of prime simplex which appears in the classical lattice systems.[7]

The configuration space of a classical lattice system is given by $T = \{0, 1\}^{\mathbf{Z}^\nu}$. A configuration of the system is described by a subset $X \subset \mathrm{Z}^\nu$ of occupied lattice sites. The C^*-algebra $\mathfrak{A}$ in this system is $\mathcal{C}(T)$, the C^*-algebra of all continuous complex function on T. Every element of $\mathfrak{A}$ can be considered a function of subsets of Z^ν. $\mathfrak{A}$ can be endowed with the quasilocal structure: Indeed, for each finite subset Λ of Z^ν, there is a corresponding subalgebra $\mathfrak{A}(\Lambda)$ of $\mathfrak{A}$. The elements of $\mathfrak{A}(\Lambda)$ are defined by $a(X) = f(X \cap \Lambda)$ for some $f \in \mathcal{C}(\Lambda)$, where X is a configuration in T, and $\mathcal{C}(\Lambda)$ is the space of all continuous complex functions on Λ. Then the union of $\mathfrak{A}(\Lambda)$, by the Stone–Weierstrauss theorem, is dense in $\mathfrak{A}$.

A state φ on $\mathfrak{A}$ can be constructed from a density destribution μ_Λ on Λ by

$$\varphi(a) = \sum_{X \subset \Lambda} \mu_\Lambda(X) f(X).$$

For the details, we refer to Ref. 7, p. 189. Let $\mathcal{S}_{\mathbf{Z}^\nu}$ be the set of all Z^ν-invariant states on $\mathfrak{A}$. Then, $\mathcal{S}_{\mathbf{Z}^\nu}$ is a Choquet simplex, since $\mathfrak{A}$ is Abelian, hence also $\mathbf{Z}^\nu$-Abelian.[7] We note that a state $\varphi \in \mathcal{E}(\mathcal{S}_{\mathbf{Z}^\nu})$ can be constructed by

$$\varphi(a) = V(x_0)^{-1} \sum_{x \in \Lambda(x_0)} \tilde{\varphi}(\alpha_x(a)),$$

for $x_0 \in \mathbf{Z}^\nu$ and $a \in \mathfrak{A}$, where $\Lambda(x_0) = \{x \in \mathbf{Z}^\nu;\ 0 \leqslant x_i < x_{0i}$ for $i = 1, \ldots, \nu\}$, $V(x_0)$ is the volume of $\Lambda(x_0)$. And, $\tilde{\varphi}$ is given by a density distribution

$$\mu_\Lambda(X)=\prod_{i=1}^{n}\hat{\mu}_{\Lambda_{n_i}}(X\cap\Lambda_{n_i})$$

with $\Lambda_{n_i}=\Lambda(x_0)+n_i x_0$ such that $\hat{\mu}_K$ is the corresponding density distribution for a given $\hat{\varphi}\in\mathcal{S}_{\mathbf{Z}^\nu}$. Under these conditions, φ tends to $\hat{\varphi}$ as $x_0\to\infty$ (Ref. 7, p. 197). Hence, $\mathcal{E}(\mathcal{S}_{\mathbf{Z}^\nu})$ is dense in $\mathcal{S}_{\mathbf{Z}^\nu}$. Therefore, by a theorem in Ref. 8, we have the following.

Observation 7: The Choquet simplex $\mathcal{S}_{\mathbf{Z}^\nu}$ in the classical lattice systems is a prime simplex.

We now turn to the cases where K can be a Bauer simplex.

As we have shown in Ref. 1, if K is a facial simplex, then it can be a Bauer simplex under a certain assumptions. However, a facial simplex, so far, seems to be no direct interest in physical systems. Therefore, K will be considered only in the cases of K_β or $K_\beta\cap\mathcal{S}_G$.

K_β can be a Bauer simplex in the following cases:

(i) K_β is a finite set. In particular, K_β is a singleton, e.g., in one-dimensional quantum lattice systems with finite-range interaction.[9]

(ii) Another example is the CCR algebra $\overline{\Delta(H,\sigma)}$ considered in Ref. 10. For a degenerate σ, it has a class of central states $\mathcal{C}$ invariant under a group of automorphisms of $\overline{\Delta(H,\sigma)}$. $\mathcal{C}$ forms a Bauer simplex. Consequently, K_β with respect to the *trivial* automorphism for $\beta=0$ is also a Bauer simplex. Moreover, if σ is nondegenerate, then it turns out that $\overline{\Delta(H,\sigma)}$ is simple and K_β degenerates to a singleton as in the case (i). This was pointed out to the author by Dr. M. Winnink.

From Corollary 5 and the above remarks, we know that if $F_\varphi^{K_\beta}$ is induced by a closed face of the state space, then K_β is either prime or Bauer. In fact, under this assumption, the following are equivalent:

(i) K_β is prime,

(ii) $A(K_\beta)$ is an antilattice,

(iii) $\mathcal{E}(K_\beta)$ is not closed.

On the other hand, the following are also equivalent:

(i) K_β is Bauer,

(ii) $A(K_\beta)$ is a lattice,

(iii) $\mathcal{E}(K_\beta)$ is closed.

Let us consider next $K=K_\beta\cap\mathcal{S}_G$, which is a Choquet simplex if K is nonempty.[5] As usual, for $\varphi\in\mathcal{S}_G$, let $\{\Pi_\varphi\mathcal{H}_\varphi,\ \xi_\varphi\}$ be the cyclic representation of $\mathfrak{A}$ induced by φ, and U_φ the unitary representation of G on $\mathcal{H}_\varphi$. We denote by $\mathcal{Z}_\varphi$ the center of $\Pi_\varphi(\mathfrak{A})^-$, by $\mathcal{U}_\varphi(G)$ the group of unitary operators $U_\varphi(g)$ for $g\in G$, and by $\mathcal{B}_\varphi$ the von Neumann algebra $\Pi_\varphi(\mathfrak{A})'\cap\mathcal{U}_\varphi(G)'$.

Proposition 8[5]: K is a face of $\mathcal{S}_G$ (resp. K_β) if and only if $\mathcal{B}_\varphi\subseteq\mathcal{Z}_\varphi$ (resp. $\mathcal{B}_\varphi\supseteq\mathcal{Z}_\varphi$).

The proof for K to be a face of K_β follows exactly the same way as K is a face of $\mathcal{S}_G$.[5]

Consequently, $\mathcal{E}(K)\subseteq\mathcal{E}(\mathcal{S}_G)$ and $\mathcal{E}(K)\subseteq\mathcal{E}(K_\beta)$, respectively.

Furthermore, we note that if $\mathfrak{A}$ is asymptotically Abelian with respect to G in the sense of Ref. 2, then K is a face of K_β if and only if $\mathcal{B}_\varphi=\mathcal{Z}_\varphi$.

Certainly, Corollary 6 holds for $K_\beta\cap\mathcal{S}_G$. Hence $K_\beta\cap\mathcal{S}_G$ can be a prime simplex if there is a complete mixture in $\mathcal{E}(K_\beta\cap\mathcal{S}_G)^{-}\setminus\mathcal{E}(K_\beta\cap\mathcal{S}_G)$. However, $K_\beta\cap\mathcal{S}_G$ can also be a Bauer simplex in the following cases:

(i) $K_\beta\cap\mathcal{S}_G$ is a finite subset. In particular it degenerates to a singleton, e.g., either K_β or $\mathcal{S}_G$ is a single point. Moreover,

(ii) For $\mathcal{B}_\varphi\supseteq\mathcal{Z}_\varphi$ (resp. $\mathcal{B}_\varphi\subseteq\mathcal{Z}_\varphi$), if K_β (resp. $\mathcal{S}_G$) is a Bauer simplex, then $K_\beta\cap\mathcal{S}_G$ is also Bauer. This can be seen as follows: The *facial topology* on $\mathcal{E}(K_\beta)$ is defined as a topology whose closed sets are exactly those $\mathcal{E}(F)$ for a closed face F of K_β. This facial topology is T_1 and compact, but, in general, not Hausdorff. It is Hausdorff iff K_β is Bauer.[3] As $\mathcal{B}_\varphi\supseteq\mathcal{S}_\varphi$, by Proposition 8, $\mathcal{E}(K_\beta\cap\mathcal{S}_G)\subseteq\mathcal{E}(K_\beta)$. Hence $\mathcal{E}(K_\beta\cap\mathcal{S}_G)$ is a topological subspace of $\mathcal{E}(K_\beta)$, which is endowed with the facial topology. In fact, any closed face of $K_\beta\cap\mathcal{S}_G$ is a closed face of K_β, since $K_\beta\cap\mathcal{S}_G$ is a face of K_β from Proposition 8. Therefore, $\mathcal{E}(K_\beta\cap\mathcal{S}_G)$ endowed with the facial topology is Hausdorff, as $\mathcal{E}(K_\beta)$ is Hausdorff by the assumption. Similar arguments hold exactly for the case, where $\mathcal{S}_G$ is Bauer and $\mathcal{B}_\varphi\subseteq\mathcal{Z}_\varphi$.

We construct another nontrivial Bauer simplex of physical states with "lower" symmetry.

Let $\mathfrak{A}$ be separable, G locally compact and separable. As before, G is acting on $\mathfrak{A}$ as a representation in the $*$-automorphisms of $\mathfrak{A}$. For $\varphi\in\mathcal{S}$, let $\mathcal{O}_\varphi$ be the orbit of φ under G, i.e., $\mathcal{O}_\varphi=\{\alpha_g^t(\varphi);\ g\in G\}$, where α_g^t is the transpose of α_g, and $G_\varphi=\{g\in G;\ \alpha_g^t(\varphi)=\varphi\}$, the stabilizer of φ. $\mathcal{S}_{G_\varphi}$ denotes the set of all G_φ-invariant states on $\mathfrak{A}$. We consider the w^*-closed convex hull K_φ of $\mathcal{O}_\varphi$.

We assume, furthermore, that G_φ is a normal subgroup of G, and $\mathfrak{A}$ is asymptotically Abelian with respect to G_φ in the sense of Ref. 2. Suppose that $\mathcal{S}_{G_\varphi}$ is nonempty, then it is a Choquet simplex. Let $\rho\in\mathcal{S}_{G_\varphi}$ μ_ρ the associated central measure, and μ_ρ be concentrated on $\mathcal{O}_\varphi$, i.e., $\mu_\rho(\mathcal{O}_\varphi)=\mu_\rho(\mathcal{S})$.

As $\mathfrak{A}$ is asymptotically Abelian with respect to G_φ, then $\mathcal{B}_\rho=\Pi_\rho(\mathfrak{A})'\cap\mathcal{U}_\rho(G_\varphi)'$ is contained in $\mathcal{Z}_\rho$. Hence from Ref. 11, $\mathcal{E}(K_\varphi)=\mathcal{O}_\varphi\subseteq\mathcal{E}(\mathcal{S}_{G_\varphi})$, which implies that K_φ is a face of $\mathcal{S}_{G_\varphi}$, since $\mathcal{S}_{G_\varphi}$ is a Choquet simplex. In fact, K_φ is also a Choquet simplex (which is called a facial simplex of $\mathcal{S}_{G_\varphi}$ in Ref. 1). Moreover, $\mathcal{E}(K_\varphi)=\mathcal{O}_\varphi$ is w^*-closed.[11] Therefore, we have shown the following:

Proposition 9: K_φ is a Bauer simplex.

ACKNOWLEDGMENTS

The author is grateful to Dr. M. Winnink for helpful remarks.

*Supported in part by NSF Grant DID71-04010-A02.

[1] E. Chen, J. Math. Phys. 13, 1130 (1972).

[2] E. Stǿrmer, Commun. Math. Phys. 5, 1 (1967).

[3] E. M. Alfson, *Compact Convex Sets and Boundary Integrals* (Springer-Verlag, Berlin, 1971).

[4] E. G. Effros and J. Kazdan, J. Differential Eqs. 8, 95 (1970).

[5] M. Takesaki and M. Winnink, Commun. Math. Phys. 30, 129 (1973).

[6] R. V. Kadison, Trans. Amer. Math. Soc. 103, 304 (1962).

[7] D. Ruelle, *Statistical Mechanics* (Benjamin, New York, 1969).

[8] Chu Cho-Ho, Math. Scand. 31, 151 (1972).

[9] H. Araki, Commun. Math. Phys. 14, 120 (1969).

[10] J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, Commun. Math. Phys. 32, 231 (1973).

[11] R. H. Herman, Trans. Amer. Math. Soc. 158, 503 (1971).