U(N) Integrals, 1/N, and the De Wit–'t Hooft anomalies

Stuart Samuel
Institute for Advanced Study, Princeton, New Jersey 08540

(Received 2 April 1980; accepted for publication 23 July 1980)

Formulas for the evaluation of all U(N) integrals are derived. Tables display the results for integrands involving up to six U's and six U†'s. The complete pole structure of De Wit–'t Hooft anomalies is unveiled. The effects of $1/N^2$ corrections and De Wit–'t Hooft anomalies on two-dimensional U(N) lattice gauge theories in the strong coupling $1/N$ expansion is discussed.

I. INTRODUCTION

This paper derives a set of formulas which immediately allow the evaluation of U(N) group integrals. These formulas have intrinsic mathematical value. But not only are they interesting from a mathematical standpoint, they are valuable from a physics standpoint. A theory of strong interactions has been proposed: quantum chromodynamics (QCD). It is in qualitative agreement with the general features of strong interactions. There appear to be no other theories having these qualities. QCD is unique in having asymptotic freedom and the possibilities of quark confinement. However, in QCD it is extremely difficult to calculate except in selected high energy processes. A proposed calculational method is to put the theory on a lattice. Immediately, the problem of doing group integrations arises. Therefore, this paper's integral formulas are valuable to anyone attempting to compute QCD on a lattice. Furthermore, this study of the U(N) integral reveals many interesting phenomena some of which may lead to new computational methods.

In addition to U(N) integrals the $1/N$ expansion and the De Wit–'t Hooft anomalies[1] are studied. Much of Sec. II is spent defining the notation. A general formula for U(N) integrals is derived. The answer [Eq. (2.10)] is expressed in terms of the characters of the permutation groups (all of which are known in closed form[2]). Hence all U(N) integrals are known. This is an important result because the integrals appear in strong coupling lattice gauge theories.

Section III derives a set of recursion relations for the U(N) integrals.

Section IV discusses the De Wit–'t Hooft anomalies. The U(N) integrals behave nonanalytically in N. For integrals involving n U's and n U†'s analytic expressions exist for $N \geqslant n$. When extrapolated to $N < n$, poles appear and invalidate the formulas. These poles are known as the De Wit–'t Hooft anomalies. Section IV gives a complete description of the pole structure. *Amazingly, not only do simple poles appear in N but for n large enough poles of arbitrary high order appear.*

Section V explains a procedure for extrapolating $N \geqslant n$ results to $N < n$ so that no anomalies occur. In other words, a correct method of handling the anomalies is found. Thus, Sec. II results can be applied to the $N < n$ case.

Section VI contains a set of tables displaying the integrals up to $n = 6$ for all N. These tables are for theorists performing lattice strong coupling expansions.

Section VII contains simple algebraic formulas for the U(2) coefficients.

Section VIII discusses the generating function for U(N) integrals.

Section IX focuses on the two-dimensional lattice gauge theory. This is a solvable model in which large N and De Wit–'t Hooft anomalies can be analyzed exactly. It is shown that *large N strong coupling expansions are bad because of the anomalies. For $1/g^2N$ small, large N is a reasonable approximation to finite N but $1/N^2$ corrections cannot and do not improve on this. For $1/g^2N$ sufficiently large, $1/N$ strong coupling expansions give erroneous results.*

The analysis of this paper suggests two trends of thought. There seems to be a connection between large N and the permutation groups. These groups naturally arise when doing U(N) integrals and they may play an important role in higher dimensions. The interplay of the permutation groups with large N deserves more consideration and might uncover a deeper relation.

Secondly, the De Wit–'t Hooft poles are extremely important for finite N. They hamper the extrapolation of large N to finite N. From this point of view they are an annoyance, a barrier to be overcome. I believe the situation should be looked at differently. Instead of being considered destructive, they should be considered as an interesting theoretical phenomenon to be taken advantage of. One should ask how can they be put to good use to obtain finite N results; how can new approximation schemes be found (can some sort of pole dominance of integrals be made?) These ideas beckon more attention.

II. THE U(N) INTEGRAL

The problem is to compute

$$I_n^N = \int d\,\mathrm{U}\, \mathrm{U}_{i_1}^{i_1} \mathrm{U}_{i_2}^{i_2} \cdots \mathrm{U}_{i_n}^{i_n} \mathrm{U}_{m_1}^{*j_1} \mathrm{U}_{m_2}^{*j_2} \cdots \mathrm{U}_{m_n}^{*j_n}, \tag{2.1}$$

for U(N). The group measure in Eq. (2.1) is the right and left invariant normalized ($\int d\,\mathrm{U} \equiv 1$) Haar measure.[2] Throughout this paper $\mathrm{U}_{il} \equiv \mathrm{U}_l^i$, $(\mathrm{U}^\dagger)_{mj} \equiv \mathrm{U}_m^{*j}$, capital N is the N of U(N), and lower case n refers to the number of U's and U†'s in Eq. (2.1). Of course, integrals are zero unless the number of U's is equal to the number of U†'s.

Because the U(N) measure is invariant under multiplication from the right [$d(V\mathrm{U}) = d\,\mathrm{U}$], each i_s index must contract with a j_t index (likewise for l and m indices). Hence Eq. (2.1) must be of the form

 0022-2488/80/122695-09$1.00

$$I_n^N = \sum_{\sigma_A}\sum_{\sigma_B} C_{\sigma_A,\sigma_B}(N)\delta^{i_1}_{j_{\sigma_A(1)}}\delta^{i_2}_{j_{\sigma_A(2)}}\cdots\delta^{i_n}_{j_{\sigma_A(n)}}$$
$$\times\delta^{l_1}_{m_{\sigma_B(1)}}\delta^{l_2}_{m_{\sigma_B(2)}}\cdots\delta^{l_n}_{m_{\sigma_B(n)}}, \tag{2.2}$$

where Σ_{σ_A} means to sum over all permutations, σ_A, of the integers 1, 2, ..., n. $C_{\sigma_A,\sigma_B}(N)$ are coefficients which depend on σ_A and σ_B. Knowledge of these coefficients is equivalent to knowing all the U(N) integrals. The main result of this section is a formula [Eqs. (2.10), (2.11), and (2.12)] for $C_{\sigma_A,\sigma_B}(N)$ in terms of the characters of S_n, the permutation group on n objects.

Another way of expressing U(N) integrals is to multiply Eq. (2.1) by $(A_1)_{l_1i_1}(A_2)_{l_2i_2}\cdots(A_n)_{l_ni_n}(B_1)_{j_1m_1}(B_2)_{j_2m_2}\cdots(B_n)_{j_nm_n}$ and sum over all indices:

$$I_n^N\{(A,B)\} = \int dU\,\mathrm{Tr}A_1U\,\mathrm{Tr}A_2\mathrm{U}\cdots\mathrm{Tr}A_nU$$
$$\times\mathrm{Tr}B_1U^\dagger\mathrm{Tr}B_2U^\dagger\cdots\mathrm{Tr}B_nU^\dagger, \tag{2.3}$$

where Tr stands for trace. Equation (2.3) must be of the form

$$I_n^N\{(A,B)\} = \sum_{\substack{\text{partitions}\\ l_1,l_2,\ldots,l_m\\ \text{of } n}}\sum_{\sigma_A}\sum_{\sigma_B}\frac{g_n(l_1,l_2,\ldots,l_m)}{n!}C^{(N)}_{l_1,l_2,\ldots,l_m}$$
$$\times\left[\mathrm{Tr}(A_{\sigma_A(1)}B_{\sigma_B(1)}A_{\sigma_A(2)}B_{\sigma_B(2)}\cdots A_{\sigma_A(l_1)}B_{\sigma_B(l_1)})\right]$$
$$\times\left[\mathrm{Tr}(A_{\sigma_A(l_1+1)}B_{\sigma_B(l_1+1)}\cdots A_{\sigma_A(l_1+l_2)}B_{\sigma_B(l_1+l_2)})\right]$$
$$\times\cdots\times\left[\mathrm{Tr}(A_{\sigma_A(l_1+l_2+\cdots+l_{m-1}+1)}B_{\sigma_B(l_1+l_2+\cdots+l_{m-1}+1)}\right.$$
$$\left.\cdots A_{\sigma_A(n)}B_{\sigma_B(n)})\right]. \tag{2.4}$$

In Eq. (2.4) and throughout this paper a partition of n is a set of integers $l_1, l_2, \ldots, l_m$ such that $l_1 + l_2 + \cdots + l_m = n$ and

$$\sum_{\substack{\text{partitions}\\ l_1,l_2,\ldots,l_m \text{ of } n}}$$

means to sum over $l_1, l_2, \ldots, l_m$ (and m) with the constraints that $n\geqslant l_1\geqslant l_2\geqslant\cdots\geqslant l_m\geqslant 1$ and $l_1 + l_2 + \cdots + l_m = n$. For a partition, $l_1, l_2, \ldots, l_m$ the right-hand side of Eq. (2.4) has the following structure: there is a trace of a product of l_1 (AB)'s, times the trace of a product of l_2 (AB)'s etc. Summing over the permutations, σ_A and σ_B, generates all possible terms with the same trace structure. $g_n(l_1, l_2, \ldots, l_m)/n!$ [given in Eq. (2.5) below] insures that each distinct term on the right-hand side of Eq. (2.4) occurs precisely once (summing over all σ_A and σ_B leads to duplication). Sometimes it is convenient to adopt an alternative expression of a partition.[2] Let α_1 be the number of 1's in $(l_1, l_2, \ldots, l_m)$; let α_2 be the number of 2's in $(l_1, l_2, \ldots, l_m)$; etc. Use the standard[2] abbreviation

$$l^p = (\underbrace{l, l, \ldots, l}_{p\ l\text{'s}}).$$

Then $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ means $(l_1, l_2, \ldots, l_m) = (1^{\alpha_1}, 2^{\alpha_2}, \ldots, n^{\alpha_n})$ [some of the α_l's will be zero indicating the absence of l from $(l_1, l_2, \ldots, l_m)$]. Furthermore, $l_1 + l_2 + \ldots l_m = n$ implies $\alpha_1 + 2\alpha_2 + 3\alpha_3 + \ldots + n\alpha_n = n$. The notation is the same as in Ref. 2. In terms of the α's

$$g_n(\alpha) = n!/1^{\alpha_1}(\alpha_1!)2^{\alpha_2}(\alpha_2!)\ldots n^{\alpha_n}(\alpha_n!). \tag{2.5}$$

In Eq. (2.5) and throughout this paper adopt the notations $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\mathbf{l} = (l_1, l_2, \ldots, l_m)$ and the conventions $g_n(l_1, l_2, \ldots, l_m) = g_n(\mathbf{l}) = g_n(\alpha) = (\alpha_1, \alpha_2, \ldots\alpha_n)$ and $C_{l_1,l_2,\ldots l_m}(N) = C_1(N) = C_\alpha(N) = C_{1^{\alpha_1},2^{\alpha_2},\ldots,n^{\alpha_n}}(N)$ when the partition associated with l notation corresponds to the one associated with α notation.

Equation (2.4) may look complicated with its vast array of indices but it is actually quite simple. For example,

$$\int dU\,\mathrm{Tr}A_1U\mathrm{Tr}B_1U^\dagger = C_1(N)\mathrm{Tr}A_1B_1,$$
$$\int dU\,\mathrm{Tr}A_1U\mathrm{Tr}A_2U\mathrm{Tr}B_1U^\dagger\mathrm{Tr}B_2U^\dagger$$
$$= C_{1^2}(N)(\mathrm{Tr}A_1B_1\mathrm{Tr}A_2B_2 + \mathrm{Tr}A_1B_2\mathrm{Tr}A_2B_1)$$
$$+ C_2(N)(\mathrm{Tr}A_1B_1A_2B_2 + \mathrm{Tr}A_1B_2A_2B_1). \tag{2.6}$$

Contact can be made between the matrix index formulation [Eqs. (2.1) and (2.2)] and the trace formulation [Eqs. (2.3) and (2.4)]. The following is true: the coefficients, $C_{\sigma_A,\sigma_B}(N)$ in Eq. (2.2) depend only on the conjugacy class[3] of $\sigma_A\circ\sigma_B^{-1}$. Recall[3] that a permutation can be uniquely specified by exhibiting its cycles and that two permutations are in the same conjugacy class if they have the same cycle structure (i.e. they leave the same number of objects invariant (1-cycles), they have the same number of 2-cycles, 3-cycles, etc.). There is a one-to-one correspondence between cycle structures (and hence conjugacy classes) and partitions. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the number of 1-cycles, 2-cycles, ..., n-cycles in $\sigma_A\circ\sigma_B^{-1}$. Then

$$C_{\sigma_A,\sigma_B}(N) = C_{1^{\alpha_1},2^{\alpha_2},\ldots,n^{\alpha_n}}(N). \tag{2.7}$$

Equation (2.7) relates the coefficients in Eq. (2.2) with those in Eq. (2.4) and bridges the two formulations.

To derive a set of equations for the C's, choose two permutations σ'_A and σ'_B, set $i_1 = j_{\sigma'_A(1)}, i_2 = j_{\sigma'_A(2)}, \ldots, i_n = j_{\sigma'_A(n)}, l_1 = m_{\sigma'_B(1)}, l_2 = m_{\sigma'_B(2)}, \ldots, l_n = m_{\sigma'_B(n)}$ in Eq. (2.2) and sum over all indices. What results is

$$\sum_{\sigma_A,\sigma_B} C_{\sigma_A,\sigma_B}N^{\{f(\sigma'_A,\sigma_A)+f(\sigma'_B,\sigma_B)\}} = N^{\{f(\sigma'_A,\sigma'_B)\}}, \tag{2.8}$$

where

$$f(\sigma,\sigma') = \#\text{ of cycles in }\sigma^{-1}\circ\sigma', \tag{2.9}$$

(that is, if $\sigma^{-1}\circ\sigma'$ has a cycle structure corresponding to $(\alpha_1, \alpha_2, \ldots, \alpha_n)$, then $f(\sigma,\sigma') = \alpha_1 + \alpha_2 + \ldots + \alpha_n$).

For $n\leqslant N$ Eqs. (2.8) are sufficient to determine the C_{σ_A,σ_B} uniquely:

$$C_{\sigma_A,\sigma_B}(N) = \sum_r \chi_r(\sigma_A\circ\sigma_B^{-1})\chi_r(e)/n!f_r(N). \tag{2.10}$$

Here Σ_r is a sum over all irreducible representations of the permutation group S_n, $\chi_r(\sigma)$ is the character of σ in the rth representation, $\chi_r(e)$ [see Eq. (2.11)] is the character of the identity element, and $f_r(N)$, a polynomial in N of order n vanishing at certain integers, is specified below in Eq. (2.12). The proof of Eq. (2.10) is presented in Appendix A. Recall that there is a one-to-one correspondence between conjugacy classes and irreducible representations so that a representation, r, of S_n can be characterized by an ordered partition $(\lambda_1, \lambda_2, \ldots, \lambda_m)$ (with $\lambda_1\geqslant\lambda_2\geqslant\ldots\geqslant\lambda_m$) of n in the standard manner.[2] A formula[2] for $\chi_r(e)$ is

$$\chi_{(\lambda_1,\lambda_2,\ldots,\lambda_m)}(e) = n!\prod_{i<j}(\lambda_i - \lambda_j + j - i)/\prod_i(\lambda_i + m - i)!. \tag{2.11}$$

Finally $f_r(N)$ is

$$f_{(\lambda_1,\lambda_2,\ldots,\lambda_m)}(N) = \prod_{i=1}^{m} (N+\lambda_i - i)!/(N-i)!$$
$$= [N(N+1)\cdots(N+\lambda_1-1)]$$
$$\times [(N-1)(N)(N+1)\cdots(N+\lambda_2-2)]$$
$$\times [(N-m)(N-m+1)\cdots(N+\lambda_m-m)]. \quad (2.12)$$

Because all the characters of S_n are known (in terms of the Frobenius generating function)[2] Eqs. (2.10) and (2.11) represent a complete solution. They allow the quick calculation of any $C_{l_1,l_2,\ldots,l_m}(N)$. For example, if $\sigma_A \circ \sigma_B^{-1} = e$ then one need only to substitute Eqs. (2.11) and (2.12) to obtain $C_{1^n}(N)$. As another example, consider $C_n(N)$. Let σ be an n cycle. From the Frobenius generating function one deduces that $\chi_{1^q,p}(\sigma) = (-1)^q$ and for other representations $\chi(\sigma) = 0$. Equation (2.11) gives

$$\chi_{1^q,p}(e) = (n-1)!/q!(p-1)!$$

(where $p+q=n$). Thus

$$C_n(N) = \frac{1}{n!}\sum_{q=0}^{n-1}(-1)^q \frac{(n-1)!}{q!(n-q-1)!}\frac{1}{N}$$
$$\times\left[\frac{1}{N+1}\cdot\frac{1}{N+2}\cdots\frac{1}{N+n-q-1}\right]$$
$$\times\left[\frac{1}{N-1}\frac{1}{N-2}\cdots\frac{1}{N-q}\right], \quad (2.13)$$

from which one concludes

$$C_n(N) = \frac{(-1)^{n-1}}{N(N^2-1)(N^2-4)\cdots(N^2-(n-1)^2)}\frac{(n-1)!(n-1)!}{n}$$
$$\times\sum_{q=0}^{n-1}\frac{1}{q!}\frac{1}{q!}\frac{1}{(n-q-1)!}\frac{1}{(n-q-1)!}. \quad (2.14)$$

All coefficients, $C_1(N)$, can be calculated in the above manner.

III. RECURSION RELATIONS

This section presents a complete set of recursion equations relating the C_1^n's to the C_1^{n-1}'s. For clarity a superscript, n, has been appended to the C_1's (i.e. C_1^n are the coefficients for I_n^N). The recursion relations provide an alternative method of computation. This method was used, for example, in Ref. 4. For small n there is little difference in computational complexity between the two methods; for large n, however, the use of character tables is much more efficient.

Recursion relations can be generated in the following manner: Take the integral in Eq. (2.3) and replace A_n and B_n by $\lambda^\gamma A_n$ and $B_n\lambda^\gamma$ and sum over γ (here the λ^γ are the N^2 generators of the Lie algebra of U(N); they satisfy $\Sigma_\gamma(\lambda^\gamma)_{ij}(\lambda^\gamma)_{lm} = \delta_{im}\delta_{jl}$). The left-hand side of Eq. (2.4) becomes

$$I_{n-1}^N(\{A,B\})\mathrm{Tr}A_nB_n, \quad (3.1)$$

while a variety of terms depending on the trace structure are generated on the right-hand side. Here is a summary of what can happen under $A_n \to \lambda^\gamma A_n$ and $B_n \to B_n\lambda^\gamma$:

(a) *Invariant Processes*: Terms such as $\mathrm{Tr}A_nXB_n$ become $N\mathrm{Tr}A_nXB_n$ and are simply multiplied by a factor of N. Here X stands for any matrix product of A's and B's and by fiat includes the case, X = the identity matrix, for which $\mathrm{Tr}A_nB_n \to N\mathrm{Tr}A_nB_n$ also.

(b) *Fission Processes*: Terms such as $\mathrm{Tr}XA_nYB_n$ break into two trace terms (fission), i.e., $\mathrm{Tr}YA_nXB_n \to \mathrm{Tr}Y\,\mathrm{Tr}A_nXB_n$. Here X and Y are arbitrary matrix products of A's and B's and X may also be just the identity matrix.

(c) *Fusion Processes*: Terms such as $\mathrm{Tr}A_nX\,\mathrm{Tr}YB_n$, in which A_n and B_n appear in different traces, fuse into a single trace: $\mathrm{Tr}A_nX\,\mathrm{Tr}YB_n \to \mathrm{Tr}YB_nA_nX$.

Collecting all terms of the form $(\mathrm{Tr}A_nB_n)\times$(something) and comparing to Eq. (3.1) results in the following recursion relations:

$$C^{n-1}_{l_1,l_2,\ldots,l_m} = N\,C^n_{l_1,l_2,\ldots,l_m,1}$$
$$+\sum_{s=1}^{m}(l_s)C^n_{l_1,\ldots,l_{s-1},l_s+1,l_{s+1},\ldots,l_m} \quad (3.2)$$

where $(l_1, l_2, \ldots, l_m)$ is a partition of $n-1$. No particular ordering is assumed for the l_s's and for convenience $C^n_{l_1,\ldots,l_m}$ is chosen to be symmetric in the l_s indices. The first term of the right-hand side of Eq. (3.2) comes from invariant processes whereas the second term comes from fission processes. Fusion processes do not contribute because they cannot lead to a $(\mathrm{Tr}A_nB_n)\times$(something) structure.

All terms not of the form $(\mathrm{Tr}A_nB_n)\times$(something) must sum to zero and lead to the following consistency conditions:

$$0 = N\,C^n_{l_1,l_2,\ldots,l_m} + \sum_{l=1}^{l_1-1} C^n_{l_1-l,l_2,\ldots,l_m,l}$$
$$+\sum_{s=1}^{m}(l_s)\,C^n_{l_1+l_s,l_2,\ldots,\hat{l}_s,\ldots,l_m}, \quad (3.3)$$

where $(l_1, l_2, \ldots, l_m)$ is any unordered partition of n for which $l_1 \geqslant 2$. The "hat" over l_s indicates the absence of that index symbol. The three terms in Eq. (3.3) are generated respectively by invariant, fusion, and fission processes.

Equations (3.3) and (3.2) are the main result of this section. They are a complete set which uniquely determine the C^n's in terms of the C^{n-1}'s.

IV. THE DE WIT–'t HOOFT ANOMALIES

In a letter,[1] De Wit and 't Hooft found poles in a certain subset of diagrams at integer values of N when attempting to do U(N) integrals in lattice gauge theory calculations. In particular, they found in low orders poles at $N=1$ and $N=2$. This phenomenon made it impossible to write the contribution of a high temperature (strong coupling) graph for arbitrary N. Separate formulas were needed for $N=1$ and $N=2$. This anomalous behavior in N, they argued, presents a serious barrier to performing strong coupling $1/N$ expansions and prevents such expansions from approximating finite values of N. This nonanalytic behavior casts doubt whether strong coupling large N expansions are relevant. This section studies the nature of the De Wit–'t Hooft anomaly. A complete description of the anomaly will be given. It will be shown that the situation is much worse: *not only do simple poles occur at all integers, but poles of arbitrarily high order.*

The pole structure is easily analyzed using Eqs. (2.10) and (2.12). In fact poles are due to the $f_r(N)$ which vanish at

integer values and appear in the denominator of Eq. (2.10). Define $D_n(N)$ to be the common denominator of the $C_1^n(N)$'s. Using the results of Sec. II:

$$
\begin{aligned}
&D_1(N) = N,\\
&D_2(N) = N(N^2-1),\\
&D_3(N) = N(N^2-1)(N^2-4),\\
&D_4(N) = N^2(N^2-1)(N^2-4)(N^2-9),\\
&D_5(N) = N^2(N^2-1)(N^2-4)(N^2-9)(N^2-16),\\
&D_6(N) = N^2(N^2-1)^2(N^2-4)(N^2-9)(N^2-16)(N^2-25).
\end{aligned}
\tag{4.1}
$$

A double pole at $N=1$ first occurs when $N=6$, i.e. in integrations involving six U's and six U†'s. In general

$$
D_n(N) = N^{m_0}\prod_{s=1}^{n-1}(N^2-s^2)^{m_s}, \tag{4.2}
$$

where

$$
m_s = \text{The biggest integer such that } m_s(m_s+s)\leqslant n. \tag{4.3}
$$

Equations (4.2) and (4.3) imply that the $C_1^n(N)$ coefficients will eventually have poles at all integers to arbitrary high powers. As an example, for U(3) no poles occur in $C_1^n(3)$ for $n=1$, 2, and 3; simple poles occur for $n=4$–9; double poles occur for $n=10$–17, triple poles for $n=18$–27, etc. In general a pole of order l will first occur at N when $l(l+N)=n$.

If De Wit and 't Hooft are correct about the nonextrapolation of large N to finite N in strong coupling $1/N$ expansions then the results of this section imply the situation is infinitely worse.

V. FINITE N

For $N<n$ the coefficients $C_1^n(N)$ are infinite. This is due to the poles in N at the integers $-(n-1)$ to $(n-1)$. These singularities are the De Wit–'t Hooft anomalies discussed in the last section. Of course, the integral in Eq. (2.1) is well defined and always finite. The source of difficulty is the lack of independence of the index structures in Eq. (2.2) [see Eq. (5.2) below]. Similarly, in the trace formalism not all the trace structures in Eq. (2.4) are independent [see Eq. (5.3) below]. Because an overly determined set of tensor structures is being used it is natural that singularities occur. Hence the formalism of Sec. II appears invalidated for $N<n$. However, Eqs. (2.2) and (2.4), as well as the solution [Eq. (2.10)] still work in the following sense:

For $N\times N$ matrices $M_1, M_2, \ldots, M_n$ and $N<n$ a symmetrized trace of n matrices, $[\Sigma_\sigma \mathrm{Tr}(M_{\sigma(1)}M_{\sigma(2)}\ldots M_{\sigma(n)})]$ can be written as sums of products of traces of less than n matrices. Examples are:

$$
\begin{aligned}
&\mathrm{Tr}M_1M_2 = \mathrm{Tr}M_1\mathrm{Tr}M_2, \quad \text{for } N=1,\\
&\mathrm{Tr}M_1M_2M_3 + \mathrm{Tr}M_1M_3M_2 = \mathrm{Tr}M_1M_2\mathrm{Tr}M_3\\
&\qquad + \mathrm{Tr}M_2M_3\mathrm{Tr}M_1 + \mathrm{Tr}M_3M_1\mathrm{Tr}M_2\\
&\qquad - \mathrm{Tr}M_1\mathrm{Tr}M_2\mathrm{Tr}M_3, \quad \text{for } N=2 \text{ or } N=1.
\end{aligned}
\tag{5.1}
$$

When this happens in Eq. (2.4) terms involving traces of more than N matrices can be regrouped into terms involving traces less than or equal to N matrices. The C's then combine and all poles cancel. Hence by expressing dependent tensor structures in terms of an independent set, a nonsingular formalism with finite coefficients results.

Call the process of decomposing a trace into products of smaller traces a decay process. The goal is to obtain all decay processes. Consider the completely antisymmetric delta function on n indices:

$$
\delta^{i_1 i_2\cdots i_n}_{j_1 j_2\cdots j_n} \equiv \sum_\sigma (\mathrm{sign}\sigma)\delta^{i_1}_{j_{\sigma(1)}}\delta^{i_2}_{j_{\sigma(2)}}\cdots\delta^{i_n}_{j_{\sigma(n)}}, \tag{5.2}
$$

for which the indices, i_s and j_t, take the values 1, 2, ..., N. If $N<n$, antisymmetry in the j_t's implies $\delta^{i_1 i_2\cdots i_n}_{j_1 j_2\cdots j_n}=0$. Multiply Eq. (5.2) by $(M_1)_{i_1 j_1}(M_2)_{i_2 j_2}\cdots(M_n)_{i_n j_n}$ and sum over all indices. The following trace identities are generated and represent a complete set of decay processes:

$$
\begin{aligned}
&\sum_{\substack{\text{partitions}\\ p_1,p_2,\ldots,p_m \text{ of } n}} \mathrm{sign}(p_1,p_2,\ldots,p_m)\frac{g_n(p_1,p_2,\ldots,p_m)}{n!}\\
&\quad\times\sum_\sigma(\mathrm{Tr}M_{\sigma(1)}M_{\sigma(2)}\ldots M_{\sigma(p_1)})(\mathrm{Tr}M_{\sigma(p_1+1)}\cdots M_{\sigma(p_1+p_2)})\\
&\quad\times\cdots\times(\mathrm{Tr}M_{\sigma(p_1+p_2+\cdots p_{m-1}+1)}\cdots M_{\sigma(n)}) = 0,
\end{aligned}
\tag{5.3}
$$

for $N<n$, where $g_n(p_1,p_2,\ldots,p_m)/n!$ [see Eq. (2.5)] serves the same purpose as in Eq. (2.4), namely to insure that each distinct term on the right-hand side of Eq. (5.3) appears once. In Eq. (5.3)

$$
\mathrm{sign}(p_1,p_2,\ldots,p_m) \equiv \prod_{i=1}^m(-1)^{p_i+1}. \tag{5.4}
$$

Equation (5.3) has the following structure: a trace of a product of p_1 M's times a trace of a product of p_2 M's etc. Equations (5.1) are examples of Eq. (5.3).

Regrouping traces according to the decay processes of Eq. (5.3) modifies the C_1^n's as follows:

$$
\begin{aligned}
&C^n_{p_1,p_2,\ldots,p_m,l_2,l_3,\ldots,l_q}(N) \to C^n_{p_1,p_2,\ldots,p_m,l_2,l_3,\ldots,l_q}(N)\\
&\quad + \frac{g_n(l_1,l_2,\ldots,l_q)g_{l_1}(p_1,p_2,\ldots,p_m)}{g_n(p_1,p_2,\ldots,p_m,l_2,l_3,\ldots,l_q)(l_1-1)!}\\
&\quad\times(-1)^{l_1}\mathrm{sign}(p_1,p_2,\ldots,p_m)C^n_{l_1,l_2,\ldots,l_q}(N),
\end{aligned}
\tag{5.5}
$$

where $l_1 = p_1+p_2+\cdots+p_m$, l_1 must be greater than N, the g's are defined in Eq. (2.5), and $(l_1,l_2,\ldots,l_q)$ form a partition of n whereas $(p_1,p_2,\ldots,p_m)$ form a partition of l_1. Equation (5.5) can be thought of as the process in which $(l_1,l_2,\ldots,l_q)$ decays into $(p_1,p_2,\ldots,p_m,l_2,l_3,\ldots,l_q)$.

When $N<n$, keep doing Eq. (5.4) until all the l_i in $C_1^n(N)$ are less than or equal to N. Then this generates a set of $C_1^n(N)$'s (which shall be denoted $C_1^{U(N)}$) without poles in N and Eq. (2.4) is valid if one sums only over those partitions $(l_1,l_2,\ldots,l_m)$ of n such that all l_s are less than or equal to N.

For example, if $n=N+1$ then $C^{U(N)}_{N+1}\to 0$ and

$$
\begin{aligned}
C^{U(N)}_{p_1,p_2,\ldots,p_m} &= C_{p_1,p_2,\ldots,p_m}(N)\\
&\quad + (-1)^N\left(\prod_{i=1}^m(-1)^{p_i+1}\right)C_{N+1}(N).
\end{aligned}
\tag{5.6}
$$

The decay processes in Eq. (5.5) can be combined with

TABLE I. Table of d_n^N.

n \ N	2	3	4	5
3	24			
4	240	2160		
5	1440	15120	161280	
6	30240	967680	21772800	435456000

the invariant, fusion, and fission processes of Sec. III to yield recursion relations for $C_1^{U(N)}$. For example, for U(2)

$$C^{U(2)}_{2^l,1^m} = 2(l+1)C^{U(2)}_{2^{l+1},1^{m-1}} + m(l-m+2)C^{U(2)}_{2^{l+1},1^{m-1}} - \tfrac{1}{2}m(m-1)(m-2)\,C^{U(2)}_{2^{l+2},1^{m-3}}, \tag{5.7}$$

for $l \geq 0$ and $m \geq 0$.

$$0 = C^{U(2)}_{2^{l-1},1^{m+2}} + (1+2m+l)C^{U(2)}_{2^l,1^m} + \frac{m(m-1)}{2}\,C^{U(2)}_{2^l,1^{m-1}}, \tag{5.8}$$

for $l \geq 1$, $m \geq 0$, and the last term is absent if $m = 0$. Section VII presents the solution to these equations yielding in closed form the $C^{U(2)}_{2^l,1^m}$'s.

VI. THE INTEGRAL TABLES

This section computes all U(N) integrals up to $n = 6$ (i.e., six U's and six U†'s) and displays the results in Tables II–VI. These tables will be particularly useful in strong coupling expansions of lattice U(N) gauge theories and other lattice U(N) field theories. Enough information is contained in these tables to do computations to at least twelfth order [i.e., $(1/g^2N)^{1,2}$].

For $N \geq n$, integrals were computed using Eq. (2.10). For $N < n$ the "decay" reduction process of Sec. V were carried out.

In reference to Tables II–VI, the $C_1^{U(N)}$'s have been written in fractional form, (numerator)/(denominator). The numerators are the entries in the tables. The denominators, denoted by d_n^N, are given in Table I for $N < n$ and are equal to $D_n(N)$ [Eq. (4.1)] for $N \geq n$. In general, for $N < n$, the d_n^N would be defined as

$$d_n^N = \left[\prod_{s=1}^{n-1}(N+s)^{m_s}\right] N^{m_0}\left[\prod_{s=1}^{N-1}(N-s)^{m_s}\right], \tag{6.1}$$

where the m_s are given in Eq. (4.3). Hence $1/d_n^N$ is $1/D_n(N)$ with the poles at $n-1, n-2, \ldots, N+1, N$ removed. d_n^N naturally arises in carrying out the decay processes. The

TABLE II. The $C_\alpha^{U(N)}$'s for $n = 2$.

	$N \geq 2$
$D_2(N)C_{1^2}^{U(N)}$	N
$D_2(N)C_2^{U(N)}$	-1

TABLE III. The $C_\alpha^{U(N)}$'s for $n = 3$.

	$N = 2$	$N > 3$
$d_3^N C_{1^3}^{U(N)}$	4	N^2-2
$d_3^N C_{1,2}^{U(N)}$	-1	$-N$
$d_3^N C_3^{U(N)}$	0	2

$C_1^{U(N)}$'s are written in fractional form to avoid the unpleasant appearance of ratio's of large numbers.

For $n = 1$,

$$C_1(N) = C_1^{U(N)} = 1/N. \tag{6.2}$$

For $n = 2$–6, the $d_n^N C_1^{U(N)}$'s are displayed in Tables II–VI.

VII. THE U(2) INTEGRALS

When $N = 1$ there is a single coefficient for each n: $C_{1^n}^{U(1)}$, and $C_{1^n}^{U(1)} = 1/n!$. The first nontrivial case is U(2). The U(2) integrals can be computed by writing the measure and integrand in terms of the four parameters needed to describe U(2): three SU(2) angles and one U(1) phase. Explicit integration then gives

$$C^{U(2)}_{1^m,2^l} = \frac{m!(-1)^l}{(n+1)!} \times \sum_{\substack{r=0,2,4,\ldots \text{if } n \text{ is even} \\ r=1,3,5,\ldots \text{if } n \text{ is odd}}}^{n-2l} \frac{1}{r!((n-r)/2)!((n-r)/2-l)!2^{((n-r)/2-l)}}, \tag{7.1}$$

where $n = m + 2l$. These coefficients represent the complete solution to Eqs. (5.7) and (5.8) and the U(2) integral.

Table VII summarized the U(2) coefficients up to $n = 12$. To avoid ratios of large fractions $(n+1)!\,C^{U(2)}_{1^m,2^l}$ is shown. Columns one, two, and three specify n, m, and l (of course $n = m + 2l$) and column four displays $(n+1)!\,C^{U(2)}_{1^m,2^l}$.

VIII. THE GENERATING FUNCTION

This section studies the generating function

$$I(AB) = \int dU \exp[\beta' \mathrm{Tr} A\,U + \beta' \mathrm{Tr} B\,U^\dagger], \tag{8.1}$$

where A and B are arbitrary matrices. $I(AB)$ is a function only of the invariants TrAB, Tr$ABAB$, ... Such an integral is interesting for several reasons:

TABLE IV. The $C_\alpha^{U(N)}$'s for $n = 4$.

	$N = 2$	$N = 3$	$N > 4$
$d_4^N C_{1^4}^{U(N)}$	17	55	N^4-8N^2+6
$d_4^N C_{1^2,2}^{U(N)}$	-3	-18	$-N^3+4N$
$d_4^N C_{1,3}^{U(N)}$	0	7	$2N^2-3$
$d_4^N C_{2^2}^{U(N)}$	1	1	N^2+6
$d_4^N C_4^{U(N)}$	0	0	$-5N$

TABLE V. The $C_\alpha^{U(N)}$'s for $n = 5$.

	$N=2$	$N=3$	$N=4$	$N\geqslant 5$
$d_5^N C_{1^5}^{U(N)}$	37	151	384	N^5-20N^3+78N
$d_5^N C_{1^3,2}^{U(N)}$	−5	−38	−130	$-N^4+14N^2-24$
$d_5^N C_{1^2,3}^{U(N)}$	0	10	64	$2N^3-18N$
$d_5^N C_{1,2^2}^{U(N)}$	1	5	32	N^3-2N
$d_5^N C_{1,4}^{U(N)}$	0	0	−26	$-5N^2+24$
$d_5^N C_{2,3}^{U(N)}$	0	−2	−2	$-2N^2-24$
$d_5^N C_5^{U(N)}$	0	0	0	$14N$

(a) It appears as an intermediate integration in lattice U(N) gauge theories and other lattice U(N) field theories. For example, the calculation of $I(AB)$ would be the first step of a real space renormalization program in which a set of link variables were integrated out. This integral also arises in other approximation methods such as that of Ref. 4.

(b) Often simple integrals [such as Eq. (8.1)] are studied to gain insight into higher dimensional field theories. For example, similar one variable integrals can be used to count the number of Feynman graphs.

(c) When A equals B equals I, the identity matrix, the integral in Eq. (8.1) becomes the vacuum functional for the two-dimensional lattice U(N) gauge theory[5,6] and is exactly solvable for all N.[6] In this model $\beta' = 1/g^2$, where g is the gauge field coupling constant. Thus $I(AB)$ contains as a subcase an interesting model.

(d) Knowledge of $I(AB)$ is commensurate to knowledge of all the integrals in Eq. (2.1): differentiating $I(AB)$ with respect to $A^{l_1}_{i_1}, A^{l_2}_{i_2}, \ldots, A^{l_n}_{i_n}, B^{j_1}_{m_1}, B^{j_2}_{m_2}, \ldots, B^{j_n}_{m_n}$ and setting $A = B = 0$ yields Eq. (2.1). This is why $I(AB)$ is called the generating function.

In general

$$I(AB) = \exp\left\{N^2 \sum_{n=1}^{\infty} \frac{(\beta)^{2n}}{n!} \times \sum_{\substack{\alpha_1,\alpha_2,\ldots,\alpha_n \\ \alpha_1+2\alpha_2+\ldots+n\alpha_n=n}} C_\alpha^c(N) N^{(2n-2)} \left(\frac{\mathrm{Tr}AB}{N}\right)^{\alpha_1} \times \left(\frac{\mathrm{Tr}ABAB}{N}\right)^{\alpha_2} \cdots \left(\frac{\mathrm{Tr}(AB)^n}{N}\right)^{\alpha_n}\right\}, \tag{8.2}$$

where $\beta = \beta'/N$ ($= 1/g^2N$ for gauge models) is the real expansion parameter and $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ (with some $\alpha_s = 0$) represents the partition of n of the form $(1^{\alpha_1}, 2^{\alpha_2}, \ldots, n^{\alpha_n})$. The N^2 in front of the sum indicates that for large N the "vacuum energy" is proportional to N^2, that is, the coefficients, $C_\alpha^c(N)$, satisfy

$$C_\alpha^c(N) N^{(2n-2)} \to \text{const} + O(1/N^2), \tag{8.3}$$

as $N \to \infty$. The superscript, c, on $C_\alpha^c(N)$ stands for connected part. The $C_\alpha^c(N)$ can be recursively related to the $C_\alpha^{U(N)}$, are the analogs of the contribution from connected Feynman graphs, and appear in the exponent because connected vacuum bubbles exponentiate.

For $n = 1$ and $N = 2$

$$C_1^c(N) = 1,$$
$$C_{1^2}^c(N) = 1/(N^2-1),$$
$$C_2^c(N) = -1/(N^2-1). \tag{8.4}$$

Tables VIII, IX, and X display the results for $n = 3, 4$, and 5. After the completion of this work, another manuscript by Bars appeared which also obtains the generating function to β^{10}.[7]

IX. DISCUSSION OF LARGE N AND DE WIT–'t HOOFT ANOMALIES IN TWO-DIMENSIONAL LATTICE U(N) GAUGE THEORIES

The two-dimensional U(N) lattice gauge theory is exactly solvable for all N finite[6] or infinite.[5,6] This provides a framework in which questions about large N and De Wit–'t Hooft anomalies can be answered. Consider large N first. Reference 6 has thoroughly analyzed the large N behavior, so only the impact on strong coupling expansions will be discussed. Let $I_N(\beta) = I(AB)$ [Eq. (8.1)] for $A = B =$ the identity matrix. Define

$$\Gamma_N(\beta) = (1/N^2)\ln I_N(\beta) = \sum_{n=1}^{\infty} \beta^{2n} c_n(N),$$
$$\Gamma_\infty(\beta) = \lim_{N\to\infty} \Gamma_N(\beta), \tag{9.1}$$

and

$$\Gamma_{p,N}(\beta) = \sum_{n=1}^{p} \beta^{2n} c_n(N). \tag{9.2}$$

TABLE VI. The $C_\alpha^{U(N)}$'s for $n = 6$.

	$N=2$	$N=3$	$N=4$	$N=5$	$N\geqslant 6$
$d_6^N C_{1^6}^{U(N)}$	246	3498	17890	69562	$N^8-41N^6+458N^4-1258N^2+240$
$d_6^N C_{1^4,2}^{U(N)}$	−27	−726	−5024	−21850	$-N^7+33N^5-254N^3+342N$
$d_6^N C_{1^3,3}^{U(N)}$	0	144	2044	11182	$2N^6-51N^4+229N^2-60$
$d_6^N C_{1^2,2^2}^{U(N)}$	4	118	986	6722	$N^6-19N^4+58N^2-160$
$d_6^N C_{1^2,4}^{U(N)}$	0	0	−586	−5750	$-5N^5+93N^3-208N$
$d_6^N C_{1,2,3}^{U(N)}$	0	−36	−149	−2650	$-2N^5+5N^3+117N$
$d_6^N C_{1,5}^{U(N)}$	0	0	0	2352	$14N^4-154N^2+140$
$d_6^N C_{2^3}^{U(N)}$	−1	−2	−288	−450	$-N^5-N^3-358N$
$d_6^N C_{2,4}^{U(N)}$	0	0	94	142	$5N^4+75N^2+40$
$d_6^N C_{3^2}^{U(N)}$	0	18	22	52	$4N^4+116N^2-360$
$d_6^N C_6^{U(N)}$	0	0	0	0	$-42N^3+42N$

TABLE VII. The $C_\alpha^{U(2)}$'s up to $n = 12$.

n	m	l	$(n+1)!C^{U(2)}_{l^m,2^l}$
1	1	0	1
2	2	0	2
	0	1	− 1
3	3	0	4
	1	1	− 1
4	4	0	8 1/2
	2	1	− 1 1/2
	0	2	1/2
5	5	0	18 1/2
	3	1	− 2 1/2
	1	2	1/2
6	6	0	41
	4	1	− 4 1/2
	2	2	2/3
	0	3	− 1/6
7	7	0	92
	5	1	− 8 1/2
	3	2	1
	1	3	− 1/6
8	8	0	208 3/8
	6	1	− 16 5/8
	4	2	1 5/8
	2	3	− 5/24
	0	4	1/24
n	m	l	$(n+1)!C^{U(2)}_{l^m,2^l}$
9	9	0	475 3/8
	7	1	− 33 5/8
	5	2	2 19/24
	3	3	− 7/24
	1	4	1/24
10	10	0	1090 3/4
	8	1	− 68 3/8
	6	2	5
	4	3	− 53/120
	2	4	1/20
	0	5	− 1/120
11	11	0	2514 1/2
	9	1	− 142 3/8
	7	2	9 1/4
	5	3	− 17/24
	3	4	1/15
	1	5	− 1/120
12	12	0	5819 5/16
	10	1	− 300 7/16
	8	2	17 9/16
	6	3	− 1 3/16
	4	4	23/240
	2	5	− 7/720
	0	6	1/720

Equation (9.1) defines $c_n(N)$. $\Gamma_N(\beta)$ is the vacuum energy density per degree of freedom for the two-dimensional U(N) lattice gauge theory. $\Gamma_{p,N}(\beta)$ is the strong coupling approximation to $\Gamma_N(\beta)$ to pth order. Take the large N limit of Eq. (9.2); write $c_n(N)$ as

$$c_n(N) = \sum_{m=0}^{\infty} c_{n,m}(1/N^2)^m. \tag{9.3}$$

TABLE VIII. The $C_\alpha^c(N)$'s for $n = 3$.

	$N=2$	$N\geqslant 3$
$C_{1^3}^c(N)$	$-\frac{2}{3}$	$\frac{8}{(N^2-1)(N^2-4)}$
$C_{1,2}^c(N)$	$\frac{1}{2}$	$\frac{-12}{(N^2-1)(N^2-4)}$
$C_3^c(N)$	0	$\frac{4}{(N^2-1)(N^2-4)}$

As long as $p < N$ no anomalies occur, $c_n(N)$ is a ratio of polynomials in N and the expansion in Eq. (9.3) can be done. Large N replaces the coefficients $c_n(N)$ in $\Gamma_{p,N}$ by $c_{n,0}$ of Eq. (9.3):

$$\Gamma_{p,\infty}(\beta) = \sum_{n=1}^{p} \beta^{2n} c_{n,0}. \tag{9.4}$$

Equation (9.4) is the strong coupling large N approximation to pth order. For sufficiently large β ($\beta > \frac{1}{2}$) the series in Eq. (9.4) does not converge to $\Gamma_\infty(\beta)$ in Eq. (9.2) as $p \to \infty$. In other words

$$\lim_{p\to\infty}\left(\lim_{N\to\infty}\Gamma_{p,N}(\beta)\right) \neq \lim_{N\to\infty}\left(\lim_{p\to\infty}\Gamma_{p,N}(\beta)\right) \equiv \Gamma_\infty(\beta) \tag{9.5}$$

for $\beta > \frac{1}{2}$. This is deduced from the large N results of Refs. 5 and 6 and the β expansion of Γ [see Eq. (9.10)]. What does Eq. (9.5) say about strong coupling large N calculations? $\Gamma_{p,N}$ is what one computes in the strong coupling lattice expansion to order p. In the large N limit, $\Gamma_{p,\infty}$ is obtained and is a bad approximation (for sufficiently large β) to the exact large N limit [the right-hand side of Eq. (9.5)]. One is ultimately interested in weak coupling (in g and hence large β) so that a continuum limit can be taken. The large N strong coupling expansions give erroneous results in precisely the most interesting region. Thus strong coupling $1/N$ expansions are of virtually no value. This does not mean that the $1/N$ expansion fails; it means that if $1/N$ expansions are to succeed that they must be done nonperturbatively in β.

Roughly what is going wrong can be seen in Sec. VIII. The $1/N^2$ corrections in the $C_\alpha^c(N)$ get big as n gets big. For example, consider $C_n^c(N)$ which is just $N(n-1)!$ times $C_n(N)$ [Eq. (2.14)]. The ratio of the leading contribution of $C_n^c(N)$ to the $1/N^2$ correction is precisely

$$\text{Correction to } C_n^c(N) \text{ in large } N = \left(\sum_{m=1}^{n-1} m^2\right)\Big/ N^2 \sim n^3/N^2. \tag{9.6}$$

Other connected coefficients seem to have similar corrections. The $1/N^2$ corrections become uncontrollably large as n increases.[8] Even for $n < N$, $1/N^2$ corrections grow like N if n is close to N. Of course nth order corrections are important only if β is sufficiently large. This explains Eq. (9.5). Of course, De Wit–'t Hooft anomalies enter in the right-hand side of Eq. (9.5) but are absent from the left-hand-side and also ruin large N. These two effects are intimately related since the reason for Eq. (9.6) is a tower of poles in N at $-(n-1)$ through $(n-1)$.

TABLE IX. The $C_\alpha^c(N)$'s for $n = 4$.

	$N=2$	$N=3$	$N\geqslant 4$
$C_{1^4}^c(N)$	$\frac{7}{15}$	$-\frac{171}{320}$	$\frac{144N^2-216}{(N^2-1)^2(N^2-4)(N^2-9)}$
$C_{1^2,2}^c(N)$	$\frac{1}{15}$	$\frac{111}{160}$	$\frac{-288N^2+432}{(N^2-1)^2(N^2-4)(N^2-9)}$
$C_{1,3}^c(N)$	0	$-\frac{1}{6}$	$\frac{120}{(N^2-1)(N^2-4)(N^2-9)}$
$C_{2,2}^c(N)$	$-\frac{17}{60}$	$\frac{-11}{320}$	$\frac{54N^2-126}{(N^2-1)^2(N^2-4)(N^2-9)}$
$C_4^c(N)$	0	0	$\frac{-30}{(N^2-1)(N^2-4)(N^2-9)}$

Now consider the effect of De Wit–'t Hooft anomalies on the strong coupling expansion. One's attitude might be as follows: the connected coefficients, $C_1^c(N)$, are ratios of polynomials in N. They have poles at $N = 1, 2, \ldots, (n-1)$ and a completely different set of $C_1^c(N)$'s must be used for $N < n$. However, by expanding in a power series in $1/N^2$, terminating it after several orders, and extrapolating to $N < n$, the connected coefficients become finite. One might hope that via this extrapolation strong coupling contributions combine to give reasonable results and that $1/N$ corrections improve on this, thereby bypassing the De Wit–'t Hooft problem. It will be shown that this does not happen. Define

$$\Gamma_{p,\infty}^{(l)}(\beta) = \frac{1}{(N^2)^l}\sum_{n=1}^{p}\beta^{2n}c_{n,l}. \tag{9.7}$$

$\Gamma_{p,\infty}^{(l)}(\beta)$ is the l th contribution in the $1/N^2$ strong coupling expansion. For β sufficiently large $\Gamma_{p,\infty}^{(0)}(\beta)$ is a bad approximation to exact results. One might hope that the $1/N^2$ correction, $\Gamma_{p,\infty}^{(1)}(\beta)$, rectifies the situation (for $\beta > \frac{1}{2}$) both for finite and infinite N and improve results for $\beta < \frac{1}{2}$. Amazingly

$$\Gamma_{p,\infty}^{(l)}(\infty) = 0 \quad \text{for all } l \geqslant 1. \tag{9.8}$$

In the two-dimensional model in a strong coupling expansion all $1/N^2$ corrections are zero. It is impossible to bridge the gap between infinite N results and finite N results. Although a strong coupling $1/N$ expansion is a reasonable approximation for $\beta < \frac{1}{2}$, there is no way to improve on this by taking into account $1/N^2$ corrections. In higher dimensions, Eq. (9.8) does not hold and some improvement can be obtained by treating $1/N^2$ corrections; however, many contributions are still ruined by trying to extrapolate due to the De Wit–'t Hooft anomalies.

Equation (9.8) is proved by using the definitions of $C_1(N)$ and $f_r(N)$ in Appendix A. The contribution in nth order [obtained by expanding the exponent in Eq. (8.1) and picking out the term proportional to $(\beta'^2)^n$] is (for $n < N$)

$$\begin{aligned}
&\int dU \frac{(\beta')^{2n}}{n!n!}(\mathrm{tr}U)^n(\mathrm{tr}U^\dagger)^n \\
&= \frac{(\beta')^{2n}}{n!n!}\sum_{\substack{\text{partitions}\\ \alpha_1,\alpha_2,\ldots,\alpha_n \\ \alpha_1+2\alpha_2+\cdots+n\alpha_n=n}} g_n(\alpha)n!C_\alpha(N)N^{\alpha_1+\alpha_2+\cdots+\alpha_n} \\
&= \frac{(\beta')^{2n}}{n!}\sum_\sigma C(\sigma)N^{f(\sigma)} \\
&= \frac{(\beta')^{2n}}{n!}\sum_\sigma\sum_r \frac{\chi_r(e)\chi_r(\sigma)N^{f(\sigma)}}{n!f_r} \\
&= \frac{(\beta')^{2n}}{n!}\left[\sum_r \frac{\chi_r(e)\chi_r(e)}{n!}\right].
\end{aligned} \tag{9.9}$$

The term in brackets is 1. The nth contribution is just the exponentiation of the first order contribution: $\Gamma_{p,\infty}^{(0)}(\beta) = \beta^2$ for all p and $\Gamma_{p,\infty}^{(l)}(\beta) = 0$ for all p and $l \geqslant 1$. The first

TABLE X. The $C_\alpha^c(N)$'s for $n = 5$.

	$N=2$	$N=3$	$N=4$	$N\geqslant 5$
$C_{1^5}^c(N)$	$\frac{62}{45}$	$\frac{51}{140}$	$\frac{-304}{945}$	$\frac{4224N^2-13824}{(N^2-1)^2(N^2-4)(N^2-9)(N^2-16)}$
$C_{1^3,2}^c(N)$	$-\frac{28}{9}$	$\frac{-9}{56}$	$\frac{110}{189}$	$\frac{-10560N^2+34560}{(N^2-1)^2(N^2-4)(N^2-9)(N^2-16)}$
$C_{1^2,3}^c(N)$	0	$\frac{11}{168}$	$\frac{-194}{945}$	$\frac{4800N^2-9600}{(N^2-1)^2(N^2-4)(N^2-9)(N^2-16)}$
$C_{1,2^2}^c(N)$	$\frac{3}{2}$	$\frac{-17}{56}$	$\frac{-33}{280}$	$\frac{4320N^2-18720}{(N^2-1)^2(N^2-4)(N^2-9)(N^2-16)}$
$C_{1,4}^c(N)$	0	0	$\frac{1}{24}$	$\frac{-1680}{(N^2-1)(N^2-4)(N^2-9)(N^2-16)}$
$C_{2,3}^c(N)$	0	$\frac{17}{168}$	$\frac{41}{3780}$	$\frac{-1440N^2+6240}{(N^2-1)^2(N^2-4)(N^2-9)(N^2-16)}$
$C_5^c(N)$	0	0	0	$\frac{336}{(N^2-1)(N^2-4)(N^2-9)(N^2-16)}$

contribution to $\Gamma_{p,N}(\beta)$ beyond β^2 occurs when the first anomaly appears (i.e. at $p = N+1$):

$$\Gamma_N(\beta) = \beta^2 + \sum_{p=N+1}^{\infty} a_p(N)\beta^{2n}. \tag{9.10}$$

For finite N, the exact high temperature expansion begins with a β^2 term but the next term does not appear until the $(N+1)$th order in β^2. This explains the statement in Ref. 6 after Eq. (50). Equation (9.10) shows why all $1/N^2$ corrections vanish as $N\to\infty$ and shows how the De Wit–'t Hooft anomalies ruin a $1/N^2$ strong coupling approximation.

ACKNOWLEDGMENTS

I thank Frits Beukers for valuable and interesting discussions. I thank Paula Bozzay for the typing. Research is supported by the department of Energy under grant EY-76-S-02-2220.

APPENDIX A

The proof of Eq. (2.10) given below is due to Frits Beukers. Let $\lambda, \mu, \nu, \sigma$, and τ be permutations and let e be the identity permutation. Denote $F(\sigma) = N^{f(\sigma)}$ where $f(\sigma)$ is the number of cycles in σ. Let $\delta_{\sigma,e}$ be the delta function of S_n, that is

$$\delta_{\sigma,e} = \begin{cases} 1 \text{ if } \sigma = e \\ 0 \text{ otherwise} \end{cases}.$$

Important ingredients in the proof are:

(a) Uniqueness of the solution, C_{σ_1,σ_2} of Eq. (2.10);

(b) $F(\sigma) = F(\tau\circ\sigma\circ\tau^{-1})$, $F(\sigma^{-1}) = F(\sigma)$;

(c) Orthogonality relations for the characters of S_n:

$$\sum_{\sigma}\chi_r(\sigma^{-1}\circ\mu)\chi_{r'}(\sigma) = \frac{n!}{\chi_r(e)}\delta_{rr'}\chi_{r'}(\mu).$$

Equation (2.8) reads

$$\sum_{\mu,\nu} C_{\mu,\nu}F(\sigma\circ\mu^{-1})F(\tau\circ\nu^{-1}) = F(\sigma\circ\tau^{-1}). \tag{A1}$$

It is easy to see that $C_{\mu\circ\lambda,\nu\circ\lambda}$ also satisfies Eq. (A1) by plugging it in, changing summations to $\mu\to\mu\circ\lambda^{-1}$ and $\nu\to\nu\circ\lambda^{-1}$, and using (b) above. Uniqueness [(a) above] implies $C_{\mu,\nu} = C_{\mu\circ\lambda,\nu\circ\lambda}$ so that $C_{\mu,\nu}$ is a function of $\mu\circ\nu^{-1}$, which will be denoted by $C(\mu\circ\nu^{-1})$. Take Eq. (A1), shift the summation variable μ to $\mu\circ\nu$, and set $\tau = e$ to get an equation for $C(\mu)$

$$\sum_{\mu,\nu} C(\mu)F(\sigma\circ\nu^{-1}\circ\mu^{-1})F(\nu^{-1}) = F(\sigma). \tag{A2}$$

Equation (A2) can be satisfied if

$$\sum_{\mu} C(\mu)F(\sigma\circ\nu^{-1}\circ\mu^{-1}) = \delta_{\sigma\circ\nu^{-1},e}, \tag{A3}$$

and by uniqueness this must be the solution. The new equation to solve is

$$\sum_{\mu} C(\mu)F(\sigma\circ\mu^{-1}) = \delta_{\sigma,e}. \tag{A4}$$

It immediately follows from (A4) that $C(\mu) = C(\tau\circ\mu\circ\tau^{-1})$ and $C(\mu) = C(\mu^{-1})$. $C(\mu)$ is a class function. Equation (A4) is a group convolution of class functions and therefore diagonalizes by group Fourier transform, i.e. by writing all class functions in character expansions:

$$\begin{aligned} C(\mu) &= \sum_r C_r\chi_r(\mu), \\ F(\sigma\circ\mu^{-1}) &= \sum_r F_r\chi_r(\sigma\circ\mu^{-1}), \\ \delta_{\sigma,e} &= \sum_r \delta_r\chi_r(\sigma). \end{aligned} \tag{A5}$$

Here C_r, F_r, δ_r represent the Fourier components of C, F, and δ. Plug Eqs. (A5) into Eq. (A4) and use orthogonality [(c) above]:

$$\sum_r C_r F_r \frac{n!}{\chi_r(e)}\chi_r(\sigma) = \sum_r \delta_r\chi_r(\sigma), \tag{A6}$$

or

$$C_r = \delta_r/f_r, \tag{A7}$$

where by definition

$$f_r = F_r n!/\chi_r(e). \tag{A8}$$

F_r and δ_r are determined by taking the inverse Fourier transform of Eqs. (A5):

$$\begin{aligned} \delta_r &= \frac{1}{n!}\chi_r(e), \\ F_r &= \frac{1}{n!}\sum_{\sigma}\chi_r(\sigma)F(\sigma). \end{aligned} \tag{A9}$$

Summarizing,

$$C(\sigma) = \sum_r \frac{\chi_r(e)}{n! f_r}\chi_r(\sigma), \tag{A10}$$

$$f_r = \sum_{\sigma}\frac{\chi_r(\sigma)F(\sigma)}{\chi_r(e)}. \tag{A11}$$

Equation (A10) is precisely Eq. (2.10). The author has guessed the solution of Eq. (A11), namely that f_r is given by Eq. (2.12). This has been verified up to $n = 7$ but a general proof of Eq. (2.12) is lacking.

[1] B. De Wit and G. 't Hooft, Phys. Lett. B **69**, 61 (1977).

[2] See, for example, M. Hamermesh, *Group Theory and Its Application to Physical Problems* (Addison-Wesley, London, 1962), Chap. 7.

[3] See, for example, I. N. Herstein *Topics in Algebra* (Xerox, Toronto, 1964), Chap. 2, Sec. 10.

[4] I. Bars and F. Green, Phys. Rev. D **20**, 3311 (1979).

[5] D. J. Gross and E. Witten, Phys. Rev. D **21**, 446 (1980); Y. Y. Goldschmidt, "Expansion in Two-Dimensional Lattice Gauge Theory," Saclay Preprint (1979).

[6] S. Wadia, "A Study of U(N) Lattice Gauge Theory in Two-Dimensions," University of Chicago Preprint (1979).

[7] I. Bars, "U(N) Integral for Generating Functional in Lattice Gauge Theory," Yale University Preprint (1980).

[8] Actually the $1/N^2$ correction in Eq. (8.1) for $A = B =$ (the identity) vanishes. One should therefore use $A = B =$ (something close to the identity) to see what is going on.

Lie groups, spin equations and the geometrical interpretation of solitons

George Reiter
Brookhaven National Laboratory, Upton, New York 11973

(Received 11 April 1980; accepted for publication 23 May 1980)

The integrable evolution equations imbeddable in SU (2) are shown to have two gauge equivalent forms; the AKNS form, and a spin form for which the field is a three-dimensional vector of unit length. These equations are the compatibility conditions for the existence of a bilocal Lie group in two distinct frames of reference. These frames are associated with moving bases on surfaces formed by the motion of the strings introduced by Lamb. Both forms of the evolution equation are derivable from a locality assumption for the generators of the bilocal Lie group. The assumption is sufficient to distinguish between integrable and nonintegrable systems imbedded in SU (2).

The interpretation of soliton equations as the compatibility condition for the existence of a Lie group[1] suggests at the same time a direction in which one might be able to extend the structure these equations manifest to higher dimensions and a cohesive perspective for interpreting the insufficiently understood aspects of these equations in two dimensions. The latter set of concerns is of interest in its own right. The equivalence of the nonlinear Schrödinger equation and the continuum limit of the classical Heisenberg chain,[2] two systems shown independently[3,4] to be integrable by inverse scattering methods and of independent interest as descriptions of apparently diverse physical systems, is an example of what we regard as an insufficiently well understood aspect of two particular soliton equations. We will show, making use of the Lie group perspective that this equivalence is in fact a general feature of a class of soliton equations [those that can be imbedded in SU (2)]. That is, that the equations have two forms, a "ψ form," in terms of a complex field, which is the familiar form for most of the historically important examples, and an "S form," in terms of a unit vector on a sphere $\mathbf{S}(x,t)$.

The two forms of the equation correspond to the compatibility conditions for the existence of a group manifold expressed in two different bases. These bases arise naturally in giving the structure of the equations a geometrical interpretation. The spin vector is the tangent in the space direction of a particular one of a family of surfaces associated with each solution of the equations. The S form of the equation is associated with a coordinate system fixed in the three dimensional space in which the surface is imbedded. The ψ form is associated with a coordinate system whose orientation is determined by the tangent curve on the surface in the x direction and a free parameter that is the eigenvalue of the inverse scattering method.

The S form may be constructed from the ψ form by an algorithm that we will present in the case that they are evolution equations, i.e.

$$\frac{\partial\psi}{\partial t} = K(\psi,\psi^*,\psi_k,\psi_x^*\cdots). \tag{0.1}$$

We conjecture that there is a spin equation for all the equations that can be imbedded in SU (2), even those, such as the sine–Gordon equation which are not evolution equations, and, indeed there is such an equation for the sine–Gordon example. This turns out to be the same equation derived previously by Pohlmeyer.[5] After completing this work, we became aware that Zakharov and Taktajan[6] had obtained the transformation given by our algorithm for the particular case of the nonlinear Schrödinger equation and Heisenberg model.

We show in addition that a certain locality condition suffices to distinguish between integrable and nonintegrable equations imbeddable in SU (2), and suffices, furthermore, to produce the associated linear operator in the case that they are integrable. The "squared eigenfunctions" of the linear problem have a natural geometric interpretation here.

The Hamiltonian structure of the equations is also interesting. If Hn is a sequence of conserved densities, corresponding to a hierarchy of integrable equations in their ψ form, so that the nth equation of motion is

$$\frac{\partial\psi}{\partial t} = -i\frac{Hn+2}{\partial\psi^*},$$

then the corresponding spin equation is

$$\frac{\partial\mathbf{S}}{\partial t} = \mathbf{S}\times\frac{\delta Hn}{\delta\mathbf{S}},$$

where Hn is the nth conserved density of the ψ form of the equation, expressed in terms of S (and the nth conserved density of the S form as well).

We begin in Sec. I with a review of the connection between the geometrical interpretation of the soliton equations and the significance of the equations as the compatibility conditions for the existence of a bilocal Lie group. The construction that defines $\mathbf{S}(x)$ is given, and the relationship to the work of Lamb[7] on the motion of strings, and the later work by Lund[8] and Sym and Corones[9] is made explicit.

Section II shows how one may construct the "S form" given the linear problem of the "ψ form."

Section III gives the construction of the linear problem for a given evolution equation, and gives a criterion by which one can decide if the equation has an inverse scattering theory or not. The example of the modified KdV equation is

 0022-2488/80/122704-11$1.00

worked out in detail. A method of calculating the S form of the equations directly is given.

Section IV discusses the sine–Gordon example and its spin equivalent, and makes contact with the work of Pohlmeyer.

Section V discusses the relation between the Hamiltonian structures of the two equations.

I. GEOMETRICAL INTERPRETATION

Suppose we have a surface in R_3, parametrized by coordinates (s,t), where ds is taken to be the arc length along a coordinate curve for fixed t. That is a vector function $\mathbf{X}(s,t)$ such that $|d\mathbf{X}/ds| = 1$. Each coordinate curve for fixed t may be regarded as the position of a space curve, and the surface regarded as the locus of the curve as it moves in time. At each point of this space curve, we can define the Serret–Frenet basis, i.e., an orthonormal coordinate system in R_3, determined by the three unit vectors $\mathbf{t}$, $\mathbf{n}$, $\mathbf{b}$.

$$\mathbf{t} \equiv \frac{\partial \mathbf{X}}{\partial s}, \tag{1.1a}$$

$$\mathbf{n} \equiv \kappa^{-1} \frac{\partial t}{\partial s}, \quad \kappa = \left| \frac{\partial \mathbf{t}}{\partial s} \right|, \tag{1.1b}$$

$$\mathbf{b} = \mathbf{t} \times \mathbf{n}. \tag{1.1c}$$

The torsion of the curve is defined by

$$\frac{\partial \mathbf{b}}{\partial s} = -\tau \mathbf{n} \tag{1.2}$$

from which it follows that

$$\frac{\partial \mathbf{n}}{\partial s} = \tau \mathbf{b} - \kappa \mathbf{t}. \tag{1.3}$$

(1.1b), (1.2), (1.3) are called the Serret–Frenet equations, and give the evolution of the basis vectors as one moves along the curve in terms of two functions κ and τ, the curvature and torsion at each point.

If we pick two points (s_0,t_0) (s,t), then evidently there is a unique rotation that takes us from the Serret–Frenet basis at (s_0,t_0) to the Serret–Frenet basis at (s,t). Denote this abstract element of the rotation group by $g(s,t;s_0,t_0)$. If we pick a third point (s',t'), then the rotation that takes the basis at (s_0,t_0) into that at (s',t') must be the same as that which is obtained by first transforming to (s,t) and then transforming to (s',t'), i.e.,

$$g(s',t';s_0,t_0) = g(s',t';s,t)\, g(s,t;s_0,t_0). \tag{1.4}$$

This must be true for every point (s,t) and defines bilocal Lie group, here the rotation group in three dimensions, introduced by Corones, Markovski, and Rizov[1] (see Fig. 1).

We now imagine specifying not the surface, but the generators of the group at each point (s,t) κ and τ for instance can be used to parametrize the generator in the s direction, and there will be two additional functions that will specify the generator in the t direction. To be specific, we will represent g as an element of SU (2), two-dimensional unitary matrices with trace 1.

Then $g(s,t;s_0,t_0)$ can be obtained by integrating

$$g_s(s,t;s_0,t_0) = A(s,t)\, g(s,t;s_0,t_0)\, g(s_0,t_0;s_0,t_0) = I,$$
$$g_t(s,t;s_0,t_0) = B(s,t)\, g(s,t;s_0,t_0), \tag{1.5}$$

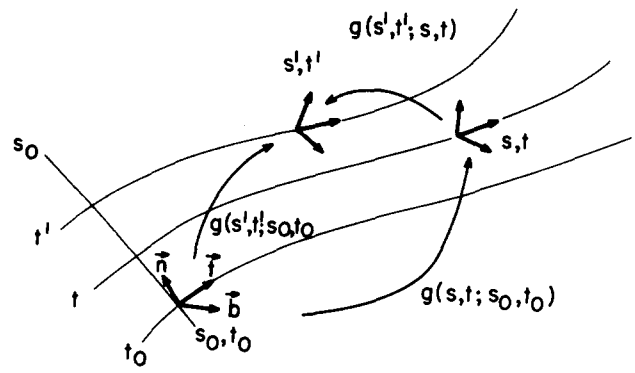


FIG. 1. A surface defines a bilocal Lie group and vice versa $g(s,t;s_0,t_0)$ rotates the basis at s_0,t_0 into the basis at s,t.

where $A(s,t)$ and $B(s,t)$ are elements of the Lie algebra of SU (2) i.e. can be represented as linear combinations of the 2×2 anti-Hermitian matrices α_i they satisfy

$$[\alpha_i,\alpha_j] = \epsilon_{ijk}\alpha_k. \tag{1.6}$$

Specifically, a representation for the α_i is

$$\alpha_1 = \tfrac{1}{2}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \alpha_2 = \tfrac{1}{2}\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}, \quad \alpha_3 = \frac{i}{2}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{1.7}$$

and

$$A(s,t) = \sum_{i=1} A_i(s,t)\alpha_i, \quad B(s,t) = \sum_{i=1} B_i(s,t)\alpha_i, \tag{1.8}$$

where the A_i, B_i are real functions.

We will write expressions such as (1.8) that map a vector field into the Lie algebra as $A = \mathbf{A}\cdot\boldsymbol{\alpha}$. To obtain the inverse mapping, we observe that

$$\alpha_i^2 = -1/4I,$$
$$\alpha_i\alpha_j = -\alpha_j\alpha_i, \quad i \neq j, \tag{1.9}$$

so that

$$A_i = -4T_r(A\alpha_i). \tag{1.10}$$

We note that

$$AB + BA = -\tfrac{1}{2}\mathbf{A}\cdot\mathbf{B} \tag{1.11}$$

and

$$\mathbf{A}\times\mathbf{B}\cdot\boldsymbol{\alpha} = [A,B]. \tag{1.12}$$

The specification of two vector fields $\mathbf{A}(s,t)$ $\mathbf{B}(s,t)$ then serves to define g, and, one may show, a surface that is associated with g. Of course, one must be able to simultaneously integrate both equations, that is, they must be compatible, so we are not free to choose $\mathbf{A}$ and $\mathbf{B}$ arbitrarily. The compatibility condition is simply

$$\frac{\partial g}{\partial s \partial t} = \frac{\partial g}{\partial t\, \partial s}, \tag{1.13}$$

which implies

$$A_t - B_x + [A,B] = 0. \tag{1.14}$$

If A and B satisfy (1.14), then one can obtain a g which has A and B as its generators.[1]

The known soliton equations can all be written in the form (1.14), for perhaps a different group, where A and B are parametrized by the field, perhaps complex, that satisfies the soliton equation. That is,

$$A(s,t) = A(\psi,\psi^*,\psi_s,\psi_s^*,\psi_{ss}\cdots)$$

and similarly for B, and then (1.14) is equivalent to the soliton equation for $\psi(s,t)$. This is what we mean by imbedding the soliton equation in a group.

We think it is important to point out that equations that are not soliton equations can also be put in this form. In fact, if we regard the surface as being traced out by a moving space curve, as in the work of Lamb, then it is clear that we will obtain a surface whatever the equation of motion for the curve, while only rather special types of equations of motion produce solitons. What makes these equations special is their relationship with a free parameter in the theory, the eigenvalue of earlier works.

The role of this parameter is not well understood in the present context. We will here introduce it in an ad hoc fashion by defining A to be of the form[3]

$$A(s,t) = \lambda\alpha_z - (i/2)\psi(s,t)\alpha^- + (i/2)\psi^*(s,t)\alpha^+, \tag{1.15}$$

where $\alpha^\pm = \alpha_x \pm i\alpha_y$. This is almost the most general form for A, we have only restricted the coefficient of α_z to be a constant, independent of x and t. ψ will be closely related with κ and τ in a manner we shall see shortly. If we want to represent the most general equation of motion for the space curve, we can parametrize B as[7]

$$B(s,t) = R(s,t)\alpha_z + (i/2)\gamma(s,t)\alpha^- - (i/2)\gamma^*(s,t)\alpha^+, \tag{1.16}$$

where $\gamma(s,t) = \gamma\{\psi\}$, and the bracket $\{\ \}$ denotes "a functional of." The compatibility conditions together with the commutation relations for the α_i imply the two equations

$$R_s = (i/2)[\gamma\psi^* - \psi\gamma^*], \tag{1.17a}$$

$$\psi_t + \gamma_x - i\lambda\gamma - i\psi R = 0. \tag{1.17b}$$

[The third equation given by (1.14) is the complex conjugate of (1.17b).] (1.17a) determines R, molulo an integration constant, in terms of ψ, and (1.17b) is an equation of motion for ψ. γ may also depend upon λ, and it is one of the remarkable feature of soliton equations that the dependence on λ cancels out of (1.17b) with an appropriate choice of the integration constant in (1.17a), and the λ dependence of γ. We do not, however, need λ at all to describe an arbitrary motion of the space curve, and if we set $\lambda = 0$ and let γ be an arbitrary functional of ψ, we can still obtain the equation of motion for ψ that will generate the surface determined by the choice of γ. It is not therefore, the ability to imbed equations of motion as the compatibility conditions for the existence of a surface that is the distinguishing feature of soliton equations, but rather the existence of a family of surfaces, corresponding to different values of the free parameter, all having the same equation of motion for the compatibility condition.

Lamb[7] gives various choices for the functional γ that lead to different known soliton equations and we refer the reader to his paper for examples.

We will now show the connection between ψ and κ and τ. (ψ in fact turns out to be the same function as defined by Lamb).

Consider a coordinate system fixed at (s_0,t_0). Then $g(s,t;\ s_0,t_0)$ provides the transformation of a vector, $\mathbf{v}$, in this coordinate system to one in the rotated frame, $\mathbf{v}'$, by

$$v' = gvg^{-1}, \tag{1.18}$$

where $v = \mathbf{v}\cdot\alpha$, $v' = \mathbf{v}'\cdot\alpha$. For the tangent vector $\mathbf{t}(s,t)$ previously introduced, we require that $\mathbf{t}' = \hat{z}$, i.e., g describes a transformation to a basis where the $\mathbf{z}$ axis coincides with the $\hat{z}$ axis of the Serret–Frenet basis. Then

$$t(s,t) = g^{-1}(s,t;\ s_0,t_0)\alpha_z g(s,t;\ s_0,t_0). \tag{1.19}$$

We define also

$$N^\pm(s,t) = g^{-1}(s,t;\ s_0,t_0)\alpha^\pm g(s,t;\ s_0,t_0). \tag{1.20}$$

The vectors associated with t, $\frac{1}{2}(N^+ + N^-)$, $\frac{1}{2}i[N^- - N^+]$ form an orthonormal basis and, as elements of the Lie algebra, these matrices satisfy the same commutation relations as do the α_i, i.e.,

$$[t,N^\pm] = \mp iN^\pm, \quad [N^+,N^-] = 2it. \tag{1.21}$$

This basis differs from the Serret–Frenet basis only by a space dependent rotation about $\mathbf{t}$. The magnitude of this rotation is determined by τ and λ. To make the identification complete, note that we have

$$t_s = g^{-1}[\alpha_z, g_s g^{-1}]g = g^{-1}[\alpha_z, A]g = \tfrac{1}{2}[\psi N^+ + \psi^* N^-], \tag{1.22a}$$

$$N_s^\pm = g^{-1}[\alpha^\pm, g_s g^{-1}] = \pm i\lambda N^\pm - \begin{Bmatrix}\psi\\ \psi^*\end{Bmatrix} t. \tag{1.22b}$$

While the Serret–Frenet equations imply

$$\mathbf{t}_s = \kappa\mathbf{n}, \tag{1.23}$$

$$\frac{\partial}{\partial s}[\mathbf{n} \pm i\mathbf{b}] = \mp i\tau[\mathbf{n} \pm i\mathbf{b}] - \kappa\mathbf{t}.$$

(1.23) and (1.22) are equivalent if we make the identification

$$\psi(s,t) = \kappa(s,t)\exp\left\{-i\int_{s_0,t}^{s,t}[\tau(s',t) + \lambda]\,ds'\right\},$$

$$N^\pm(s,t) = [\mathbf{n}(st) \pm i\mathbf{b}(s,t)]\cdot\boldsymbol{\alpha} \times \exp\left\{\pm i\int_{s_0,t}^{s,t}[(s',t) + \lambda]\,ds'\right\},$$

$$t(s,t) = \mathbf{t}(s,t)\cdot\alpha. \tag{1.24}$$

(1.24) differs from Lamb's work only in that λ is called $-\tau_0$ there. The transformation that maps the Serret–Frenet basis at (s_0,t_0) onto the Serret–Frenet basis at (s,t) is generated by

$$A'(s,t) = [-\tau(s,t)\alpha_z + \kappa(s,t)\alpha_y]. \tag{1.25}$$

The two generators are related by a gauge transformation. That is, if $g(s,t_0;\ s_0,t_0)$ is the solution of (1.5) with the A of (1.15), and $g'(s,t_0;\ s_0,t_0)$ is the solution of (1.5) with the A' of (1.25) then

$$g'(s,t_0;\ s_0,t_0) = g_1(s,t_0;\ s_0,t_0)\,g(s,t_0;\ s_0,t_0),$$

with

$$g_1(s,t_0;\ s_0,t_0) = \exp\left\{ - \int_{s_0,t_0}^{s,t_0} [\tau(s',t_0) + \lambda\ ]\alpha_z\ ds' \right\} \tag{1.26}$$

and A' and A are related by

$$A'(s,t) = g_{1,s} g_1^{-1} + g_1 A g_1^{-1}. \tag{1.27}$$

II. EXISTENCE OF INTEGRABLE SPIN EQUATIONS

We have seen that associated with each soliton equation that can be imbedded in SU (2) there is a family of surfaces, characterized by the eigenvalue parameter λ. The wavefunction ψ that satisfies the soliton equation parametrizes the generators of the bilocal Lie group associated with the surface and determines directly the infinitesimal rotation angle in a basis closely associated with the Serret–Frenet basis. We wish to show now how to construct the equivalent spin equation.

For evolution equations, that is, equations of the form (0.1), this construction is an algorithm, and leads to spin equations of the form

$$\frac{\partial \mathbf{S}}{\partial t} = \mathbf{S} \times \mathbf{K}'(\mathbf{S}, \mathbf{S}_x \cdots) = \frac{\partial \mathbf{B}}{\partial x}(S, S_x \cdots), \tag{2.1}$$

where we will determine K' and B explicitly in terms of K. For the sine–Gordon equation,

$$\frac{\partial^2 \psi}{\partial x \partial t} = \sin \psi \tag{2.2}$$

there is an equivalent spin equation with the spin vector identified in a similar way, although there is no algorithm for obtaining it. (We will call the variables of the previous section x henceforth.)

We define

$$S(x,t) = g_0^{-1}(x,t;x_0,t_0)\alpha_z g_0(x,t;x_0,t_0), \tag{2.3}$$

where

$$g_{0,x}(x,t;\ x_0,t_0) = [\ -(i/2)\psi(x,t)\alpha^- + (i/2)\psi^*(x,t)\alpha^+\ ] \times g_0(x,t;\ x_0,t_0) \tag{2.4}$$

and

$$g_0(x_0,t_0;x_0,t_0) = I.$$

That is, $\mathbf{S}(x,t)$ is the tangent vector to the surface generated with the parameter $\lambda = 0$, and with the tangent vector $\mathbf{t}$ at x_0,t_0 oriented along the z direction in a basis fixed in R_3.

The Serret–Frenet equations, in the form (1.22) with $\lambda = 0$ become

$$\begin{aligned} S_x &= \tfrac{1}{2}[\psi N + \psi^* N^+], \\ N_x^{\pm} &= -\begin{Bmatrix} \psi \\ \psi^* \end{Bmatrix} S. \end{aligned} \tag{2.5}$$

Since

$$(N^+)^2 = (N^-)^2 = 0,$$

while

$$[N^+ N^- + N^- N^+] = -I, \tag{2.6}$$

$$(S_x)^2 = -\tfrac{1}{4}|\psi|^2 I = -\tfrac{1}{4}\kappa^2 I.$$

Also

$$S_{xx} = \tfrac{1}{2}[\psi_x N^- + \psi_x^* N^+] - |\psi|^2 S, \tag{2.7}$$

$$[S_x, S_{xx}] = +(i/2)[\psi_x^* \psi - \psi_x \psi^*] S = -\kappa^2 \tau S \tag{2.8}$$

and since $S^2 = -\tfrac{1}{4}$,

$$S\,[S_x, S_{xx}] = \tfrac{1}{4}\kappa^2 \tau. \tag{2.9}$$

(2.6) and (2.9) are Lakshamanan's equations[2] relating S and ψ. They permit the inversion of (2.3), (2.4) to obtain ψ in terms of S. For the nonlinear Schrödinger equation, Lakshmanan was able to use them to show the equivalence of the soliton equation in the ψ form and the equation in its S form, the Heisenberg chain, by direct substitution. This procedure requires that one know the spin equation thought to be equivalent to the equation in its ψ form, and is not suitable for determining the spin equation. It also does not reveal the connection between the linear problems of the two forms of the equation, which in fact are related by a gauge transformation.

From (2.4) we have

$$\frac{\partial S}{\partial t} = g_0^{-1}[\alpha_z, g_{0,t}\, g_0^{-1}] g. \tag{2.10}$$

But

$$g_{0,t}\, g_0^{-1} = B(\psi, \lambda = 0) = \mathbf{B}(\psi, \lambda = 0) \cdot \alpha,$$

where by $B(\psi,\lambda)$ we mean a matrix of the form 1.16. Hence

$$\frac{\partial S}{\partial t} = [S, g_0^{-1} B(\psi, \lambda = 0) g_0] = [S, K'\{S\}], \tag{2.11}$$

where $K\{S\}$ is obtained by using the Lakshmanan equations to eliminate ψ in (2.4).

(2.11) shows in principle that there is an S form of the equations, although the means of calculating $K'\{S\}$ is so far purely formal. Furthermore, there is no reason to suspect that $\mathbf{K}'(S, S_x \cdots)$ will be a local function of S, i.e., involve only S and its derivatives. In fact, it appears that this is only the case if the original equation was a soliton equation, although one could in principle construct B as a functional of S, using the Lakshmanan equations, for any evolution equation. This locality requirement is tied up with the invariance of the form of the equation to a choice of λ in ways that are not clear to us at the moment.

(2.11) may also be interpreted as the compatibility conditions for the existence of a surface, in a frame related to that in which we have expressed the generator (1.15) (which we will call the ψ frame) by a gauge transformation. Specifically, let $g(\lambda)$ be the solution of (1.5) with (1.15) and a compatible expression for B, and let g_0 be defined as in (1.24). Then we introduce g_1 defined by

$$g(\lambda) = g_0 g_1(\lambda). \tag{2.12}$$

From (1.26) and (1.27) appropriately reinterpreted we find that

$$g_{1,x}(\lambda) = \lambda S(x,t) g_1(\lambda), \tag{2.13}$$

where S is defined by (2.3). This is the form of the linear eigenvalue problem used by Takhtajan[4] to integrate the Heisenberg chain. We will refer to the frame obtained by apply-

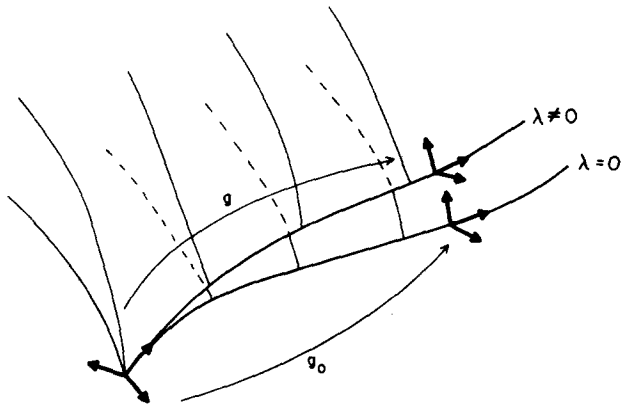


FIG. 2. Relationship between the elements g_0,g and the surfaces generated with $\lambda = 0$ and $\lambda \neq 0$. The moving frames are at the same distance, measured along their respective surfaces, from the fixed frame.

ing g_0^{-1} to the ψ frame as the S frame. Geometrically, if we have the two surfaces defined by the generators of (1.15) and (2.4), i.e. one for nonzero λ, one for $\lambda = 0$, then $g_0^{-1}(x,t;s_0,t_0)$ takes the ψ frame at any point t, for the surface with $\lambda = 0$, back to a fixed frame at x_0,t_0. When applied to the ψ frame for the $\lambda \neq 0$ surface at x,t, it returns the frame to one which is rotated from the fixed frame at x_0,t_0 by an amount that is determined by λ, and $\mathbf{S}(x,t)$ along the curve between x_0 and x (see Fig. 2). The generator for time translations associated with (2.13) is given by

$$B(S,\lambda) = g_0^{-1}[B(\psi,\lambda) - B(\psi,\lambda = 0)]g_0. \tag{2.14}$$

The compatibility condition (1.14) becomes

$$\lambda \frac{\partial S}{\partial t} - \frac{\partial B(S,\lambda)}{\partial x} + \lambda\ [S,B(S,\lambda)] = 0. \tag{2.15}$$

From (2.11) it follows that

$$\frac{\partial B(S,\lambda)}{\partial x} = \lambda\ [S,g_0^{-1}B(\psi,\lambda)g_0]. \tag{2.16}$$

If the equation in its ψ form is an evolution equation, then $B(S,\lambda)$ will be a polynomial in λ beginning with λ, as we will show later, so that

$$B(S,\lambda) = \lambda B_1 + \lambda^2 B_2 + \cdots \lambda^n B_n. \tag{2.17}$$

The highest power of λ is equal to the order of the highest derivative appearing in the evolution equation. The equation of motion (2.11) can also be written as

$$\frac{\partial S}{\partial t} = \frac{\partial B_1}{\partial x}, \tag{2.18}$$

where (2.15) implies

$$\frac{\partial Bn}{\partial x} = [S,Bn-1], \quad n \geqslant 1 \tag{2.19}$$

if we define B_0 to be $g_0^{-1}B(\psi,\lambda = 0)g_0$.

The vector associated with B_1 by the mapping (1.10) has the further property that it is the tangent in the t direction to the surface generated when $\lambda = 0$. To see this, observe that (2.18) is equivalent to

$$\frac{\partial}{\partial t}\frac{\partial \mathbf{X}}{\partial x}\cdot\boldsymbol{\alpha} = \frac{\partial}{\partial x}\frac{\partial \mathbf{X}}{\partial t}\cdot\boldsymbol{\alpha} = \frac{\partial(B_1)}{\partial x}. \tag{2.20}$$

Therefore

$$\frac{\partial \mathbf{X}}{\partial x}\cdot\boldsymbol{\alpha} = B_1 + C(t). \tag{2.21}$$

We will see later that B_1 always involves at least one derivative of S, and for the problem we are considering, $S \to \alpha_z$ as $x \to -\infty$. Evaluating $C(t)$ at $x \to -\infty$, we see that it is zero if we assume that $\partial \mathbf{X}/\partial t = 0$ at $x \to -\infty$, that is, that the end of our space curve is fixed in the reference frame we are considering. This seems sufficiently general for the boundary conditions we are considering, and we have then

$$\frac{\partial \mathbf{X}}{\partial x}\cdot\boldsymbol{\alpha} = B_1. \tag{2.22}$$

It is satisfying that the transformation of the "ψ form" into the "S form" of the equation also produces explicitly the information needed to construct the surface.

We do not yet have a self contained spin equation, since g_0 and $B(\psi,\lambda)$ are given in terms of ψ_λ with $\lambda = 0$.

To actually construct the spin equation for an arbitrary evolution equation requires that we first construct the appropriate linear operator $B(\psi,\lambda)$.

III. CONSTRUCTION OF INTEGRABLE SPIN EQUATIONS

Given a nonlinear evolution equation thought to be a soliton equation, the problem of finding the linear operators associated with it has generally been solved by guesswork and intuition. As pointed out by Corones,[9] if one assumes that the equation will be the compatibility conditions for the existence of a bilocal Lie group, then one can construct A and B if one knows, or thinks one knows, the group the equation corresponds to, by a procedure that appears to work generally. If one assumes that the group is SU (2), and one has chosen A to be in the form (1.15) one can in fact derive all possible integrable evolution equations associated with SU (2) and A in the form (1.15), from a simple locality condition on B. The locality condition, which states that B can only be a function of ψ and its derivatives at a given point, and cannot depend upon an integral of ψ over a region seems to us to be a necessary condition for the system to be integrable. For if B evaluated at $+\infty$ depended upon the values of ψ for all x, then the scattering data would not evolve simply in time with a frequency depending only on λ, and no decomposition into action angle variables labeled by λ would be possible.

In any event, we show here that the locality condition implies that the most general evolution equation, having a given linear dispersion relation, imbeddable in SU (2), with A given by (1.15), is of the AKNS[10] form. Furthermore, the B associated with a particular equation is provided automatically by the procedure we use to show this.

If g is any solution of

$$g_x = Ag, \tag{3.1}$$

and we wish to find a B such that

$$A_t - B_x + [A,B] = 0 \tag{3.2}$$

we can represent B as

$$B = g B_0 g^{-1}. \tag{3.3}$$

Then

$$B_x = [A,B] + g B_{0,x} g^{-1}. \tag{3.4}$$

If 3.2 holds, then

$$B_0(x,t) = B_0(x_0,t) + \int_{x_0}^{x} g^{-1} A_t g \, dx, \tag{3.5}$$

which is the result in AKNS. Now in the case we are considering

$$A_t = -\frac{i}{2}\psi_t \alpha^- + \frac{i}{2}\psi_t^* \alpha^+ \tag{3.6}$$

and

$$g^{-1} A_t g = -\frac{i}{2}\psi_t N^- + \frac{i}{2}\psi_t N^+ \tag{3.7}$$

if we take for g the solution which is the identity at x_0,t_0. If ψ_t satisfies an evolution equation, then $\psi_t = K(\psi,\psi_x,\psi_{xx},\cdots)$. Using the generalized Serret–Frenet equations (1.22), we can integrate (3.5) by parts repeatedly. If the equation is a soliton equation, we conjecture that the integration can be done completely and B_0 will be of the form

$$B_0 = P(\psi,\psi_x\cdots,\lambda)N^+ + P^*(\psi,\psi_x\cdots;\lambda)N^- + R(\psi,\psi_x\cdots,\lambda)t, \tag{3.8}$$

where the P,R are polynomial functions of their arguments.

Or stated differently, we assume that

$$-\tfrac{1}{2}i\psi_t N^- + \tfrac{1}{2}i\psi_t^* N^+ = \frac{\partial}{\partial x}[P(\psi,\psi_x\cdots\lambda)N^- + P^*(\psi,\psi_x\cdots\lambda)N^+ + R(\psi,\psi_x\cdots\lambda)t]. \tag{3.9}$$

Using 1.22 we find that the left hand side of (3.9) is equivalent to

$$\left[\frac{\partial P}{\partial x} - i\lambda P + \tfrac{1}{2}R\psi\right]N^- + \left[\frac{\partial P^*}{\partial x} + i\lambda P^* + \tfrac{1}{2}R\psi^*\right]N^+ + \left[\frac{\partial R}{\partial x} - P\psi^* - P^*\psi\right]t. \tag{3.10}$$

Since $N^{\pm},t$ are linearly independent matrices, (and correspond to an orthogonal basis of vectors) we have

$$\frac{\partial R}{\partial x} = P\psi^* + P^*\psi \tag{3.11}$$

or

$$R = R(-\infty) + \int_{-\infty}^{x} (P\psi^* + P^*\psi)\, dx. \tag{3.12}$$

Assuming that ψ vanishes at $x = -\infty$, $R(-\infty)$ is determined by the linear dispersion relation. For, if for sufficiently small $\psi \propto e^{i\lambda x}$

$$\psi_t = i\Omega(\lambda)\psi, \tag{3.13}$$

the the left-hand side of (3.9) is

$$\Omega(\lambda)\tfrac{1}{2}(\psi N^- + \psi^* N^+) = \Omega(\lambda)\frac{\partial t}{\partial x} \tag{3.14}$$

and (3.9) can be satisfied by choosing $R(-\infty) = \Omega(\lambda)$.

The two remaining equations that come from identifying the components of $N^{\pm}$ on the right- and left-hand sides of (3.9) can be written as

$$\mathscr{L}\begin{pmatrix} P \\ P^* \end{pmatrix} = \lambda\begin{pmatrix} P \\ P^* \end{pmatrix} + \frac{i}{2}\Omega(\lambda)\begin{pmatrix} \psi \\ -\psi^* \end{pmatrix} - \tfrac{1}{2}\begin{pmatrix} \psi_t \\ \psi^*_t \end{pmatrix}, \tag{3.15}$$

where

$$\mathscr{L} = \frac{1}{i}\begin{bmatrix} \dfrac{\partial}{\partial x} + \tfrac{1}{2}\psi\displaystyle\int_{-\infty}^{x}\psi^* & \tfrac{1}{2}\psi\displaystyle\int_{-\infty}^{x}\psi \\ -\tfrac{1}{2}\psi^*\displaystyle\int_{-\infty}^{x}\psi & -\dfrac{\partial}{\partial x} + \tfrac{1}{2}\psi\displaystyle\int_{-\infty}^{x}\psi \end{bmatrix}. \tag{3.16}$$

Let us assume that $\Omega(\lambda) = C_N\lambda^N$. Then the solution of (3.15) for $\binom{P}{P^*}$ is of the form

$$\begin{pmatrix} P \\ P^* \end{pmatrix} = \sum_{n=1}^{N} \lambda^{n-1}\begin{pmatrix} P_n \\ P^*_n \end{pmatrix}, \tag{3.17}$$

where

$$\begin{pmatrix} P_N \\ P^*_N \end{pmatrix} = -\frac{i}{2}C_N\begin{pmatrix} \psi \\ -\psi^* \end{pmatrix} \tag{3.18}$$

and

$$\begin{pmatrix} P_{n-1} \\ P^*_{n-1} \end{pmatrix} = \mathscr{L}\begin{pmatrix} P_n \\ P^*_n \end{pmatrix} \tag{3.19}$$

i.e.

$$\begin{pmatrix} P_n \\ P^*_n \end{pmatrix} = -\frac{i}{2}C_N\mathscr{L}^{N-n}\begin{pmatrix} \psi \\ -\psi^* \end{pmatrix}. \tag{3.20}$$

The term independent of λ leads to the equation of motion

$$\begin{pmatrix} \psi_t \\ \psi^*_t \end{pmatrix} - iC_N\mathscr{L}^N\begin{pmatrix} \psi \\ -\psi^* \end{pmatrix} = 0. \tag{3.21}$$

For an arbitrary polynomial dispersion relations it is easy to see that this generalizes to

$$\begin{pmatrix} \psi_t \\ \psi^*_t \end{pmatrix} - i\Omega(\mathscr{L})\begin{pmatrix} \psi \\ -\psi^* \end{pmatrix} = 0, \tag{3.22}$$

which is, with the definition (3.15), the AKNS equation appropriate to the special case we are considering. The expressions for P,P^* generalize simply to the linear combinations of the expressions for each power of λ appearing in the dispersion relation. AKNS derived these equations by considering the equations of motion for certain squared eigenfunctions. The relation between their method and ours can be seen by noting that if we denote the elements of g by

$$g = \begin{pmatrix} \phi_1 & \bar{\Phi}_1 \\ \phi_2 & \bar{\Phi}_2 \end{pmatrix}, \tag{3.23}$$

then the elements of $N^{\pm},t$ are quadratic products of the elements of g. Since the entries in the matrices correspond to components of the vectors associated with these matrices, the squared eigenfunctions can be thought of as the components of the basis vectors of the frame moving along the surface, in a fixed frame. For instance, if we integrate the equation for t_x in (1.15), and eliminate t from the remaining equations, we obtain

$$\mathscr{L}\begin{Bmatrix} N^+(x) \\ N^-(x) \end{Bmatrix} = \lambda\begin{Bmatrix} N^+(x) \\ N^-(x) \end{Bmatrix} + i\begin{pmatrix} \psi\alpha_z \\ \psi^*\alpha_z \end{pmatrix}. \tag{3.24}$$

This is nothing but the evolution equation for the basis vectors in the moving frame, with the condition that $t(-\infty) = \alpha_z$. We observe that

$$T_r\alpha = \begin{pmatrix} N^+ \\ N^- \end{pmatrix} = -\begin{pmatrix} \phi_1{}^2 \\ \phi_2{}^2 \end{pmatrix} \tag{3.25}$$

and hence the generalized Serret–Frenet equations (1.15) imply the eigenvalue equation

$$\mathscr{L}\begin{pmatrix} \phi_1{}^2 \\ \phi_2{}^2 \end{pmatrix} = -\lambda \begin{pmatrix} \phi_1{}^2 \\ \phi_2{}^2 \end{pmatrix}, \tag{3.26}$$

which, with slight changes of notation, is the starting point for the AKNS derivation of the form of the soliton equations. One may obtain the other evolution equations for different products of the ϕ_i, such as appear in Flaschka and Newell[11] by taking appropriate components of the equations (1.15), rewritten in the form (3.24).

Returning to our main theme, we see that we have succeeded in characterizing the possible integrable equations imbeddable in SU (2), with our choice of A, as well as obtaining the compatible expression for B. For, from (3.5) and (3.3) we see that

$$B(\psi,\lambda) = P\alpha^- + P^*\alpha^+ + R\alpha_z. \tag{3.27}$$

It is not obvious that P,P^*,R are in fact polynomials in ψ and its derivatives, since L is an integrodifferential operator, but this is the case. We do not have any proof of this for the moment, but observe that it can be shown by direct calculation for the lowest few terms. For instance, taking

$$\Omega(\lambda) = +\lambda^3, \tag{3.28}$$

leads to the modified KdV equation, for which

$$\begin{pmatrix} \psi_t \\ \psi^*_t \end{pmatrix} = i\mathscr{L}^3\begin{pmatrix} \psi \\ -\psi^* \end{pmatrix} = -\begin{pmatrix} \psi_{xxx} + 3/2|\psi|^2\psi_x \\ \psi^*_{xxx} + 3/2|\psi|^2\psi^*_x \end{pmatrix}, \tag{3.29}$$

and for which

$$P_3 - \tfrac{1}{2}i\psi, \quad P_2 = -\psi_x, \quad P_1 = +\tfrac{1}{2}i(\psi_{xx} + 1/2|\psi|^2\psi). \tag{3.30}$$

Defining R_n by

$$R = \sum_{n=0}^{N} \lambda^n R_n \tag{3.31}$$

we find, using (3.12) and (3.16)

$$R_3 = 1, \quad R_2 = 0, \quad R_1 = -\tfrac{1}{2}|\psi|^2,$$
$$R_0 = (i/2)[\psi_x\psi^* - \psi_x\psi]. \tag{3.32}$$

Having obtained $B(\psi,\lambda)$ for the general evolution equation, we can obtain $B(S,\lambda)$ from (2.14), and the equivalent spin equation follows from either (2.18) or (2.11), which we rewrite as

$$\frac{\partial S}{\partial t} = [S,B_0(\psi,\lambda = 0)]. \tag{3.33}$$

Of course, one has B_0 expressed in terms of ψ, and one wants it in terms of S. $B_0(S,\lambda = 0)$ is actually a simple functional of S in this case, and in the general case. It is rather easy to see how to do this, for our particular example.

We have, (since $t = S$ when $\lambda = 0$)

$$B_0(\psi,0) = [\tfrac{1}{2}i\psi_{xx} + \tfrac{1}{4}i|\psi|^2\psi]N^- + \{\text{c.c.}\} + \tfrac{1}{2}i(\psi_x\psi^* - \psi^*{}_x\psi)S. \tag{3.34}$$

From (2.7)

$$S_{xxx} = [\tfrac{1}{2}\psi_{xx} - \tfrac{1}{2}|\psi|^2\psi]N^- + \{\text{c.c.}\} - 3/2\,[\psi_x\psi^* + \psi_x\psi]S. \tag{3.35}$$

The term with the highest derivative in (3.20) can therefore be represented as, making use of (1.21)

$$[S,S_{xxx}] = \tfrac{1}{2}i[\psi_{xx}|\psi|^2\psi]N^- + \{\text{c.c.}\}. \tag{3.36}$$

Inasmuch as it is only the coefficients of $N^\pm$ that detemine the equation of motion, we have only to correct for the incorrect coefficient of $|\psi|^2\psi$ in (3.36) to obtain the desired spin equivalent of $B_0(\psi,0)$.

Using (2.5) and (2.7), we have

$$[S_x,S_{xx}] = \tfrac{1}{2}i|\psi|^2\psi N^- + \{\text{c.c.}\} + \tfrac{1}{2}i[\psi\psi^*{}_x - \psi^*{}_x\psi^*]S. \tag{3.37}$$

Thus, using (2.6) as well,

$$B_0(\psi,0) = +[S,S_{xxx}] + \tfrac{3}{2}[S_x,S_{xx}] + \{S\,[S_x,S_{xx}] + [S_x,S_{xx}]S\}S. \tag{3.38}$$

Hence, the equivalent spin equation for the modified KdV equation is

$$\frac{\partial S}{\partial t} = [S,\{[S,S_{xxx}] + \tfrac{3}{2}[S_x,S_{xx}]\}]. \tag{3.39}$$

To convert this back to an equation for the vector **S**, we note that

$$[A,B] = \mathbf{A}\times\mathbf{B}\cdot\alpha \tag{3.40}$$

from which we conclude that

$$\frac{\partial \mathbf{S}}{\partial t} = \mathbf{S}\times|\,|\,\mathbf{S}\times\mathbf{S}_{xxx} + \tfrac{3}{2}\mathbf{S}_x\times\mathbf{S}_{xx}|\,|. \tag{3.41}$$

Although we have used an apparently ad hoc procedure to pass from $B_0(\psi,0)$ expressed in terms of ψ to its form in terms of S for this particular example, the procedure may be systematized. We observe that $\mathbf{S}, \mathbf{S}_x, \mathbf{S}\times\mathbf{S}_x$ are an orthogonal basis, and we have

$$S_x \pm i[S,S_x] = \begin{Bmatrix} \psi^* \\ \psi \end{Bmatrix} N^\pm. \tag{3.42}$$

Differentiating once, we have

$$S_{xx} \pm i[S,S_{xx}] = \begin{Bmatrix} \psi_x \\ \psi_x \end{Bmatrix} N^\pm - |\psi|^2 S. \tag{3.43}$$

Since $|\psi|^2 = -4S_xS_x$, (3.4) expresses terms of the form $\binom{\psi^*}{\psi}N^\pm$ in terms of S and its derivatives. Taking another derivative, we see that $\binom{\psi^*_x}{\psi_x}N^\pm$ can be expressed in terms of S and its derivatives, plus a term that is

$$\begin{pmatrix} \psi^*{}_x\psi \\ \psi_x\psi^* \end{pmatrix} S.$$

But this can be expressed as for instance $[\psi_x N^+,\psi N^-]$, which have both been previously expressed in terms of S and its derivatives. Continuing in this way we see that we can always express any term of the form $\partial^n\psi N^-$, $\partial^n\psi^* N^+$ in terms of S and its derivatives.

If we assign an index to polynomials in ψ and ψ^* and their derivatives, $P(\psi,\psi^*,\psi_x\cdots)$, according to how they

transform under an ordinary gauge transformation, i.e.,

$$\psi \to e^{i\alpha}\psi,$$
$$\psi^* \to e^{-i\alpha}\psi^*, \tag{3.44}$$
$$P(\psi,\psi^*,\psi_x\cdots) \to e^{in\alpha}P(\psi,\psi^*,\psi_x\cdots),$$

then one sees immediately from the specific form of the AKNS evolution operator (3.16) that only terms with index $+1$ enter the equation for ψ_t, -1 for ψ^*_t. The most general term that will be needed to express B_0 in terms of S and its derivatives will be of the form

$$\left[\left(\frac{\partial^n\psi}{\partial_x{}^n}\right)^\alpha\left(\frac{\partial^m\psi^*}{\partial_x{}^m}\right)^\beta\psi^\gamma\psi^{*\delta}\right]N^- \tag{3.45}$$

or its complex conjugate, where $\alpha+\gamma-\beta-\delta=1$. If $\gamma>\delta$, we can replace $|\psi|^{2\delta}$ by $(S_xS_x)^\delta$ leaving a term

$$\left(\frac{\partial^n\psi}{\partial_x{}^n}\right)^\alpha\left(\frac{\partial_{\psi^*}}{\partial_x{}^m}\right)^\beta\psi^{\gamma-\delta}N^- \tag{3.46}$$

still to be represented in terms of S and its derivatives.

If $\delta>\gamma$, we obtain

$$\left(\frac{\partial^n\psi}{\partial\chi^n}\right)^\alpha\left(\frac{\partial^m\psi^*}{\partial\chi^m}\right)\psi^{*\delta-\gamma}N^-. \tag{3.47}$$

In the case of (3.46), we can construct the expression by commuting α factors of $(\delta^n\psi/\delta x^n)N^-$, $(\gamma-\delta)$ factors of ψN^- and β factors of $(\delta^m\psi/\delta x^m)N^-$, alternating factors proportional to N^- and N^+. Since $\alpha+(\gamma-\delta)=\beta+1$, we will have one more N^- term than N^+ term, and the result will be proportional to N^-. Similarly for (3.47). In all cases, we can reduce the terms that appear in the expression for $B_0(\psi,\lambda)$ arising from an AKNS evolution equation to terms involving S and its derivatives. Hence, one can always obtain the spin equivalent of an AKNS evolution equation in SU (2) by the procedure outlined above. We have, therefore, an algorithm for the construction of the spin equivalents.

The expression for $B(S,\lambda)$ that is associated with $A(S,\lambda)=\lambda S$ is obtained by subtracting from $B_0(\psi,\lambda)$ the value of $B_0(\psi,\lambda=0)$ and converting the remainder to its spin equivalent. The equation of motion can be obtained this way as well, using (2.18). The result will not be manifestly in the form (2.11) however, and it will generally require some manipulation of the identities that follow from differentiating $S^2=-\frac{1}{4}$ to put it in that form. For instance, for the modified KdV equation

$$\begin{aligned} B_1 &= -\tfrac{1}{2}\psi_xN^- - \tfrac{1}{2}\psi^*{}_xN^+ - \tfrac{1}{2}|\psi|^2S \\ &= -[S_{xx}-6S_xS_xS], \end{aligned} \tag{3.48}$$

and

$$\frac{\partial S}{\partial t} = -[S_{xxx}-6(S_xS_{xx}+S_{xx}S_x)S-6S_xS_xS_x], \tag{3.49}$$

which is not self-evidently the same equation as (3.41). Using

$$\begin{aligned} &SS_x+S_xS=0, \\ &2S_xS_x+SS_{xx}+S_{xx}S=0, \\ &SS_{xxx}+S_{xxx}S+3(S_xS_{xx}+S_{xx}S_x)=0, \end{aligned} \tag{3.50}$$

one can, however, transform (3.50) into (3.41).

The full expression for $B(S,\lambda)$ is, for the KdV equation

$$B(S,\lambda) = -\lambda[S_{xx}-6S_xS_xS]-\lambda^2[S,S_x]+\lambda^3S. \tag{3.51}$$

For the nonlinear Schrödinger equation $[\Omega(\lambda)=\lambda^2]$, we have from (3.18)

$$B_1=\left[\frac{i}{2}\psi N^- - \frac{i}{2}\psi^*N^+\right]=[S,S_x], \tag{3.52}$$

which leads to the equations for the Heisenberg model,

$$\frac{\partial S}{\partial t}=[S,S_{xx}] \quad \text{or} \quad \frac{\partial\mathbf{S}}{\partial t}=\mathbf{S}\times\mathbf{S}_{xx}. \tag{3.53}$$

The full expression for $B(S,\lambda)$ in this case is

$$B(S,\lambda)=\lambda[S,S_x]-\lambda^2S=2\lambda SS_x-\lambda^2S. \tag{3.54}$$

(3.43) is equivalent to the expression for B given by Taktajan.[4] In fact, we have shown that all the AKNS evolution equations that can be imbedded in SU (2) have equivalent spin equations with a linear eigenvalue problem in the Taktajan form, i.e., $A=\lambda S$.

eigenvalue problem in the Taktajan forms i.e. $A=\lambda S$.

We have seen that the assumption of locality in the ψ frame leads directly to the AKNS equations, which may then be converted to an S form. The S form may be obtained directly by requiring the locality to hold in the S frame. That is, we require that there exist a $B(\lambda)$ depending only on S and its derivatives, such that

$$g_1{}^{-1}\lambda S_t\,g_1=\frac{\partial}{\partial x}(g_1{}^{-1}B(\lambda)g_1). \tag{3.55}$$

This leads immediately to (2.15) and the relations (2.18) and (2.19) for the coefficients Bn defined in (2.17). These recursion relations may be solved for the Bn. We have

$$Bn-1=-\left[S,\frac{\partial Bn}{\partial x}\right]+(\mathbf{S}\cdot\mathbf{B}n-1)S. \tag{3.56}$$

Since $\partial\mathbf{B}n-1/\partial x$ has no component in the S direction, we

TABLE I. The first few elementary tangent vectors in the time direction, and their projections along the tangent in the space direction.

j	$\mathbf{B}_{N-j}$	$\mathbf{S}\cdot\mathbf{B}_{N-j}$
0	$\mathbf{S}$	1
1	$-\mathbf{S}_x\mathbf{S}_x-\mathbf{S}\times\mathbf{S}_x$	0
2	$-\mathbf{S}_{xx}-\frac{3}{2}(\mathbf{S}_x\cdot\mathbf{S}_x)\mathbf{S}$	$-\frac{1}{2}\mathbf{S}_x\cdot\mathbf{S}_x$
3	$\mathbf{S}\times\mathbf{S}_{xxx}+\frac{3}{2}(\mathbf{S}_x\cdot\mathbf{S}_x)\mathbf{S}\times\mathbf{S}_x$ $-(\mathbf{S}\cdot\mathbf{S}_x\times\mathbf{S}_{xx})\mathbf{S}$	$-\mathbf{S}\cdot\mathbf{S}_x\times\mathbf{S}_{xx}$
4	⋮	$\frac{15}{8}(\mathbf{S}_x\cdot\mathbf{S}_x)^2-\frac{3}{2}\mathbf{S}_{xx}\cdot\mathbf{S}_{xx}-\frac{1}{3}\frac{\partial}{\partial x}(\mathbf{S}\cdot\mathbf{S}_{xxx})$

have

$$\frac{\partial(\mathbf{S}\cdot\mathbf{B}n-1)}{\partial x}=S_x\times\frac{\partial\mathbf{B}n}{\partial x}\cdot\mathbf{S}. \tag{3.57}$$

Again, it is not obvious that the right-hand side of (3.57) is in fact a perfect derivative, but that does prove to be the case. With $\mathbf{B}_N=\mathbf{S}$, we have for $\mathbf{B}_{N-j}$, the results shown in the first column of Table I. In column two we give the associated expression for $\mathbf{S}\cdot\mathbf{B}_{N-j}$.

IV. SINE–GORDON GEOMETRY

The sine–Gordon equation is not amenable to being cast into a spin equation by the method above as it is not an evolution equation. We will treat it here as a special case, and show that nevertheless, a spin equation exists, and is in fact identical with that already obtained by Pohlmeyer. The derivation will make clear the relationship between the Lie group approach and the various geometrical interpretations of the equation.

We begin with the linear problem for a curve of constant torsion, $\tau=-\lambda$, and we define the curvature as $\sigma_x(x,t)$. Then ψ can be taken as real and equal to $\sigma_x(x,t)$, and the generator of g in the space direction becomes

$$A=\lambda\alpha_z+\sigma_x\alpha_y. \tag{4.1}$$

In this case, the frame of reference is identical with the Seret–Frenet frame.

If the curvature is to satisfy the sine–Gordon equation,

$$\sigma_{xt}=\sin\sigma, \tag{4.2}$$

then the generator in the time direction must be

$$B=-1/\lambda\,((\cos\sigma)\alpha_z+(\sin\sigma)\alpha_x). \tag{4.3}$$

Let us define

$$S(x,t)=g^{-1}(x,t)\alpha_y g(x,t), \tag{4.4}$$

where we have suppressed the initial coordinate (x_0,t_0) in the definition of g, and otherwise it satisfies (1.5). S is the binormal to the curve in this case, rather than the tangent.

Then we readily verify that

$$S_x=\lambda g^{-1}\alpha_x g, \tag{4.5a}$$

$$S_t=-(1/\lambda)g^{-1}(\cos\sigma\alpha_x-\sin\sigma\alpha_z)g, \tag{4.5b}$$

$$S_xS_x=-\tfrac{1}{4}\lambda^2, \tag{4.5c}$$

$$S_tS_t=-\tfrac{1}{4}(1/\lambda^2), \tag{4.5d}$$

$$S_xS_t+S_tS_x=\tfrac{1}{2}\cos\sigma. \tag{4.5e}$$

Thus, for the vector $\mathbf{S}$ we have

$$\mathbf{S}_x\cdot\mathbf{S}_x=\lambda^2,\quad \mathbf{S}_t\cdot\mathbf{S}_t=1/\lambda^2,\quad \mathbf{S}_x\cdot\mathbf{S}_t=-\cos\sigma. \tag{4.6}$$

Inasmuch as $\mathbf{S}_x,\mathbf{S}_t$ must be perpendicular to $\mathbf{S}$, the relationships between the three vectors and σ is as shown in Fig. 3. We have finally, the equation of motion for S, obtained from (4.5a) and (4.3) or (4.5b) and (4.1),

$$S_{xt}=(\cos\sigma)S \tag{4.7}$$

or

$$\mathbf{S}_{xt}+(\mathbf{S}_x\cdot\mathbf{S}_t)\mathbf{S}=0. \tag{4.8}$$

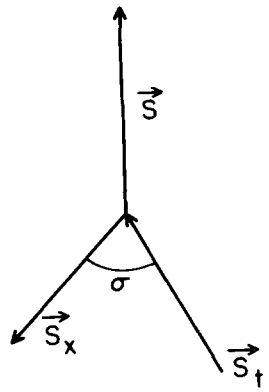


FIG. 3. The relationship between the spin vector and its derivatives when σ satisfies the sine–Gordon equation.

This equation was first derived by Pohlmeyer from the Lagrangian density

$$\mathscr{L}=\frac{\partial\mathbf{S}}{\partial t}\cdot\frac{\partial\mathbf{S}}{\partial t}-\frac{\partial\mathbf{S}}{\partial x}\cdot\frac{\partial\mathbf{S}}{\partial x}, \tag{4.9}$$

with the constraint $\mathbf{S}\cdot\mathbf{S}=1$. The σ of his work is $\pi-\sigma$ of ours.

The parameter λ serves to change the lengths of S_x, S_t, and can be regarded as arising from a lorentz transformation of the coordinates. That is, if we define x',t' by

$$x\to\lambda^{-1}x',\quad t\to\lambda t', \tag{4.10}$$

the linear problem in the primed coordinates reduces to that one obtains by setting $\lambda=1$ in (4.11) and (4.3). Evidently $\lambda=1$ has a special role in the problem, analogous to that for $\lambda=0$ in the case of evolution equations. As in that case also, we can obtain the $\lambda\neq1$ case by a gauge transformation from $\lambda=1$.

If we define

$$g(\lambda)=g(1)g', \tag{4.11}$$

then a straightforward calculation shows that

$$\begin{aligned} g'_x&=(1-\lambda)[\tilde{S},\tilde{S}_x]g',\\ g'_t&=(1-1/\lambda)[\tilde{S},\tilde{S}_t]g', \end{aligned} \tag{4.12}$$

where

$$\begin{aligned} &\tilde{S}=g(1)^{-1}\alpha_y g(1),\\ &S=g'^{-1}\tilde{S}g',\\ &\tilde{S}_x\cdot\tilde{S}_x=\tilde{S}_t\cdot\tilde{S}_t=1. \end{aligned} \tag{4.13}$$

g is the coadjoint representation in SU (2) of the elements of O (3) denoted as R_λ by Pohlmeyer.

(4.12) is analogous to the linear problem of Taktajan for the S form of the equation, for the compatibility equations imply that

$$\mathbf{S}_{xt}=\lambda\tilde{\mathbf{S}}. \tag{4.14}$$

But $\mathbf{S}\cdot\mathbf{S}=1$ implies $\tilde{\mathbf{S}}\cdot\tilde{\mathbf{S}}_{xt}=-\tilde{\mathbf{S}}_x\cdot\tilde{\mathbf{S}}_t$ and hence we have the equation of motion (4.8) as the compatibility condition.

V. HAMILTONIAN STRUCTURE

The ψ form of the equation we have been considering can be written in terms of a Poisson bracket,[11,12] defined for

two functionals of ψ as

$$\{A\{\psi\},B\{\psi\}\}_1 = i\int dx\left[\frac{\partial A}{\partial\psi}\frac{\partial B}{\partial\psi^*} - \frac{\partial A}{\partial\psi^*}\frac{\partial B}{\partial\psi}\right] \tag{5.1}$$

and a Hamiltonian Hn such that the nth equation of motion is

$$\psi_t = \{Hn,\psi\} = -i\frac{\partial Hn}{\partial\psi^*}. \tag{5.2}$$

The Hn are conserved quantities for all the equations that are integrable using the linear problem associated with (1.15). The lowest few Hn are $Hn = \int\mathscr{H}\,ndx$, $\mathscr{H}_1 = |\psi|^2$, $\mathscr{H}_2 = i/2(\psi_x\psi^* - \psi_x{}^*\psi)$, $\mathscr{H}_3 = |\psi_x|^2 - \frac{1}{4}|\psi|^4$. The first two constants yield linear equations when inserted in (5.2).

Since S is a functional of ψ, we must have

$$\frac{\partial\mathbf{S}}{\partial t} = \{Hn,\mathbf{S}\}_1. \tag{5.3}$$

A straightforward calculation, using a result readily obtained by functional differentiation of (1.5) using (1.15), shows that

$$\frac{\partial S(x)}{\partial\psi(x')} = \left[S(x), -\frac{i}{2}N^-(x')\right], \quad x>x',$$
$$= 0, \quad x<x'. \tag{5.4}$$

From (5.4) and its conjugate relation, using (5.2), we have

$$\frac{\partial S}{\partial t} = \left[S(x), \int_{-\infty}^{x}\left[-\frac{i}{2}N^-(x')\psi_t + \frac{i}{2}N^+(x')\psi_t{}^*\right]dx'\right], \tag{5.5}$$

which from (3.7) and (3.3) is equivalent to (2.11), and is the correct equation of motion for S.

The constants of the motion can all be expressed in terms of spin fields. There is another Poisson bracket defined for functionals of a spin degree of freedom from which one typically obtains the equations of motion for spin fields in physical applications,

$$\{A\{\mathbf{S}\},B\{\mathbf{S}\}\}_2 = \epsilon_{ijk}\frac{\partial A}{\partial S_i}\frac{\partial B}{\partial S_j}S_K. \tag{5.6}$$

As we have seen, the integrand appearing in (5.5) is actually a perfect derivative, and the integral can be expressed entirely in terms of the field S and its derivatives at x. Remarkably, when this is done, we find that Eq. (5.5) can also be written as

$$\frac{\partial S}{\partial t} = \left[S, \frac{\partial H'n-2}{\partial\mathbf{S}}\cdot\alpha\right], \tag{5.7}$$

where $H'n = CnHn$, Cn a constant. (5.7) is equivalent to

$$\frac{\partial\mathbf{S}}{\partial t} = \{H'n-2,\mathbf{S}\}_2 = \mathbf{S}\times\frac{\partial H'n-2}{\partial\mathbf{S}}. \tag{5.8}$$

That is, for any $n>2$, we conjecture that (5.2) and (5.8) are the equivalent pair of equations derived previously. We have no proof of this at the moment, but show in Table II that it holds for the first few densities. (It is well known that H_4 gives the modified KdV equation, and one may check that the equations associated with H_5 gives the same spin equation as H'_3 by observing that (5.8) can also be obtained from the B_{N-j} of Table I using Eq. (2.18), and the results agree with Table II.

TABLE II. The first few conserved densities in their ψ and S form. The spin equation of motion are $\partial\mathbf{S}/\partial t = \mathbf{S}\times(\partial H'n/\partial\mathbf{S})$.

Hn	$H'n$	$\partial H'n/\partial\mathbf{S}$
$\|\psi\|^2$	$-\frac{1}{2}\mathbf{S}_x\cdot\mathbf{S}_x$	$\mathbf{S}_{xx}$
$(i/2)(\psi_x{}^*\psi - \psi_x\psi^*)$	$\frac{1}{2}\mathbf{S}\cdot\mathbf{S}_x\times\mathbf{S}_{xx}$	$\mathbf{S}\times\mathbf{S}_{xxx} + \frac{1}{2}\mathbf{S}_x\times\mathbf{S}_{xx}$
$\frac{1}{4}\|\psi\|^4 - \|\psi_x\|^2$	$\frac{1}{2}\mathbf{S}_{xx}\cdot\mathbf{S}_{xx} - \frac{5}{4}(\mathbf{S}_x\cdot\mathbf{S}_x)^2$	$\mathbf{S}_{xxxx} + 5(\mathbf{S}_x\cdot\mathbf{S}_{xx})\mathbf{S}\times\mathbf{S}_x + \frac{5}{2}(\mathbf{S}_x\cdot\mathbf{S}_x)\mathbf{S}\times\mathbf{S}_{xx}$

Comparing Tables I and II, we make one further conjecture, that the $S\cdot Bn$ are to within a multiplicative constant and a divergence, the conserved densities. If true, this would provide a simple geometric interpretation for the constants of the motion.

VI. EXTENSIONS

Some of the results presented here have extensions to more general settings. The requirement of locality as a means of constructing the B operators has a natural generalization to other groups, and clearly generates the AKNS equations associated with SL $(2,R)$. It would be interesting to compare its predictions for SL $(3,R)$ with the results of the Gel'fand–Dikii[14] analysis.

The notion of strings moving in space–time clearly generalizes to that of surfaces moving in space–time, and the problem then is to find a parametrization of the surface such that the compatibility conditions can be fulfilled simultaneously. One would expect the space directions to be equivalent, and the compatibility conditions for these directions to be satisfied identically.

Since the Miura transformation maps the generalized KdV equations onto the modified KdV equations, these also have spin equivalents, and we suspect the Miura transformation can be given a geometrical interpretation.

ACKNOWLEDGMENTS

I would like to thank Jim Corones for his encouragement, support and innumerable valuable interactions throughout this work, George Wilson and Boris Kuperschmidt who are largely responsible for focusing my attention on the Hamiltonian structure of the theory, and Sharon Hoeffner for her careful attention to detail in preparing the manuscript. Research has been performed under Contract DE-AC02-7600016 with the U. S. Department of Energy, Division of Basic Energy Sciences.

[1]J. Corones, B. Markovski, and V. Rizov, J. Math. Phys. **11**, 2207 (1977).
[2]M. Lakshmanan, Phys. Letts. A **61**, 53 (1977).
[3]V. E. Zakharov and A. B. Shabat, Zh. Eksp. Teor. Fiz. **61**, 118 (1971).
[4]L. A. Takhtajan, Phys. Lett. A **64**, 235 (1977).
[5]K. Pohlmeyer, Commun. Math. Phys. **46**, 207 (1976).
[6]V. E. Zakharov and L. A. Takhtajan, Teor. Mat. Fiz. **38**, 26 (1979).
[7]G. L. Lamb, Phys. Rev. Lett. **37**, 235 (1976); J. Math. Phys. **18**, 1654 (1977).
[8]F. Lund, Phys. Rev. Lett. **38**, 1175 (1977); Phys. Rev. D **15**, 1540 (1977).

[9]A. Sym and J. Corones, Phys. Rev. Lett. **42**, 1099 (1979).
[10]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[11]H. Flaschka and A. Newell, *Integrable Systems of Nonlinear Evolution Equations, Dynamical Systems, Theory and Applications*, edited by J. Moser (Springer, New York, 1975), pp. 355–440.
[12]V. E. Zakharov and S. V. Manakov, Teor. Mat. Fiz. **28**, 38 (1976).
[13]W. Symes, J. Math. Phys. **20**, 4 (1979).
[14]I. M. Gel'fand and L. A. Dikii, Funct. Anal. Pril. **12**, 8 (1976).

Generalized groups as global or local symmetries

J. G. Taylor
Department of Mathematics, King's College, London, England

(Received 31 August 1979; accepted for publication 9 November 1979)

We extend global particle symmetries from the traditional group framework to that of generalized groups. The nature of these latter are presented, and various invariants constructed for them. The problem of gauging generalized groups is discussed and a no–go theorem proved under reasonable conditions on the generalized group structure.

I. INTRODUCTION

Since group theory has been so useful in analyzing the natural world it is of interest to ascertain if any more generalized notion than that of a group would also be of value. In particular one can ask if it is possible that the symmetries of elementary particles could be clarified by such a generalization. It is our purpose in this paper to attempt to answer this latter question in the case of generalized groups. These replace the requirement that the (binary) product of two elements of a group belong to the group by the condition that an n-fold product belong to the generalized n-group. Thus for $n = 3$, a generalized 3-group G_3 is essentially a set of elements $(g_1, g_2, \cdots)$ such that for any g_1, g_2, and g_3 in G_3 the generalized product $(g_1g_2g_3)$ is also in G_3 (though a product of any pair of elements need not even be defined).

Generalized groups have been considered at an abstract level[1,2] but we will follow our earlier work[3,4] and consider them in a more concrete form. In particular we will consider problems associated with their representations and invariants, and of their putative gauging. We will also restrict our discussion solely to that of 3-groups, though much of it is very similar for other n-groups.

One of the most important concepts in applications of group theory is that of an infinitesimal group element. In order for such a concept to exist it will be necessary to require the existence of an identity element e, which we define by the condition

$$(e^2g) = (ege) = (ge^2) = g, \tag{1.1}$$

for any $g \in G_3$. We may define the inverse g^{-1} of g by

$$(geg^{-1}) = e. \tag{1.2}$$

A 3-group G_3 is thus defined as a set of elements with the binary product $g_1, g_2, g_3 \to (g_1g_2g_3)$ in G_3 satisfying (1.1) and for which all elements g in G_3 have an inverse g^{-1} in G_3 satisfying (1.2).

A concrete example of a 3-group is the three-dimensional array g_{ijk}, where i, j, and k are integers (though continuous variables could be included), with 3-product one of the four possible expressions for $(g^{(1)}\ g^{(2)}\ g^{(3)})_{ijk}$ (the summation convention is used)

$$g^{(1)}_{ilm}g^{(2)}_{mjn}g^{(3)}_{lnk}, \tag{1.3a}$$

$$g^{(1)}_{ilm}g^{(2)}_{njm}g^{(3)}_{lnk}, \tag{1.3b}$$

$$g^{(1)}_{ilm}g^{(2)}_{mjn}g^{(3)}_{nlk}, \tag{1.3c}$$

$$g^{(1)}_{ilm}g^{(2)}_{njm}g^{(3)}_{nlk}. \tag{1.3d}$$

We note that there are other possible definitions of the 3-product besides (1.3). If we wish to keep i, j, and k in the appropriate places for an identity to exist these can only correspond to interchanging the suffices in (1.3), so do not need separate consideration.

It is possible to interpret the elements of the 3-group as "vertex functions" with three external legs denoting the three possible labels i, j, and k in the same way that a matrix can be represented by a two point function, as in Fig. 1. The matrix product now becomes the Feynman diagram with one internal line, whilst the 3-group products (1.3) can be represented by the triangle diagram, as shown in Fig. 2. We can see immediately from this graphical approach that the 3-product (1.3) is nonassociative, as seen by the differences between $[g^{(1)}g^{(2)}(g^{(3)}g^{(4)}g^{(5)})]$ and $[(g^{(1)}g^{(2)}g^{(3)})g^{(4)}g^{(5)}]$ in Fig. 3. This may cause difficulties in applications to particle physics, though it may alternatively be important in algebraic confinement, as has been suggested by Gürsey and others.[5] We can also reduce the problem of nonassociativity by working with infinitesimal elements.

We may consider the quantities g_{ijk} as the 3-group analogues of elements of GL(n,R), and so expect to need sub-3 groups, the analogues of SO(n) or SU(n), which will preserve quadratic scalars. These latter must also be constructed to be positive definite in order that they have physical import. Given such constructs we would be ready to analyze detailed physical applications of these results. For example we could determine if there are suitable groupings of particles to fill irreps of suitable 3- (or higher) groups. We would then attempt to gauge such n-groups, as was remarked earlier.

We start our analysis in the next section by considering the problem of the existence of the identity defined by (1.3). In order for this to exist for the generalized 3-group with elements $g_{i_1,\cdots i_s}$, for $1 \leqslant i_r \leqslant N$, we find we need $N = 3$, and then

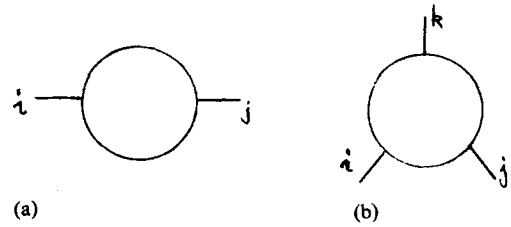


FIG. 1. (a) A graphical representation of the matrix g_{ij} as a propagator; (b) A graphical representation of the 3-group element g_{ijk} as a vertex function.

 0022-2488/80/122715-04$1.00

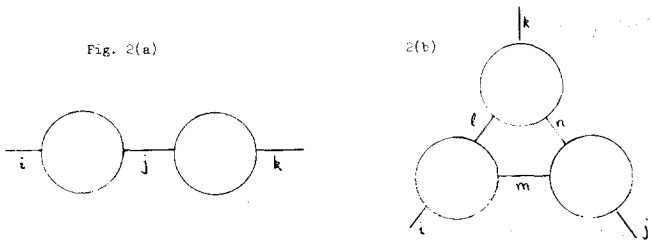


FIG. 2. (a) A graphical representation of the matrix product $g^{(1)}_{ij}\, g^{(2)}_{jk}$. (b) A graphical representation of the 3-products $(g^{(1)}g^{(2)}g^{(3)})$ of (1.3), the orders of the labels differing according to the choices (1.3a) to (1.3d).

only for the product (1.3d). Furthermore this identity is only an infinitesimal one, in the sense made more precise in Sec. 2. In the following section we consider infinitesimal symmetry operations on the n-group (the analog of the adjoint representation) and construct a quadratic invariant; the question of representations is also discussed in this section. In Sec. 4 we attempt to make the global symmetry into a local one, but find that this is not possible. Possible directions for further analysis and applications of n-groups are discussed in the last section.

2. EXISTENCE OF AN IDENTITY

Our analysis will be carried out in this paper only for the concrete case of the three-dimensional array of real numbers g_{ijk} with 3-product (1.3). While this is a severe limitation we have not been able to develop detailed results for any other case, though some of our restrictions will be expressed in a form independent of the 3-product actually chosen.

It does not appear possible to construct an identity for the labels i, j, and k taking more than three values. For the only nontrivial numerical 3-index quantities available are the Kronecker and permutant symbols ϵ_{ijk}, δ_{ijk} defined to be $(-1)^{\rho}$ and 1 when ijk is a permutation of signature ρ of 1,2,3, and zero otherwise; these are only defined if $1 \leqslant i, j, k \leqslant 3$.

Let us construct the identity e_{ijk} as a linear combination of ϵ_{ijk} and δ_{ijk},

$$e_{ijk} = a\epsilon_{ijk} + b\delta_{ijk}. \tag{2.1}$$

We will attempt to choose a and b so that (1.1) is valid under one or other of the product rules (1.3). Let us consider first the 3-product rule (1.3a). Then (1.1) becomes

$$[a^2(\delta_{ij}\delta_{ln} - \delta_{in}\delta_{lj}) + ab\,(\epsilon_{ilm}\delta_{mjn} + \epsilon_{mjn}\delta_{ilm}) + b^2\delta_{mil}\delta_{mjn}]\, g_{lnk}, \tag{2.2a}$$

and so we require the square bracket in (2.2) to be proportional to $\delta_{il}\delta_{jn}$. This is impossible, since, for example, when $i = l \neq j = n$ the bracket in (2.2) vanishes. A similar situation arises for the 3-product (1.3b). For (1.3c) the bracket in (2.2) must be proportional to $\delta_{in}\delta_{jl}$, whilst for (1.3d) the bracket becomes

$$[-a^2(\delta_{ij}\delta_{ln} - \delta_{in}\delta_{jl}) + ab\,(\epsilon_{ilm}\delta_{njm} + \delta_{ilm}\epsilon_{njm}) + b^2\delta_{mil}\delta_{mjn}], \tag{2.2b}$$

and must again be proportional to $\delta_{in}\delta_{jl}$. But this again cannot in general be satisfied for any nonzero b. The only way to satisfy (1.1) appears to be to take $a = 1$, $b = 0$, and also require

$$g_{iij} = g_{iji} = g_{jii} = 0. \tag{2.3}$$

For the choice (2.3) we have $e_{ijk} = \epsilon_{ijk}$.

The 3-group product (1.3d) does not preserve (2.3). We will therefore restrict our discussion to infinitesimal elements of form $(e + g)$, for g satisfying (2.3), and with products approximated as

$$((e + g_1)(e + g_2)(e + g_3)) \simeq e + (g_1 + g_2 + g_3), \tag{2.4}$$

where $(g_1 + g_2 + g_3)$ also satisfies (2.3). The set of such infinitesimals will be all that is required for our further discussion of symmetry transformations. We say that e acts as an infinitesimal identity.

We conclude that there is a unique 3-product, (1.3d), for which there exists the infinitesimal identity ϵ_{ijk}. Furthermore, this exists only for this particular 3-group. We now need to determine if it is possible to use this infinitesimal identity to construct symmetry transformations.

3. GENERALIZED SYMMETRIES

For any elements U, g of the 3-group we can define the symmetry transformation of g by U in the usual manner,

$$g \rightarrow (U\, g\, U^{-1}). \tag{3.1}$$

This can be analyzed infinitesimally, for $U = (e + T)$, $U^{-1} = (e - T)$, with T satisfying (2.3), as

$$\delta g = (U\, g\, U^{-1}) - g = (Tge) - (egT). \tag{3.2}$$

The rhs of (3.2) thus plays the role for our 3-group of the commutator for a 2-group. We note that in order that $(e - T)$ is the inverse of $(e + T)$ to 1st order in T, it is only necessary that e act as an infinitesimal inverse, in the manner we discussed in the previous section.

We now consider the quadratic expression (denoted by Tr)

$$\mathrm{Tr}\, g^2 = g_{ijk}\, g_{ijk}. \tag{3.3}$$

The variation of $\mathrm{Tr}\, g^2$ is given by (3.2) and (1.3d) as

$$\delta\, \mathrm{Tr}\, g^2 = 2g_{ijk}\, [T_{ilm}g_{njm}e_{nlk} - e_{ilm}g_{njm}T_{nlk}] = 0.$$

Thus (3.3) is a positive definite quadratic, invariant under (3.2).

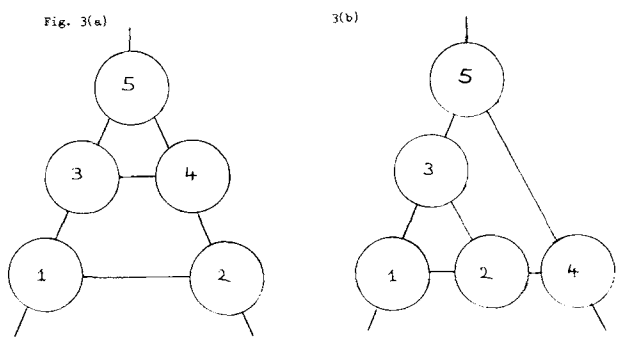


FIG. 3. (a) A graphical representation of the 5-product $(g^{(1)}g^{(2)}(g^{(3)}g^{(4)}g^{(5)}))$. (b) A graphical representation of the 5-product $((g^{(1)}g^{(2)}g^{(3)})\, g^{(4)}g^{(5)})$.

We must also consider the problem of defining the equivalent of the fundamental representation for a 2-group. We see that if $\psi_\alpha(x)$ is introduced as a complex-valued field, its transformation under U, obtained by saturating indices, generates a companion 2-index field $\psi_{\alpha\beta}(x)$. Thus we take $\{\psi_\alpha(x),\psi_{\alpha\beta}(x)\} = \Psi(x)$ as defining the analog of the fundamental representation of a 2-group, with

$$\delta\psi_\alpha = iT_{\alpha\beta\gamma}\psi_{\beta\gamma}, \quad \delta\psi_{\alpha\beta} = iT_{\alpha\beta\gamma}\psi_\gamma . \tag{3.4}$$

We can construct the positive definite quadratic form

$$\Psi^*\Psi = \psi^*_\alpha\psi_\alpha + \psi^*_{\alpha\beta}\psi_{\alpha\beta} , \tag{3.5}$$

and find that under (3.4)

$$\delta(\Psi^*\Psi) = i\psi^*_{\alpha\beta}\psi_\gamma(T_{\alpha\beta\gamma} - T_{\gamma\alpha\beta}) + i\psi^*_\alpha\psi_{\beta\gamma}(T_{\alpha\beta\gamma} - T_{\beta\gamma\alpha}).$$

This is zero provided $T_{\alpha\beta\gamma}$ is invariant under cyclic permutations,

$$T_{\alpha\beta\gamma} = +T_{\beta\gamma\alpha} = +T_{\gamma\alpha\beta} . \tag{3.6}$$

The factor i in (3.4) is essential; without it (3.6) becomes $T_{\alpha\beta\gamma} = -T_{\beta\gamma\alpha} = -T_{\gamma\alpha\beta}$, whose only solution is $T_{\alpha\beta\gamma} \equiv 0$. Thus we are forced into a complex fundamental representation.

The set of elements $T_{\alpha\beta\gamma}$ satisfying (3.6) and the trace condition (2.3) is a seven real-parameter subset of the 3-group. Furthermore it acts on the 12 complex-dimensional space Ψ by (3.4), and on the 27 real-dimensional space of g_{ijk} by (3.2). We may use the space of Ψ's to describe Dirac spinors, and can write down a Lagrangian $\bar{\Psi}\partial\Psi$ invariant under the global transformations (3.4). We have thus constructed the beginnings of a suitable global symmetry theory for elementary particles based on a 3-group. We propose to discuss detailed applications of this elsewhere.

4. GAUGING THE 3-GROUP

In view of the recent great successes of gauge theories based on the 2-group $\mathrm{Su}(3)_c \times \mathrm{SU}(2) \times \mathrm{U}(1)$ it is natural to determine if we can allow the seven-parameter infinitesimal set $T_{\alpha\beta\gamma}$ satisfying (2.3) and (3.6) to be space–time dependent. To achieve this we would require the presence of a gauge vector field $A_\mu(x)$, with values in the 3-group, and transforming under the local version of (3.2) as

$$\delta A_\mu = (TA_\mu e) - (eA_\mu T) + \partial_\mu T. \tag{4.1}$$

We will attempt to construct a field strength $F_{\mu\nu}$ transforming covariantly as (3.2), so that $-\mathrm{Tr}(F_{\mu\nu}F^{\mu\nu})$ defined by (3.3) will be a satisfactory Lagrangian. Let us consider

$$\delta\partial_{[\mu}A_{\nu]} = (T\partial_{[\mu}A_{\nu]}e) - (e\partial_{[\mu}A_{\nu]}T) + (\partial_{[\mu}TA_{\nu]}e) - (eA_{[\nu}\partial_{\mu]}T). \tag{4.2}$$

We wish to add to $\partial_{[\mu}A_{\nu]}$ the analog of the commutator bracket $[A_\mu, A_\nu]$ for 2-groups. This analog could be chosen as $(A_\mu eA_\nu) - (A_\nu eA_\mu)$, but its variation under the last term of (4.1) gives $(\partial_{[\mu}TeA_{\nu]}) - (A_{[\nu}e\partial_{\mu]}T)$, which does not cancel with the last two terms of (4.2) unless there are the identities

$$(TeA_\mu) = (TA_\mu e) \tag{4.3a}$$

and

$$(eA_\mu T) = (A_\mu eT). \tag{4.3b}$$

Using (1.3d), (4.3) becomes in components

$$T_{ilm}e_{njm}A_{\mu_{nlk}} = T_{ilm}A_{\mu_{njm}}e_{nlk} \tag{4.4a}$$

and

$$e_{ilm}A_{\mu_{njm}}T_{nlk} = A_{\mu_{ilm}}e_{njm}T_{nlk} . \tag{4.4b}$$

The solution of (4.4a) for A_μ is given by the trivial solution

$$A_{\mu_{nlk}} = e_{nlk}A_\mu , \tag{4.5}$$

where A_μ is a single vector field. This choice of generalization of the commutator bracket is therefore unsatisfactory. But this is also true of the other choices, being $(A_\mu A_\nu e) - (A_\nu A_\mu e)$, $(eA_\mu A_\nu) - (eA_\nu A_\mu)$, or their linear combinations. Similar conditions to (4.3) arise, such as

$$(A_\mu Te) = (eA_\mu T), \tag{4.6}$$

which again can only be satisfied by (4.5). We therefore conclude that it is not possible to obtain a satisfactory local theory under the gauge transformation (4.1). The above difficulty is absent for the modified transformation

$$\delta A_\mu = (TeA_\mu) - (A_\mu eT) + \partial_\mu T, \tag{4.7}$$

for which

$$\delta\partial_{[\mu}A_{\nu]} = (Te\partial_{[\mu}A_{\nu]}) - (\partial_{[\mu}A_{\nu]}eT) + (\partial_{[\mu}TeA_{\nu]}) - (A_{[\nu}e\partial_{\mu]}T). \tag{4.8}$$

The last two terms in (4.8) now agree with those arising in $(A_{[\mu}eA_{\nu]})$. Thus if we define

$$F_{\mu\nu} = \partial_{[\mu}A_{\nu]} - (A_{[\mu}eA_{\nu]}), \tag{4.9}$$

then under (4.7) $F_{\mu\nu}$ will transform without the inhomogeneous term,

$$\delta F_{\mu\nu} = (TeF_{\mu\nu}) - (F_{\mu\nu}eT), \tag{4.10}$$

provided we have the identity

$$(Te(A_\mu eA_\nu)) - ((TeA_\mu)eA_\nu) + (A_\mu e(A_\nu eT)) - ((A_\mu eA_\nu)eT) = (A_\mu e(TeA_\nu)) - ((A_\mu eT)eA_\mu). \tag{4.11}$$

Yet again, by inspection it only seems possible to satisfy (4.11) by the trivial solution (4.5). We conclude that there is a no-go theorem for gauging 3-groups.

5. DISCUSSION

We have only presented here the bare preliminaries of the framework for n-groups and their applications in particle symmetries. We have found that a sensible framework can be constructed when the transformations are global and belong to a particular 3-group. Undoubtedly similar results should be possible for higher generalized groups, and the detailed construction of such cases would be of interest. There are also many related questions as to the definition and nature of higher representations and the construction of alternate invariants. Indeed one might attempt to extend all of the standard technology for Lie groups and their associated algebras to the generalized group setting; our results (together with those in Ref. 4) can be regarded as a preliminary step in that direction.

There are also numerous questions as to the possible applications of these ideas to elementary particles. Can any

clue be detected as to the existence of a 3-group or higher global symmetry in the particle mass spectrum? This would seem a difficult question to answer until a reasonable analog of the representation theory of Lie algebras has been developed.

One of the purposes of this paper has been constructive: to point out the possible generalization of the idea of a particle symmetry, and to sketch its possible nature. However the other purpose is also constructive, but involves the no–go theorem of Sec. 4. If it is not possible to sensibly gauge the 3-group (and, by implication, higher generalized groups) then nature would not have used these objects to describe the fundamental forces. Such a result supports the recent successes of electroweak and color gauge theories, and indicates that there may well be few alternatives to them. At least the alternative discussed here does not seem viable.

[1]R.H. Bruck, *A Survey of Binary Systems* (Springer, Berlin, 1958).
[2]P.M. Cohn, *Universal Algebra* (Harper and Row, New York, 1965).
[3]J.G. Taylor, J. Math. Phys. **6**, 1148 (1965).
[4]J.G. Taylor, *Bootstraps, Fields and Generalised Groups, Lectures in Theoretical Physics XA*, edited by A.O. Barut and W.E. Brittin (Gordon and Breach, New York, 1968).
[5]M. Günaydin and F. Gürsey, Phys. Rev. **9**, 3387 (1974); Contributions in the Second Workshop on "Current Problems in High Energy Particle Theory," edited by G. Domokos and S. Kövesi-Domokos, John Hopkins University, Baltimore, Maryland,1978 (unpublished).

Group actions on principal bundles and invariance conditions for gauge fields [a)]

J. Harnad
Centre de Recherches de Mathématiques Appliqués–Université de Montréal, Canada

S. Shnider
Department of Mathematics, McGill University, Montreal, Canada

Luc Vinet
Center de Recherches de Mathématiques Appliqués–Université de Montréal, Canada

(Received 20 March 1980; accepted for publication 20 June 1980)

Invariance conditions for gauge fields under smooth group actions are interpreted in terms of invariant connections on principal bundles. A classification of group actions on bundles as automorphisms projecting to an action on a base manifold with a sufficiently regular orbit structure is given in terms of group homorphisms and a generalization of Wang's theorem classifying invariant connections is derived. Illustrative examples on compactified Minkowski space are given.

In the study of gauge field equations at the classical level a standard method of simplification involves the requirement that the fields be invariant under a group of space-time transformations.[1] Such a requirement leads to a reduction in the dimension of the free variables and a reduction of the gauge freedom to those changes of gauge which preserve the invariance condition. The specification of how the transformation group acts on the fields may involve an auxiliary gauge transformation. In local terms this gauge transformation will be determined by a function which we shall call a transformation function, depending on the group element and the space-time point and subject to an appropriate composition law. A change in gauge changes the local expression for the transformation function to an equivalent one. Since the form of the transformation function determines the form of the invariance equations and thus affects the difficulty in finding the invariant fields it is useful to have a reduction procedure for simplifying the invariance equations. An associated problem is determining all inequivalent transformation functions for a given transformation group. In this paper we study these problems and show how to find the most general gauge fields possessing a given symmetry using the language and methods of fiber bundle theory. Forgács and Manton[2] have studied the same problem from another point of view. For further applications to problems in symmetry breaking and dimensional reduction see Refs. 3–6.

Since a change of gauge can be interpreted as a change of fiber coordinates in a fiber bundle, our first step will be to formulate the problem in coordinate free language. So expressed, the problem of determining all inequivalent transformation functions is seen to be essentially the same as determining all inequivalent lifts of the transformation group action from the base to automorphisms on the bundle. For a homogeneous space, a known result[7] reduces the problem to a classification of group homomorphisms. For the general case, no result is known, however, provided the orbit structure is regular enough we can solve the problem under the additional hypothesis that the gauge group is compact. The gauge fields determine a connection on the bundle and the symmetry problem is equivalent to the classification of G-invariant connections. Again, for a homogeneous G space the solution is standard and may be extended to certain more general cases.

1. BASIC RESULTS FOR HOMOGENEOUS SPACES

Let H be the gauge group with Lie algebra $\mathcal{H}$, M a differentiable manifold, and G a Lie transformation group acting on M such that the map

$$G \times M \to M\,(g,x) \to f_g(x)$$

is differentiable and satisfies

$$f_e(x) = x, \quad f_{g_1}(f_{g_2}(x)) = f_{g_1 g_2}(x). \tag{1}$$

When no confusion can arise we shall write gx for $f_g(x)$.

The gauge fields which we consider are defined on an open covering $\{U_\alpha\}$ of M by a set of $\mathcal{H}$ valued 1-forms ω_α on U_α related by

$$\omega_\beta = \mathrm{Ad}k_{\alpha\beta}^{-1}\omega_\alpha + k_{\alpha\beta}^{-1}dk_{\alpha\beta},$$

where the functions $k_{\alpha\beta}: U_\alpha \cap U_\beta \to H$ satisfy $k_{\alpha\alpha} \equiv e$, $k_{\alpha\beta}k_{\beta\gamma} = k_{\alpha\gamma}$ on $U_\alpha \cap U_\beta \cap U_\gamma$. The $k_{\alpha\beta}$ are transition functions for a principal H bundle E over M trivial over each U_α, that is, there are functions

$$\tau_\alpha: U_\alpha \times H \to E,$$

with $\tau_\beta^{-1}\tau_\alpha: U_\alpha \cap U_\beta \times H \to U_\alpha \cap U_\beta \times H$, such that

$$\begin{aligned}\tau_\beta^{-1}\tau_\alpha(x,h) &= (x, k_{\alpha\beta}(x)^{-1}h) \\ &= (x, k_{\beta\alpha}(x)h).\end{aligned} \tag{2}$$

The right action of the gauge group H on E is given by

$$R_k\tau_\alpha(x,h) = \tau_\alpha(x,hk), \quad \text{for } x \in M; \quad h,k \in H. \tag{3}$$

Define a local section σ_α by

$$\sigma_\alpha(x) = \tau_\alpha(x,e).$$

When there is no possibility of confusion we write $\sigma_\alpha(x)h$ for $R_h\sigma_\alpha(x)$. The form ω_α is the pull-back under σ_α of a connec-

[a)]Research supported in part by the National Sciences and Engineering Research Council of Canada.

 0022-2488/80/122719-06$1.00

tion form ω on E. The pull-back of ω under τ_α is given by

$$(\tau_\alpha^*\omega)_{(x,h)} = \mathrm{Ad}h^{-1}(\omega_\alpha)_x + h^{-1}\,dh, \tag{4}$$

which in fact defines ω.

If the open sets U_α are G invariant the condition for G invariance of the ω_α up to gauge transformation is

$$(f_g^*\omega_\alpha)_x = \mathrm{Ad}\rho_\alpha(g,x)^{-1}(\omega_\alpha)_x + \rho_\alpha^{-1}(g,x)\,d\rho_\alpha(g,x), \tag{5}$$

where the differential in ρ_α is in the x variable. The function ρ_α is what we call a transformation function. The $\rho_\alpha: G\times U\to H$ satisfy

$$\rho_\alpha(g_1g_2,x) = \rho_\alpha(g_2,x)\,\rho_\alpha(g_1,g_2x) \tag{6}$$

in order to satisfy the group composition law (1) and the compatibility condition, and

$$\rho_\alpha(g,x)k_{\alpha\beta}(gx) = k_{\alpha\beta}(x)\,\rho_\beta(g,x) \tag{7}$$

for the consistency of (5) under change of section. The functions ρ_α define a G action on E

$$G\times E\to E \quad (g,\tau_\alpha(x,h))\to\tilde f_g\tau_\alpha(x,h) = \tau_\alpha(gx,\ \rho_\alpha(g,x)^{-1}h). \tag{8}$$

[This is a valid G-action on E by virtue of (6) and independent of the local trivialization τ_α by virtue of (7).] Again writing $g\sigma_\alpha(x)$ for $\tilde f_g(\sigma_\alpha(x))$,

$$g\sigma_\alpha(x) = \sigma_\alpha(gx)\,\rho_\alpha(g,x)^{-1}. \tag{9}$$

The invariance condition (5) implies that the connection defined in (4) satisfies

$$\tilde f_g^*\omega = \omega. \tag{10}$$

This is the coordinate-free form of the invariance condition which we shall study.

Before proceeding, note that if the open sets U_α over which E is trivial cannot be chosen so that they are G invariant, then given $x\in U_\alpha$, we must restrict the $g\in G$ appearing in Eq. (5) to those for which $gx\in U_\alpha$. Alternatively we can find an infinitesimal invariance equation which can be expressed in local coordinates as follows.

Let $V(M)$ be the smooth vector fields on M. Denoting by $\mathcal{G}$ the left invariant vector fields on G (identified with the Lie algebra) define mappings:

$$\varphi:\mathcal{G}\to V(M) \quad \varphi(\xi)_x = \left.\frac{d}{dt}\right|_0 \exp(-t\xi)x$$

and

$$r_\alpha:\mathcal{G}\times M\to\mathcal{h} \quad r_\alpha(\xi,x) = -\left.\frac{d}{dt}\right|_0 \rho_\alpha(\exp t\xi,x).$$

The invariance equation in infinitesimal form becomes

$$\mathcal{L}_{\varphi(\xi)}\omega_\alpha = [r_\alpha(\xi,x),\omega_\alpha] - dr_\alpha(\xi,x), \tag{11}$$

where the left hand side denotes the Lie derivative and the differential on the right is in the x variable.

The function r_α satisfies the composition law

$$r_\alpha([\xi,\eta],x) = [r_\alpha(\xi,x),r_\alpha(\eta,x)] + \varphi(\xi)_x r_\alpha(\eta,x) - \varphi(\eta)_x r_\alpha(\xi,x) \tag{12}$$

and the compatibility condition

$$r_{\alpha\beta}(\xi,x) = \mathrm{Ad}k_{\alpha\beta}(x)^{-1}r_\beta(\xi,x) + k_{\alpha\beta}^{-1}\,dk_{\alpha\beta}. \tag{13}$$

The interpretation of the infinitesimal invariance condition on the bundle level is as follows.[8] Let

$$\Phi(\xi)_{\tau_\alpha(x,h)} = \tau_{\alpha*}(\varphi(\xi) + \mathrm{Ad}h^{-1}r_\alpha(\xi,x)). \tag{14}$$

Equation (13) guarantees that this defines unambiguously a vector field on E and Eq. (12) implies that $\Phi:\mathcal{G}\to V(M)$ is an algebra homomorphism

$$\Phi([\xi,\eta]) = [\Phi(\xi),\Phi(\eta)].$$

One checks that (11) is equivalent to

$$\mathcal{L}_{\Phi(\xi)}\omega = 0. \tag{15}$$

This infinitesimal form seems more general since it does not assume the existence of a group action in finite (integrated) form. However, if the infinitesimal action on M integrates and if the gauge group is compact the infinitesimal action on E given by Φ integrates.

We can now formulate the problem in terms of fiber bundles as the determination of all principal H bundles with G action (as automorphisms) projecting to the given action on M and all invariant connections on such bundles. However the question posed in this form is too general since it involves the topological problem of classifying all H bundles over M. We restrict attention to the structure of the bundle over a neighborhood of an orbit in M and begin with the structure of E over a single orbit.

For $x\in M$ let G_x be the isotropy group at x and let $G(x)$ be the orbit through x. Assume the orbit is an imbedded submanifold of M then G/G_x is diffeomorphic to $G(x)$ and the structure of E over $G(x)$ is determined by (see e.g. Ref. 7).

Proposition 1: There is a one-to-one correspondence between

(a) Equivalence class of principal H bundles E over G/G_x admitting a G action which projects to left multiplication of G on G/G_x; and

(b) Conjugacy classes of homomorphisms $\lambda:G_x\to H$.

Equivalence in (a) means an isomorphism of bundles which commutes with the action of G and projects to the identity mapping.

We shall sketch a proof in order to clarify the result and establish notations.

Proof: Given a bundle E from one of the equivalence classes in (a) any $g\in G_x$ maps the fiber E_x over $x = eG_x$ into itself. If we pick a point $p\in E_x$ we have

$$gp = p\lambda(g),$$

where $\lambda:G_x\to H$. One sees immediately that λ is a homomorphism since the G and H actions commute and that if p is right translated by h then λ is conjugated by h. Also if $\varphi:E\to E'$ is a G equivariant bundle isomorphism so that E and E' are equivalent, the points p and $\varphi(p)$ determine the same homomorphism λ.

Conversely given $\lambda:G_x\to H$ we can construct a principal H bundle E_λ over G/G_x. On the set $G\times H$ define an equivalence relation

$$(g,h)\sim(gg_1,\lambda(g_1)^{-1}h), \quad \text{for } g_1\in G_x.$$

Let $[g,h]$ be the equivalence class of (g,h) and let E_λ be the set of equivalence classes. Another notation often used for E_λ is $G\times_{G_x}H$. Projection on the first factor $G\times H\to G$ defines a

projection

$$\pi: G \times_{G_x} H \to G/G_x.$$

The left action of G and right action of H defined by

$$(g_1,(g,h)) \to (g_1 g,h),$$

$$((g,h),h_1) \to (g,hh_1)$$

preserve the equivalence relation and so define group actions of G and H on E_λ. The action of G on E_λ projects by π to left multiplication on the coset space G/G_x. The right action of H is transitive on the fibers of π. To verify the bundle structure, let $U \subset G/G_x$ be an open set on which there is a cross-section $\sigma: U \to G$ of $G \to G/G_x$. Then we can define a cross-section of E_λ over U by

$$y \to [\sigma(y),e]$$

and a corresponding local trivialization

$$(y,h) \to [\sigma(y),h].$$

Since $G \to G/G_x$ itself has a bundle structure, there exists a covering of G/G_x by such open sets U. Having shown how to go from (a) to (b) and (b) to (a), we show that the composite in either order gives back the same equivalence class. If we pick the point $[e,e]$ in the fiber of E_λ over $x = eG_x$ we have for $g \in G_x$

$$g[e,e] = [g,e] = [e,\lambda(g)] = [e,e]\lambda(g).$$

Thus we recover the homomorphism λ from the bundle E_λ. Finally if E is a bundle and for $p \in E_x$ the associated homomorphism is λ, we define a G equivalent isomorphism:

$$E_\lambda = G \times_{G_x} H \to E,$$

$$[g,h] \to gph.$$

In local terms we can use this result to show how the transformation function depends on the homomorphism λ and the section σ of $G \to G/G_x$,

$$\begin{aligned} g[\sigma(y),e] &= [g\sigma(y),e] = [\sigma(gy)\sigma(gy)^{-1}g\sigma(y),e] \\ &= [\sigma(gy),\lambda(\sigma(gy)^{-1}g\sigma(y))] \\ &= [\sigma(gy),e]\lambda(\sigma(gy)^{-1}g\sigma(y)). \end{aligned}$$

Thus

$$\rho^{-1}(g,y) = \lambda(\sigma(gy)^{-1}g\sigma(y)) \tag{16}$$

if we use the section of E_λ

$$y \to [\sigma(y),e].$$

For a given homomorphism λ, the bundle E_λ need not be trivial and therefore the transformation function may not be defined throughout the orbit. The case when it *can* be is given by:

Corollary 1: The bundle E is trivial over G/G_x if and only if the homomorphism $\lambda: G_x \to H$ extends to a smooth function $\Lambda: G \to H$ such that

$$\Lambda(gg_1) = \Lambda(g)\lambda(g_1), \quad \text{for } g \in G, \quad g_1 \in G_x.$$

Proof: If σ is a section of E defined over all of G/G_x, define $\Lambda(g)$ by $g\sigma(x) = \sigma(gx)\Lambda(g)$. Conversely given Λ satisfying the hypotheses, $\sigma: gG_x \to [g,\Lambda^{-1}(g)]$ defines a section of the bundle $G \times_{G_x} H$ over G/G_x.

The section σ satisfies

$$\begin{aligned} g_1\sigma(gG_x) &= [g_1 g,\Lambda^{-1}(g)] \\ &= [g_1 g,\Lambda^{-1}(g_1 g)]\Lambda(g_1 g)\Lambda^{-1}(g) \end{aligned}$$

for $g_1, g \in G$ and so the associated transformation function is

$$\rho(g_1,gG_x) = \Lambda(g)\Lambda(g_1 g)^{-1}, \quad g_1,g \in G. \tag{17}$$

The condition for ρ to be independent of its second variable, the point in the orbit, is given by the following corollary.

Corollary 2: The following two conditions are equivalent.

(a) The bundle $E \to G/G_x$ is trivial with gauge function $\rho(g_1,gG_x)$ independent of the point gG_x.

(b) The homomorphism $\lambda: G_x \to H$ extends smoothly to a homomorphism $\Lambda: G \to H$.

Proof: Equation (17) shows that $\rho(g_1,gG_x) = \rho(g_1,G_x) = \Lambda(g_1)^{-1}$ if and only if Λ is a homomorphism.

The simplest transformation function is just the identity, the criterion for which is the following.

Corollary 3: The transformation function $\rho(g_1,gG_x)$ reduces to the trivial function $\equiv e$ if and only if it is trivial when restricted to the isotropy group G_x. That is, the image of λ in H is e.

One case in which this always occurs is when the G-action on M is free, i.e., $G_x = e$.

We continue with the discussion of G-invariant connections on $E \to G/G_x$. We shall give a proof of the theorem of Wang[9] classifying these connections, in which we make use of the bundle $E_\lambda = G \times_{G_x} H$.

Proposition 2: Let $\mathcal{G}$ be the Lie algebra of G, $\mathcal{G}_0$ the Lie algebra of $G_0 \subset G$ and $\mathcal{H}$ the Lie algebra of H. The G invariant connections on the bundle E_λ determined by $\lambda: G_0 \to H$ are in one to one correspondence with linear mappings $A: \mathcal{G} \to \mathcal{H}$ satisfying the following two equations:

$$A(\xi) = \lambda_*(\xi), \quad \text{for } \xi \in \mathcal{G}_0 \quad \text{and } \lambda_*, \tag{18a}$$

the homomorphism $\lambda_*: \mathcal{G}_0 \to \mathcal{H}$ determined by the differential of λ.

$$A(\mathrm{Ad}g^{-1}\xi) = \mathrm{Ad}\lambda(g)^{-1}(A(\xi)), \quad \text{for } \xi \in \mathcal{G} \text{ and } g \in G_0. \tag{18b}$$

Proof: Let ω be a G-invariant connection on $G \times_{G_0} H$, let $\psi: G \times H \to G \times_{G_0} H$ be defined by $\psi(g,h) = [g,h]$ and let $j: G \to G \times H$ be $j(g) = (g,e)$. Then $\psi^*\omega$ is a G-invariant connection on the trivial H bundle $G \times H$ and $j^*\psi^*\omega$, its pull-back to the base space G, is a left G-invariant $\mathcal{H}$ valued form and thus is determined by its value at $T_e G$ which can be identified with $\mathcal{G}$. We conclude that if $\theta_{\mathcal{G}}$ is the left-invariant Maurer–Cartan form on G then there is a linear map $A: \mathcal{G} \to \mathcal{H}$ such that

$$j^*\psi^*\omega = A \circ \theta_{\mathcal{G}}.$$

Let $\psi^*\omega = \omega_1 + \omega_2$, where ω_1 acts on the tangents to the first factor and ω_2 on the tangents to the second factor in $G \times H$. If $\eta \in \mathcal{H}$ and $\tilde{\eta}$ is the vertical vector field on $G \times_{G_0} H$ generated by $R_{\exp t\eta}$, then $\omega(\tilde{\eta}) = \eta$ which implies ω_2 is the Maurer–Cartan form on $\mathcal{H}$, $\theta_{\mathcal{H}}$. From the equivariance condition

$$R_h^* \omega = \mathrm{Ad}h^{-1}\omega$$

we conclude

$$\psi^*\omega_{(g,h)} = \mathrm{Ad}h^{-1}(A \circ \theta_{\mathcal{G}}) + \theta_{\mathcal{H}}.$$

Proof: The argument is very close to that in the proof of Proposition 2 so we will omit most of the details. As in that proof we define $\psi: G\times H\times S\to G\times_{G_0} H\times S$ by $\psi(g,h,s)=([g,h],s)$ and find

$$\psi^*\omega(g,h,s)=\mathrm{Ad}h^{-1}(A_s\circ\theta_{\mathcal{G}}+\mu)+\theta_{\mathcal{H}},$$

where μ is a one-form on S. Left G invariance shows that there is no "g dependence" in the form μ. The conditions that the right-hand side define the pull-back of a form on $G\times_{G_0}H\times S$ impose in addition to Eqs. (18a) and (18b) on the linear mappings A_s the additional equation,

$$\mu=\mathrm{Ad}\lambda(g)^{-1}\mu,\quad \text{for } g\in G_0.$$

Thus μ must take values in the subalgebra of $\mathcal{H}$ of elements invariant under the adjoint action of $\lambda(G_0)$.

3. EXAMPLES

We now illustrate these results with some examples. let M_0 be compactified Minkowski space which we identify[1] with U(2), let $M=\mathrm{SU}(2)\times\mathrm{U}(1)$ be the twofold covering and let the gauge group H be SU(2). For the transformation group G also equal to SU(2) consider the following actions of G on M. Given $g\in G=\mathrm{SU}(2)$ and $(x,e^{i\psi})\in M=\mathrm{SU}(2)\times\mathrm{U}(1)$ define

$$\alpha_g(x,e^{i\psi})=(gx,e^{i\psi}),$$
$$\beta_g(x,e^{i\psi})=(xg^{-1},e^{i\psi}),$$
$$\gamma_g(x,e^{i\psi})=(gxg^{-1},e^{i\psi}).$$

Both α and β define simple actions with special cross-sections through $(x,e^{i\psi})$ given by $\varphi(s)=(x,e^{i(\psi+s)})$. The action defined by γ is not simple since there are two orbit types for the conjugation action of SU(2) on itself. Therefore we restrict to the open submanifold $M_1=(\mathrm{SU}(2)-\{\pm I\})\times\mathrm{U}(1)$ on which the action γ is simple, where

$$I=\begin{pmatrix}1&0\\0&1\end{pmatrix}\in\mathrm{SU}(2).$$

On M_1

$$\varphi(s,t)=\left(x\begin{pmatrix}e^{is}&0\\0&e^{-is}\end{pmatrix},\ e^{i(\psi+t)}\right)$$

defines a special cross-section through $(x,e^{i\psi})$.

Since α and β commute there is a well defined action $\alpha\times\beta$ of $\mathrm{SU}(2)\times\mathrm{SU}(2)$ on M which we shall also consider.

Example α: The isotropy group is the identity I and therefore the orbits are identifiable with G. By Corollary 3 the bundle structure over any orbit is trivial. By Theorem 1 the same is true over a neighborhood of an orbit and by Theorem 2 the connection form pulled back to the base space M under any G invariant section is given by

$$\tilde{\omega}=A_\psi\circ\theta_{\mathcal{G}}+B d\psi,$$

where A_ψ is a smoothly parameterized family of linear maps $\mathcal{G}=\mathfrak{su}(2)\to\mathcal{H}=\mathfrak{su}(2)$, B is a smooth $\mathfrak{su}(2)$ valued function of ψ, and the Maurer–Cartan form $\theta_{\mathcal{G}}$ is regarded as defined, on a neighborhood of orbits, on the first term in $M\sim\mathrm{SU}(2)\times U(1)$. The triviality of the bundle in this case may be proved to be global (see Ref. 1).

Example β: This is completely equivalent to the previous example with the left invariant Maurer–Cartan form $\theta_{\mathcal{G}}$ replaced by the right invariant Maurer–Cartan form in the expression for $\tilde{\omega}$.

(For the above two examples, the gauge group SU(2) may be replaced by arbitrary H with algebra $\mathcal{H}$, with A_ψ interpreted as any smooth family of linear maps $A_\psi:\mathfrak{su}(2)\to\mathcal{H}$.

Example $\alpha\times\beta$: The transformation group G is $\mathrm{SU}(2)\times\mathrm{SU}(2)$ and along the cross-section $\varphi(s)=(I,e^{is})$ the isotropy group is the diagonal subgroup
$\Delta=\{(g,g)|g\in\mathrm{SU}(2)\}\subset G$
$=\mathrm{SU}(2)\times\mathrm{SU}(2)$. Up to conjugacy in $H=\mathrm{SU}(2)$ there are two homomorphism $\lambda:\Delta\to H$

$$\lambda_0(g,g)\equiv I \text{ and } \lambda_1(g,g)=g.$$

These both extend to homomorphisms of $G\to H$ by choosing the extension independent of the second factor; therefore, by Corollary 2, there exists a section of E_{λ_0} and E_{λ_1} over the entire orbit. The bundle E_{λ_0} is defined by the equivalence relation

$$(g_1,g_2,h)\sim(g_1g_3^{-1},g_2g_3^{-1},h)$$

and E_{λ_1} is defined by the equivalence

$$(g_1,g_2,h)\sim(g_1g_3^{-1},g_2g_3^{-1},g_3h).$$

The group action on both is given by

$$\alpha\times\beta_{(g_1',g_2')}[g_1,g_2,h]=[g_1'g_1,g_2'g_2,h],$$

where $[\ ]$ denotes an equivalence class. We can identify the orbits with G/Δ and G/Δ can be identified with SU(2) by

$$x\mapsto(x,I)\Delta\in G/\Delta,\quad \text{for } x\in\mathrm{SU}(2).$$

Define sections σ_0 and E_{λ_0} and σ_1 of E_{λ_1} over an orbit by

$$\sigma_0(x)=[x,I,I] \text{ and } \sigma_1(x)=[x,I,I].$$

(The different notations used, distinguish between the two definitions of equivalence.) Then

$$\begin{aligned}\alpha\times\beta_{(g_1',g_2')}\sigma_0(x)&=[g_1'x,g_2',I]\\&=[g_1'xg_2'^{-1},I,I]\\&=\sigma_0(g_1'xg_2'^{-1})\end{aligned}$$

and

$$\begin{aligned}\alpha\times\beta_{(g_1',g_2')}\sigma_1(x)&=[g_1'x,g_2',I]\\&=[g_1'xg_2'^{-1},I,g_2']\\&=\sigma_1(g_1'xg_2'^{-1})g_2'.\end{aligned}$$

If we consider $E_{\lambda_0}\times S$ and $E_{\lambda_1}\times S$ and define σ_0 and σ_1 by

$$\sigma_0(x,e^{i\psi})=([x,I,I],e^{i\psi}).$$
$$\sigma_1(x,e^{i\psi})=([x,I,I],e^{i\psi}).$$

We get the same transformation equations. Write the Maurer–Cartan form $\theta_{\mathcal{G}}$ as $\theta_1+\theta_2$ corresponding to the direct product decomposition. An invariant connection on $E_{\lambda_0}\times S$ or $E_{\lambda_1}\times S$ pulled back to $G\times H\times S$ looks like

$$\omega_{(g_1,g_2,h,e^{i\psi})}=\mathrm{Ad}h^{-1}(A_\psi(\theta_1+\theta_2)+B_\psi\,d\psi)+\theta_{\mathcal{H}}$$

subject to compatibility with λ_0 or λ_1. For λ_0 the condition (18b) implies

$$A_\psi(\mathrm{Ad}(g,g)^{-1}(\theta_1+\theta_2))=A_\psi(\theta_1+\theta_2),$$

which implies $A_\psi\equiv 0$. The $\mathfrak{su}(2)$ valued function B_ψ is arbitrary and the connection pulled back to the base by σ_0 is just

The mapping ψ defines a fibration of $G \times H$ over $G \times_{G_0} H$ with G_0 acting on the fibers. The conditions for a form on $G \times H$ to be the pull-back by ψ of a form on $G \times_{G_0} H$ are first that it vanish on tangents to the fibers and second that it be invariant under the action of G_0. The tangents to the fibers are given by differentiating in t the expression $[g \exp(t\xi)], \lambda(\exp(-t\xi))h]$ hence are of the form

$$[\xi, -\mathrm{Ad}h^{-1}\lambda_*(\xi)]_{(g,h)}, \quad \text{for } \xi \in \mathcal{G}.$$

The condition $\psi^*\omega\, \xi, -\mathrm{Ad}h^{-1}\lambda_*(\xi) = 0$ implies

$$A(\xi) = \lambda_*(\xi).$$

The G_0 action is given by

$$g_{1*}(\xi,\eta)_{(g,h)} = (\mathrm{Ad}g_1^{-1}\xi, \eta)_{(gg_1, \lambda(g_1)^{-1}h)}.$$

The invariance of $\psi^*\omega$ implies

$$\mathrm{Ad}\lambda(g)A(\mathrm{Ad}g^{-1}\xi) = A(\xi).$$

Since these conditions are necessary and sufficient the proposition is proved.

2. GENERALIZATION TO INTRASITIVE GROUP ACTIONS

Now we study the situation when the base space M is not a homogeneous space. The structure of the bundle over the orbit through x is determined by a homomorphism $\lambda_x: G_x \to H$. To put together the information over a set of orbits we need a smooth cross section, a submanifold intersecting each orbit in one point. Such a cross section may not exist even locally if the conjugacy class of isotropy group changes from orbit to orbit.

For $x \in M$, let $G(x)$ be the orbit of G through x, if $y \in G(x)$ then $y = gx$ and $G_y = gG_xg^{-1}$; the isotropy groups are conjugate. Associated to each orbit is a unique conjugacy class which we call the type of the orbit. If the action of G on M has just one orbit type one can often show that for all $x \in M$ there is a smooth imbedding of an open set $S \subset R^k$ ($k = \dim M - \dim G/G_0$) into M $\varphi: S \to M$ with $\varphi(0) = x$ and $\varphi(S)$ intersecting each orbit in a unique point, further the isotropy group of all the points $\varphi(S)$ is the same, $G_{\varphi(p)} = G_x$ for all $p \in S$. We call such a situation a simple G action and such an imbedding a special cross section. For a simple G action we can formulate a reasonable theorem without involved technical conditions on the orbit structure. One can best deal with the more complicated cases involving several orbit types individually.

Given an H bundle $\pi: E \to M$ with G action projecting to a simple G action on M, let φ be a special cross section and $\sigma: S \to E$ be a "section of E over φ" that is $\pi\sigma(s) = \varphi(s)$. Define $\lambda_s: G_0 \to H$ by

$$g\sigma(s) = \sigma(s)\lambda_s(g), \quad g \in G_{\varphi(s)} = G_0.$$

Lemma: If G_0 and H are compact the section σ can be chosen so that λ_s is independent of s, equal to its value at $s = 0$.

Proof: Let T_0 be a maximal torus in G_0 and let t be an element such that $\{t^n\}$ is dense in T_0. We shall show that there is a smooth function $h: S' \to H$ $h(0) = e$, with $S' \ni 0$ open in S, such that $h(s)\lambda_s(t)h(s)^{-1}$ is constant. Then the homomorphism λ'_s corresponding to the section $\sigma(s)h(s)^{-1}$ is constant in S on T_0.

First observe that if χ is any character on H then $\chi \circ \lambda_s$ is a trigonometric polynomial on T_0 whose coefficients are integer valued functions continuous in s, hence constant. Thus $\chi \circ \lambda_s(t)$ is constant and letting χ vary over all characters we see that all $\lambda_s(t)$ are in the same conjugacy class. Let Z be the centralizer of $\lambda_0(t)$. What we have shown is that S is mapped smoothly into one orbit of the conjugation action of H on itself. That orbit is diffeomorphic to H/Z. Hence we have a smooth map $S \to H/Z$ and composing with a local section of $H \to H/Z$ we have, after possibly restricting to an open subset $S' \subset S$, a function $h: S \to H$ such that

$$h(s)\lambda_s(t)h(s)^{-1} = \lambda_0(t).$$

Since G_0 and H are compact each is a product of a torus and a compact semisimple group — $G = A \times F$ and $H = B \times K$ with A, B tori and F, K compact semisimple groups. We can assume that $A \subset T_0$ and $T_0 \cap F = T_1$ is a maximal torus in F and that for some maximal torus T in H, $B \subset T$ and $T \cap K = T_2$ is a maximal torus in K. The restricted homomorphism (that is, $\lambda_s: T_0 \to T$ composed on the right with the inclusion $T_1 \to T_0$ and on the left with projection $T \to T_2$) $\lambda_s: T_1 \to T_2$ is constant in S and using the results of Dynkin[10] this shows that all the subgroups $\lambda_s(F)$ are conjugate in K. More precisely, the condition that $\lambda_s: T_1 \to T_2$ is constant in s implies that all $\lambda_s(F)$ are equivalent in every representation of K. For all semisimple groups there are at most finitely many conjugacy classes of semisimple subgroups of K which are equivalent in every representation. The continuity in s of λ_s implies that the conjugacy classes cannot vary with s and therefore all the $\lambda_s(F)$ are conjugate. That the conjugacy can be carried out with smooth dependence on s follows from the existence of smooth sections of $K/N(\lambda_0(F))$ the coset space of K by the normalizer of the subgroup $\lambda_0(F)$.

Combining this lemma with Propositions 1 and 2 gives us the following two theorems.

Theorem 1: Let M be a manifold with simple G-action and compact isotropy groups. Let E be a principal H bundle with G-action projecting on the G-action on M. Assume H is compact. Let $\varphi: S \to M$ be a special cross-section through $x \in M$ and $U = G \cdot \varphi(S) \subset M$. Then there is an isomorphism

$$E|_U \cong E_\lambda \times S, \quad \text{for some } \lambda: G_x \to H.$$

This theorem together with Proposition 1 and its corollaries completely analyzes the structure of a bundle with G action over the neighborhood of an orbit in space with a simple G action.

Proof: Let $\varphi: S \to M$ be a special cross-section and σ a section of E over φ such that the homomorphism λ_s is independent of s. Then define a mapping

$$f: E_{\lambda_0} \times S \to E|_U,$$

$$f([g,h],s) = g\sigma(s)h.$$

It is immediate to check that this is a G equivariant isomorphism.

Theorem 2: The G invariant connections on $E_\lambda \times S$ are determined by

(i) A family of linear maps $A_s: \mathcal{G} \to \mathcal{H}$ depending smoothly on s and satisfying (18a) and (18b).

(ii) A one-form μ on S with values in the subalgebra of $\mathcal{H}$ of elements invariant under the adjoint action of $\lambda_0(G_0)$.

$B_\psi\, d\psi$. For λ_1 the condition that B_ψ takes values in the Ad-invariants of the image of λ_1 implies $B_\psi \equiv 0$. The remaining conditions are (18a)

$$A_\psi(\mathrm{Ad}(g,g)^{-1}(\theta_1+\theta_2)) = \mathrm{Ad}g^{-1}A_\psi(\theta_1+\theta_2),$$

which implies

$$A_\psi(\theta_1+\theta_2) = a_\psi\theta_1 + b_\psi\theta_2$$

for a_ψ, b_ψ scalar function of ψ and the right-hand side interpreted as taking values in su(2). By condition 18b, for $\xi \in \mathrm{su}(2)$,

$$A(\theta_1+\theta_2)(\xi,-\xi) = \lambda_*(\xi,-\xi) = \xi,$$

thus $a_\psi - b_\psi = 1$. The connection pulled back to the base is

$$\tilde{\omega} = a_\psi\theta_1.$$

Example γ: This is an example which, in view of the orbit structure in M, goes beyond the scope of Theorems 1 and 2. We therefore begin by considering only the dense submanifold M_1.

Using the special cross section defined for $0 < s < \pi$ and $0 \leqslant t < 2\pi$

$$\varphi(s,t) = \left(\begin{pmatrix} e^{is} & 0 \\ 0 & e^{-is}\end{pmatrix}, e^{it}\right), \quad \text{we find}$$

$$G_0 = \left\{\begin{pmatrix} e^{is} & 0 \\ 0 & e^{-is}\end{pmatrix}\right\} \subset \mathrm{SU}(2).$$

For $p \in M_1$ and q in the fiber of E over p define $\lambda_q : G_0 \to H$ by

$$\tilde{\gamma}_g q = q\lambda_q(g),$$

where $\tilde{\gamma}$ is the action on E. The homomorphism λ_q is conjugate to some $\mu_n : G_0 \to H$, $n \in Z$, defined by

$$\mu_n \begin{pmatrix} e^{is} & 0 \\ 0 & e^{-is}\end{pmatrix} = \begin{pmatrix} e^{ins} & 0 \\ 0 & e^{-ins}\end{pmatrix}$$

hence, by continuity, the integer n characterizing the homomorphism of the isotropy group into the gauge group is independent of p. There exists an extension of λ_q to $\bar{\lambda}_q : G \to H$ if and only if $n(p) = 0$ or 1. Suppose the bundle E over M_1 extends to $\bar{E}$ over M and the G action on E extends to an action on $\bar{E}$ projecting to the γ action on M. We can find a section σ of $\bar{E}$ near $p(\psi) = (I, e^{i\psi})$ and since the isotropy group at $p(\psi)$ is SU(2) we have a homomorphism $\lambda_\psi : \mathrm{SU}(2) \to \mathrm{SU}(2)$. Restricted to G_0

$$\lambda_\psi \begin{pmatrix} e^{is} & 0 \\ 0 & e^{-is}\end{pmatrix} = \begin{pmatrix} e^{in(\psi)s} & 0 \\ 0 & e^{-in(\psi)s}\end{pmatrix}$$

where $n(\psi) \equiv 1$ or $n(\psi) \equiv 0$.

By continuity the first case implies $n(p) \equiv 1$ for $p \in M_1$ and the second case implies $n(p) \equiv 0$. In either case the homomorphism λ extends to $\bar{\lambda}_q : G \to H$ and we conclude that the transformation function can be chosen independent of the point in M_1. This implies that we can choose the transformation function to be either

$$\rho_0(g,y) \equiv I \in \mathrm{SU}(2),$$

$$\rho_1(g,y) = g^{-1} \in \mathrm{SU}(2).$$

The invariant connections corresponding to these transformation functions may be determined through Theorem 2 or by applying the theory of orthogonal invariants directly in the base space. The pull-backs ω_0 and ω_1 of the generic invariant connections corresponding to ρ_0 and ρ_1 respectively may be expressed as:

$$\omega_0 = M\, ds + N\, dt$$

and

$$\omega_1 = A dt + B\omega + C[U,\omega] + D(U,\omega)U,$$

where M, N are $\mathfrak{su}(2)$-algebra valued functions and A,B,C,D scalar functions depending on the invariants s and t only, U is an $\mathfrak{su}(2)$-valued function on M defined in the standard anti Hermitian representation[1] by

$$U(x,e^{i\psi}) = \tfrac{1}{2}(x - \tfrac{1}{2}\,\mathrm{Tr}x),$$

and ω is the Maurer–Cartan form in the first factor under the identification $M \sim \mathrm{SU}(2) \times \mathrm{U}(1)$.

ACKNOWLEDGMENTS

The authors would like to thank S. Drury for help with the proof of the Lemma preceding Theorems 1 and 2.

[1] J. Harnad, S. Shnider, and Luc Vinet, J. Math. Phys. **20**, 931 (1979); also in *Complex Manifold Techniques in Theoretical Physics*, edited by D. Lerner and P. Sommers (Pitman, New York 1979), pp. 219–30 and Refs. therein.
[2] P. Forgács and N. Manton, Comm. Math. Phys. **72**, 15 (1980).
[3] R. Jackiw and N. Manton, Ann. Phys.
[4] N. Manton, Nucl. Phys. B **158**, 141 (1979).
[5] M. Mayer, Univ. Cal. Irvine preprint.
[6] J. Harnad, S. Shnider, and J. Tafel, Lett. Math. Phys. **4**, 107 (1980).
[7] G. Bredon, *Introduction to Compact Transformaton Groups* (Academic, New York, 1972), Chap. 2.
[8] For 1-parameter groups, the infinitesimal transformation conditions on vector bundles are given by P. G. Bergmann and E. J. Flaherty, J. Math. Phys. **19**, 212 (1978).
[9] H. C. Wang, Nagoya Math. J. **13**, 1 (1958), cited in S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1969), Vol. 1, p. 106.
[10] E. B. Dynkin, A.M.S. Translations, Series 2 vol. 6, p. 111–214.

A matrix representation of the translation operator with respect to a basis set of exponentially declining functions

Eckhard Filter and E. Otto Steinborn[a)]
Institut für Chemie, Universität Regensburg, D-8400 Regensburg, West Germany

(Received 25 April 1980; accepted for publication 18 July 1980)

The matrix elements of the translation operator with respect to a complete orthonormal basis set of the Hilbert space $L_2(\mathbb{R}^3)$ are given in closed form as functions of the displacement vector. The basis functions are composed of an exponential, a Laguerre polynomial, and a regular solid spherical harmonic. With this formalism, a function which is defined with respect to a certain origin, can be "shifted", i.e., expressed in terms of given functions which are defined with respect to another origin. In this paper we also demonstrate the feasibility of this method by applying it to problems that are of special interest in the theory of the electronic structure of molecules and solids. We present new one-center expansions for some exponential-type functions (ETF's), and a closed-form expression for a multicenter integral over ETF's is given and numerically tested.

1. INTRODUCTION

The problem of how to perform spatial transformations, i.e. rotations and translations, of physical fields often arises in various branches of theoretical physics. A useful concept for the mathematical treatment of such transformations has been the introduction of operators which establish a mapping of a given function onto a new function: The operator maps the function which represents a given field onto the function which represents the transformed physical field. This concept is especially valuable if the functions, which describe the physical field, are subject to the condition of being elements of certain Hilbert spaces, because in this case, the translation and rotation operators often cause unitary transformations of the appropriate function spaces.

For rotational transformations this method has been used successfully in connection with the theory of angular momentum. The well-known results provide an easy possibility to rotate a physical field if its angular dependence can be represented by an element of the Hilbert space $L_2(\Omega)$.[1] This function space contains all square-summable functions, which are defined on the surface of a sphere in the three-dimensional Euclidian space.

Translational transformations can be described by a mapping of a function f, which represents the original field, onto a function F, which represents the shifted field. The mapping of f onto F can be formulated with the help of the translation operator $\mathcal{T}_{\mathbf{R}}$ defined by $\mathcal{T}_{\mathbf{R}}: f \to F$ or

$$\mathcal{T}_{\mathbf{R}} f(\mathbf{r}) = f(\mathbf{r} - \mathbf{R}) = F(\mathbf{r}). \tag{1.1}$$

We consider the case when the function f is an element of the Hilbert space $L_2(\mathbb{R}^3)$, which is of special importance in quantum mechanics. If f is an element of $L_2(\mathbb{R}^3)$ the function F is also an element of this space, and the translation operator causes a unitary transformation of the space $L_2(\mathbb{R}^3)$ which is closely connected to the Fourier–Plancherel transformation.[2] This connection is usually utilized to represent the translation operator by an integral-operator which provides the possibility to shift a physical field by applying two successive Fourier transformations to the function which represents the original field. However, this method has severe drawbacks. For instance, when applied to the difficult multicenter integral problem which plays an important role in the theory of the electronic structure of molecules and solids,[3] it has led to rather impractical results, although the calculations have been performed in a very sophisticated manner.[4]

In this paper we want to present an alternative method for the treatment of translational transformations of physical fields which are represented by functions of the space $L_2(\mathbb{R}^3)$. Because the translation operator is unitary, it is possible to represent it completely by its matrix, if the matrix elements are defined with respect to a complete and orthonormal basis set of the space $L_2(\mathbb{R}^3)$.[5] Since this Hilbert space is separable, only a countable set of matrix elements is needed for the representation. In practical calculations, a set of matrix elements can usually be handled rather easily by methods established in linear algebra. Therefore, a representation of an opertor in terms of matrix elements is often more practical than the representation by means of an integral operator, which can lead to serious analytical and numerical difficulties. The method of transforming functions with the help of a matrix representation of the translation operator has the further advantage that the basis functions, which are required, can be chosen properly according to the nature of the problems under consideration. This is rather important because the choice of the basis set will determine the rate of convergence of the resulting series expansions. As a complete and orthonormal basis set, we have chosen a set that consists of functions which are the product of an exponential, a Laguerre polynomial, and a regular solid spherical harmonic. The choice of spherical harmonics for the description of the angular part of the basis set functions offers the possibility to perform at once, if necessary, a rotational transformation of the field with the help of the well-known rotation matrix,[6] before the translation is considered. Utilizing results for the convolution of exponential-type functions which we derived recently,[7] we are able to find the complete matrix representation of the translation operator with respect to the chosen basis set in a rather compact analytical form. Applying this method to the problem of finding one-center expansions of given functions, we also derive new ad-

[a)]Author to whom correspondence should be directed.

 0022-2488/80/122725-12$1.00

dition theorems for "reduced Bessel–", Laguerre–, and Slater–type functions which have some striking advantages compared to results which have been given in the literature so far.

2. REALIZATIONS OF THE TRANSLATION OPERATOR

A given scalar field in three-dimensional space may mathematically be represented by a function $f(\mathbf{r})$ which is defined with respect to a certain coordinate system; the functional form of f depends on the choice of this coordinate system. If the field is subject to a translation (without rotation) defined by a displacement vector $\mathbf{R}$, then the field is mathematically represented by a new function $F(\mathbf{r})$ which is defined with respect to the same coordinate system. Hence, the "new function" F has the same value at the point defined by the local vector $\mathbf{r}$ as the "old function" f has at the point defined by the local vector $(\mathbf{r}-\mathbf{R})$, i.e.,

$$f(\mathbf{r}-\mathbf{R}) = F(\mathbf{r}). \tag{2.1a}$$

As this equality holds for any point in three-dimensional space, the relationship Eq. (2.1a) is an identity among the values of the functions, which is valid for each point. If the new function f is described with the help of an operator $\mathcal{T}_{\mathbf{R}}$, where $\mathcal{T}_{\mathbf{R}}$ changes f into F, the *functional* relationship between f and F, which corresponds to Eq. (2.1a), is given by

$$\mathcal{T}_{\mathbf{R}} f(\mathbf{r}) = F(\mathbf{r}). \tag{2.1b}$$

In Eq. (2.1b), the *functions* do not depend on the displacement vector $\mathbf{R}$ any more, as they did in Eq. (2.1a): Now, the dependence upon $\mathbf{R}$ is put into the $\mathcal{T}_{\mathbf{R}}$ operator only. If f can be expanded into a three-dimensional Taylor series the translation operator $\mathcal{T}_{\mathbf{R}}$ can be represented by the differential operator $\mathcal{T}_{\mathbf{R}} = \exp(-\mathbf{R}\cdot\partial/\partial\mathbf{r})$. For quantum mechanical investigations, it is often sufficient to consider $\mathcal{T}_{\mathbf{R}}$ as an operator that operates on such functions f as are elements of certain Hilbert spaces. Then, the relationships Eqs. (2.1a,b), which are pointwise valid, can be replaced by the equation

$$\mathcal{T}_{\mathbf{R}}|f\rangle = |F\rangle. \tag{2.2}$$

For a given Dirac–ket $|f\rangle$, the ket $|F\rangle$ will depend on $\mathbf{R}$ as a parameter. In order to analyze this dependence, the operator $\mathcal{T}_{\mathbf{R}}$ has to be specified in a way which exhibits its $\mathbf{R}$ dependence in explicit form. If this is done, it will be possible to execute analytical calculations with the help of $\mathcal{T}_{\mathbf{R}}$.

A well-known realization of the translation operator $\mathcal{T}_{\mathbf{R}}$, which can be used for the shifting of functions that are elements of the Hilbert space $L_2(V)$, is the matrix representation of $\mathcal{T}_{\mathbf{R}}$ with respect to a basis of plane wave functions $|\mathbf{k}\rangle = V^{-1/2}\exp(i\mathbf{k}\cdot\mathbf{r})$, where $\mathbf{k} = 2\pi(n_1/a, n_2/b, n_3/c)$ and $V = a\cdot b\cdot c$ specifies a normalization volume:

$$\begin{aligned}\mathcal{T}_{\mathbf{R}} &= \sum_{\mathbf{k}}\sum_{\mathbf{k}'}|\mathbf{k}\rangle\langle\mathbf{k}|\mathcal{T}_{\mathbf{R}}|\mathbf{k}'\rangle\langle\mathbf{k}'| \\ &= \sum_{\mathbf{k}}|\mathbf{k}\rangle e^{-i\mathbf{k}\cdot\mathbf{R}}\langle\mathbf{k}|.\end{aligned} \tag{2.3}$$

Here it is assumed that every function which is an element of $L_2(V)$ is extended to a function defined in the whole three-dimensional space $\mathbb{R}^3$ by periodic continuation.

The situation becomes more complicated if one considers the translation operator acting on the space $L_2(\mathbb{R}^3)$, because then it is no longer possible to decompose the space into a denumerable direct sum of invariant subspaces.[8] A realization which is formally similar to Eq. (2.3) can be obtained with the help of the Fourier–Plancherel transformation given by

$$\bar{f}(\mathbf{k}) = \mathcal{U} f(\mathbf{r}) = (2\pi)^{-3/2}\int d\mathbf{r}\,\exp(-i\mathbf{k}\cdot\mathbf{r})f(\mathbf{r}), \tag{2.4}$$

utilizing a Fourier integral operator $\mathcal{U}$. Because $\mathcal{U}$ defines a unitary transformation due to Plancherel,[2] it is possible to obtain a representation of $\mathcal{T}_{\mathbf{R}}$ from the unitary equivalent operator $\bar{\mathcal{T}}_{\mathbf{R}}$ defined by $\bar{\mathcal{T}}_{\mathbf{R}}\bar{f}(\mathbf{k}) = \bar{F}(\mathbf{k})$. Because $\bar{\mathcal{T}}_{\mathbf{R}} = \exp(-i\mathbf{k}\cdot\mathbf{R})$ it follows that

$$\begin{aligned}\mathcal{T}_{\mathbf{R}} = \mathcal{U}^{\dagger}\bar{\mathcal{T}}_{\mathbf{R}}\mathcal{U} = (2\pi)^{-3}&\int d\,\mathbf{k}\exp(i\mathbf{k}\cdot\mathbf{r}) \\ &\times\exp(-i\mathbf{k}\cdot\mathbf{R})\cdot\int d\mathbf{r}'\exp(-i\mathbf{k}\cdot\mathbf{r}').\end{aligned} \tag{2.5}$$

This realization of the translation operator via Fourier integrals, however, often leads to serious integration problems if the integrals are to be evaluated for practical purposes.

In this article we present a new analytical realization of the translation operator in $L_2(\mathbb{R}^3)$ by means of its matrix elements. The details of the results and the derivations will be given in Sec. 4. Now we discuss some more general aspects.

The translation operator $\mathcal{T}_{\mathbf{R}}$ is unitary if it acts on the space $L_2(\mathbb{R}^3)$. Therefore, $\mathcal{T}_{\mathbf{R}}$ is also a linear and bounded operator, and the following realization of the operator must be possible[5]:

$$\mathcal{T}_{\mathbf{R}} = \sum_{n_1}\sum_{n_2}|\phi_{n_1}\rangle\langle\phi_{n_1}|\mathcal{T}_{\mathbf{R}}|\phi_{n_2}\rangle\langle\phi_{n_2}|. \tag{2.6}$$

As expansion basis $\{\phi_n\}$, any complete orthonormal set of $L_2(\mathbb{R}^3)$ can be used. As the space $L_2(\mathbb{R}^3)$ cannot be decomposed into a direct sum of translationally invariant subspaces, no basis set $\{\phi_n\}$ exists which would reduce the matrix to block-diagonal form. However, this disadvantage will often be compensated by the possibility to use methods of linear algebra when the representation as it is given by Eq. (2.6) is applied. In Eq. (2.6) the $\mathbf{R}$–dependence of $\mathcal{T}_{\mathbf{R}}$ is expressed completely by a set of matrix elements

$$(\mathcal{T}_{\mathbf{R}})_{n_1,n_2} = \langle\phi_{n_1}|\mathcal{T}_{\mathbf{R}}|\phi_{n_2}\rangle. \tag{2.7a}$$

Each matrix element can be considered as a function S of $\mathbf{R}$, i.e.,

$$(\mathcal{T}_{\mathbf{R}})_{n_1,n_2} = S_{n_1,n_2}(\mathbf{R}). \tag{2.7b}$$

The determination of these functions is essential for the applicability of Eq. (2.6).

A rather general method to obtain formal expressions for these functions is provided by the Fourier transform convolution theorem,[9] which states that the Fourier transforms $\bar{S}_{n_1,n_2}$, $\bar{\phi}_{n_1}$ and $\bar{\phi}_{n_2}$ are related by

$$\bar{S}_{n_1,n_2}(\mathbf{k}) = (2\pi)^{3/2}\bar{\phi}_{n_1}(\mathbf{k})\bar{\phi}_{n_2}(\mathbf{k}). \tag{2.8}$$

This relationship converts the matrix elements given by Eq. (2.7a), which are in fact two-centric convolution integrals, into one-centric Fourier integrals

$$\begin{aligned}S_{n_1,n_2}(\mathbf{R}) &= (\mathcal{T}_{\mathbf{R}})_{n_1,n_2} \\ &= \int d\,\mathbf{k}\,e^{i\mathbf{k}\cdot\mathbf{R}}\bar{\Phi}_{n_1}(\mathbf{k})\bar{\Phi}_{n_2}(\mathbf{k}).\end{aligned} \tag{2.9}$$

This equation can also be obtained if in Eq. (2.7a) the opera-

tor $\mathcal{T}_{\mathbf{R}}$ is substituted according to Eq. (2.5).

Another rather general method is the procedure that one tries to find the function $S(\mathbf{R})$ as a series in terms of functions of $\mathbf{R}$ which constitute an orthonormal set. If one chooses the original functions ϕ_n, one obtains the formal expansion

$$S_{n_1,n_2}(\mathbf{R}) = \sum_{\nu} c_{\nu}^{n_1,n_2} \phi_{\nu}(\mathbf{R}). \tag{2.10}$$

Such an expansion is always possible if the product $\bar{\phi}_{n_1} \cdot \bar{\phi}_{n_2}$ of the Fourier transforms $\bar{\phi}_{n_1}$ and $\bar{\phi}_{n_2}$ is an element of $L_2(\mathbb{R}^3)$. Then, $\bar{S}$ as given by Eq. (2.8), and, therefore, also S must be elements of this space. In this case, the series expansion as given by Eq. (2.10) is at least convergent in the mean, and the expansion coefficients are given by the integrals

$$c_{\nu}^{n_1,n_2} = (2\pi)^{3/2} \langle \bar{\phi}_{\nu} | \bar{\phi}_{n_1} | \bar{\phi}_{n_2} \rangle, \tag{2.11}$$

as can be seen with the help of Eq. (2.9).

From Plancherel's theorem it is clear that the expansion coefficients $\langle \bar{\phi}_{\nu} | \bar{\phi}_{n_1} | \bar{\phi}_{n_2} \rangle$ are integrals over three orthogonal functions. This sort of integral is often hard to evaluate. Only a few special results are available even if the functions $\bar{\phi}_n$ consist of classical orthogonal polynomials multiplied by appropriate weight functions.[10] The same is true for the kind of integrals given by Eq. (2.9). Therefore, it will often only be possible to evaluate the matrix elements of the translation operator with respect to a given basis set if specific mathematical relationships are available which make it possible to avoid the rather general but complex integral representations given by Eqs. (2.9) and (2.11).

An explicit matrix realization of the translation operator requires the choice of a basis set. Obviously, if the matrix representation of the translation operator is used for the translational transformation of a given function, the rate of convergence of the resulting series expansion will depend on the choice of the basis functions which are used. We will consider the case when the functions of the complete and orthonormal basis set consist of the product of an exponential, a Laguerre polynomial, and a regular solid spherical harmonic. It is to be expected that the representation of the translation operator with respect to this set of exponential-type functions (ETF's) will lead to rapidly convergent series, if it is used for the shifting of fields which are described by exponentially declining or similar "strongly localized" functions, because then the main character of the original function is preserved by the basis functions. Therefore, the choice of ETF's as basis functions for the matrix representation of the translation operator should be especially appropriate for applications to problems in which ETF's are to be transformed, as it is the case in the context of various fields of theoretical physics, as especially in atomic, molecular, and solid-state theory.

3. ORTHOGONAL AND NONORTHOGONAL BASIS SETS OF EXPONENTIALLY DECLINING FUNCTIONS

In this section, different orthogonal and nonorthogonal exponentially declining basis functions will be defined for later use. For the different classes of ETF's studied here, new relationships will be given which make it possible to transform a given type of ETF into another type of ETF. These relationships enhance the applicability of ETF's, and they also make it possible to transform the formulas given in the present paper into formulas which hold for those ETF's which are not used in this article.

As an exponential-type function (ETF) we denote a function of the form

$$\phi_{n,l}^{m}(\mathbf{r}) = e^{-r} p_n(r) \mathcal{Y}_l^m(\mathbf{r}), \tag{3.1}$$

where $p_n(r)$ is an arbitrary polynomial of order n. The regular solid spherical harmonic $\mathcal{Y}_l^m(\mathbf{r})$ stands for the product $r^l Y_l^m(\theta,\phi)$, where for the surface spherical harmonic Y_l^m the definition of Condon and Shortley is used.[11] The function sets considered in the following differ from each other only by the choice of the polynomial part $p_n(r)$; each choice of a certain kind of polynomial $p_n(r)$ leads to a certain set of ETF's. The various sets obtained in this way exhibit different properties as far as orthogonality and completeness is concerned.

Well-known ETF's of the type defined by Eq. (3.1) are the bound-state wave functions of the electron in the hydrogen atom. These functions are orthogonal but do not form a complete set of functions in $L_2(\mathbf{R}^3)$. A complete set can be obtained only if the Coulomb functions which belong to the continuous spectrum are included.[12] However, these functions are not of the form Eq. (3.1). Therefore, the solutions of the one-center-one-electron Coulomb problem do not form a countable exponential-type basis set and, therefore, the "basis" consisting of all hydrogen functions (including the continuum) is not suitable for our purposes.

A well established exponential-type function set is given by the system of (unnormalized) Slater–type functions (STF's) which are defined by

$$\chi_{N,L}^{M}(\alpha r) = (\alpha r)^{N-1} e^{-\alpha r} Y_L^M(\Omega_r). \tag{3.2}$$

These functions are a complete but not orthogonal basis set for the space $L_2(\mathbb{R}^3)$.[13] Slater–type functions (or Slater–type orbitals, STO's) are widely used in atomic, molecular, and solid-state theory.

Investigating the integral and convolutional as well as the transformational properties of Slater–type functions and their applicablity in electronic structure calculations,[7,14] we have recently introduced the so-called B functions[15] which have some remarkable advantages over other kinds of ETF's:

$$B_{N,L}^{M}(\alpha \mathbf{r}) = \hat{k}_{N-1/2}(\alpha r) \mathcal{Y}_L^M(\alpha \mathbf{r}) [(2N+2L)!!]^{-1}, \tag{3.3a}$$

$$\hat{k}_{n-1/2}(r) = r^{-1} e^{-r} \sum_{p=1}^{N} \frac{(2N-p-1)!}{(p-1)!(N-p)!} 2^{p-N} r^p$$

$$= (2/\pi)^{1/2} r^{N-1/2} K_{N-1/2}(r). \tag{3.3b}$$

Here, $\hat{k}_{\nu}$ is the so-called reduced Bessel function[7,14] (RBF), which is closely related to the modified Bessel function of the second kind K_{ν}.[16] These B functions are closely connected with Slater–type functions, and can like these be used for the calculation of properties of small molecules.[17] The set of all B functions with different n,l,m is complete but not orthogonal. The completeness follows from the fact that the basis set of B functions can be obtained from the basis set of Slater–type functions by a linear transformation with a triangular matrix.[13]

Another set of ETF's which is a complete and also orthogonal basis set for $L_2(\mathbf{R}^3)$ is given by the Laguerre func-

tions $\Lambda^M_{N,L}$ which we define by the equation

$$\Lambda^m_{n,l}(\alpha\mathbf{r}) = \mathcal{N}(n,l)\alpha^{3/2} L^{(2l+2)}_{n-l-1}(2\alpha r) e^{-\alpha r} \mathcal{Y}^m_l(2\alpha\mathbf{r}). \quad (3.4a)$$

The normalization factor is given by

$$\mathcal{N}(n,l) = 2^{3/2}[(n-l-1)!/(n+l+1)!]^{1/2}. \quad (3.4b)$$

For the Laguerre polynomial we use the definition[18]

$$L^{(\alpha)}_n(x) = \sum_{m=0}^{n} (-1)^m \binom{n+\alpha}{n-m} \frac{x^m}{m!}. \quad (3.5)$$

Functions of this kind have also successfully been used in electronic structure calculations.[19] For our purposes the Λ functions are especially useful because it is possible to express the unit operator in $L_2(\mathbb{R}^3)$ as

$$\sum_{N,L,M} |\Lambda^M_{N,L}\rangle \langle \Lambda^M_{N,L}| = 1. \quad (3.6)$$

The decomposition of the unit operator with respect to a complete basis set that consists of functions which are not orthogonal is usually much more complicated. This is the reason why we used the set of Λ functions for the expansion of the translation operator in Sec. 4. It is important, however, to be able to change from one basis set to another one. Therefore, we are now going to derive the necessary transformation formulas which relate the Slater-type, B, and Λ functions to each other.

The formulas describing the transformation of Slater-type functions into B functions and the inverse transformation read[14]

$$\chi^M_{N,L}(\mathbf{r}) = \sum_{p=\min(p)}^{N-L} \left[\frac{(-1)^{N-L-p}(N-L)!2^{L+p}(L+p)!}{(2p-N+L)!(2N-2L-2p)!!} B^M_{p,L}(\mathbf{r}) \right], \quad (3.7a)$$

$$\min(p) = \begin{cases} (N-L)/2 & \text{for } N-L \text{ even} \\ (N-L+1)/2 & \text{for } N-L \text{ odd} \end{cases}, \quad (3.7b)$$

$$B^M_{N,L}(\mathbf{r}) = [(2N+2L)!!]^{-1} \times \sum_{p=1}^{N} \frac{(2N-p-1)!2^{p-N}}{(p-1)!(N-p)!} \chi^M_{p+L,L}(\mathbf{r}). \quad (3.8)$$

The transformation of Λ functions into Slater-type functions can be obtained directly by representing the Laguerre polynomial in terms of powers of r according to Eq. (3.5), leading to

$$\Lambda^M_{N,L}(\mathbf{r}) = \mathcal{N}(N,L) \sum_{p=0}^{N-L-1} (-2)^p 2^L (p!)^{-1} \times \binom{N+L+1}{N-L-p-1} \chi^M_{p+L+1,L}(\mathbf{r}). \quad (3.9)$$

Because any power of r can be expressed in terms of Laguerre polynomials $L^{(\alpha)}_\nu(r)$,[20] we obtain immediately for the inverse relationship

$$\chi^M_{N,L}(\mathbf{r}) = (N-L-1)!2^{-2L} \sum_{p=0}^{N-L-1} (-1)^p \times \binom{N+L+1}{N-L-1-p} \mathcal{N}^{-1}(p+L+1,L) \times \Lambda^M_{p+L+1,L}(\mathbf{r}). \quad (3.10)$$

The expansion of B functions in terms of Λ functions can be obtained by starting from the interesting relationship

$$\hat{k}_{n+1/2}(r) = (-2)^{-n} n!\, e^{-r} L^{(-2n-1)}_n(2r), \quad (3.11)$$

which follows from a comparison of the definitions of the $\hat{k}$ functions with the Laguerre functions. Then, with the help of the relationship[20]

$$L^{(\alpha)}_n(x) = \sum_{m=0}^{n} \frac{(\alpha-\beta)_m}{m!} L^{(\beta)}_{n-m}(x), \quad (3.12)$$

the $\hat{k}$ function can be expressed as a linear combination of Laguerre polynomials with arbitrary upper index times an exponential function according to

$$\hat{k}_{n+1/2}(r) = (-2)^{-n} n!\, e^{-r} \sum_{m=0}^{n} (-1)^m \binom{2n+1+\beta}{n-m} \times L^{(\beta)}_m(2r). \quad (3.13)$$

Setting $\beta = 2L+2$ and multiplying Eq. (3.13) by a regular solid spherical harmonic $\mathcal{Y}^M_L$, we find

$$B^M_{N+1,L}(\mathbf{r}) = \frac{N!}{(2N+2L+2)!!} \sum_{p=0}^{N} (-1)^p \times \binom{2N+2L+3}{N-p} \mathcal{N}^{-1}(p+L+1,L) \times \Lambda^M_{p+L+1,L}(\mathbf{r}). \quad (3.14)$$

In order to express Λ by B functions, which results in the inversion of the transformation Eq. (3.14), we start from the relationship

$$x^{-1} e^{-x} L^{(\alpha)}_n(2x) = \sum_{t=0}^{n} \frac{(-2)^t \Gamma(n+\alpha+t+1)}{t!(n-t)!\Gamma(\alpha+2t+1)} \hat{k}_{t-1/2}(x), \quad (3.15)$$

which can be obtained as follows: In the Laguerre polynomial Eq. (3.5) given in terms of powers of x, with the help of Eq. (3.7) the various terms of the form $x^\nu e^{-x}$, which occur, are expressed by reduced Bessel functions $\hat{k}$. If we collect the $\hat{k}$ functions with equal index, we obtain a linear combination of $\hat{k}$ functions, where the coefficients are given as finite sums. These sums can be expressed in closed form by Vandermonde's theorem,[21] yielding Eq. (3.15). Now, making use of the recursion relation[22]

$$xL^{(\alpha+1)}_n(x) = (n+\alpha+1)L^{(\alpha)}_n(x) - (n+1)L^{(\alpha)}_{n+1}(x), \quad (3.16)$$

the factor x^{-1} in formula Eq. (3.15) can be eliminated, leading to[23]

$$e^{-x} L^{(\alpha)}_n(2x) = (2n+\alpha+1) \times \sum_{t=0}^{n} \frac{(-2)^t \Gamma(n+\alpha+t+1)}{t!(n-t)!\Gamma(\alpha+2t+2)} \hat{k}_{t+1/2}(x). \quad (3.17)$$

For $\alpha = 2l+2$ we finally obtain the relationship which enables us to transform Λ functions into B functions:

$$\Lambda^M_{N,L}(\mathbf{r}) = \sum_{t=L+1}^{N} b^{N,L}_t B^M_{t-L,L}(\mathbf{r}), \quad (3.18a)$$

with

$$b_t^{N,L} = (-1)^{t-L-1}2^{t-1}(N+t)! \times [(t-L-1)!(N-t)!(2t+1)!!]^{-1}(2N+1) \times \mathcal{N}(N,L). \quad (3.18b)$$

4. MATRIX ELEMENTS OF THE TRANSLATION OPERATOR

The matrix representation of the translation operator $\mathcal{T}_{\mathbf{R}}$ in terms of a general expansion basis $\{\Phi_n\}$ was introduced by Eq. (2.6). Now, as a specific expansion basis we choose the Λ basis, which consists of all $\Lambda_{N,L}^{M}$ functions as given by Eqs. (3.4a,b). Then, in Dirac's notation the matrix representation of $\mathcal{T}_{\mathbf{R}}$ can be written as[24]

$$\mathcal{T}_{\mathbf{R}} = \sum_{N_1L_1M_1} \sum_{N_2L_2M_2} |\Lambda_{N_1L_1}^{M_1}\rangle \langle \Lambda_{N_1L_1}^{M_1}|\mathcal{T}_{\mathbf{R}}|\Lambda_{N_2L_2}^{M_2}\rangle \langle \Lambda_{N_2L_2}^{M_2}|. \quad (4.1)$$

The applicability of this expansion to practical problems depends strongly on the efficient calculation of the coefficients

$$(\mathcal{T}_{\mathbf{R}})_{N_1L_1,N_2L_2}^{M_1,M_2} = \langle \Lambda_{N_1L_1}^{M_1}|\mathcal{T}_{\mathbf{R}}|\Lambda_{N_2L_2}^{M_2}\rangle \quad (4.2)$$

that occur in the expansion. In this section we will show that with the chosen basis it is possible to obtain very compact analytical formulas for these matrix elements which are well suited for practical applications. In Sec. 4A we will list the main results. The derivation of these formulas will be given in Sec. 4B.

A. Results

The matrix elements of the translation operator $\mathcal{T}_{\mathbf{R}}$ are three-dimensional convolution products, i.e., overlap integrals, of the Λ functions according to

$$(\mathcal{T}_{\mathbf{R}})_{N_1L_1,N_2L_2}^{M_1,M_2} = \int d\mathbf{r}\, \Lambda_{N_1L_1}^{M_1\dagger}(\mathbf{r})\mathcal{T}_{\mathbf{R}}\Lambda_{N_2L_2}^{M_2}(\mathbf{r}) = \int d\mathbf{r}\, \Lambda_{N_1L_1}^{M_1\dagger}(\mathbf{r})\Lambda_{N_2L_2}^{M_2}(\mathbf{r}-\mathbf{R}). \quad (4.3)$$

These matrix elements can be expressed as linear combinations of functions $\Lambda_{\nu,l}^{m}(\mathbf{R})$ which depend on the displacement vector $\mathbf{R}$. By doing so, the following simple analytical expression is obtained:

$$(\mathcal{T}_{\mathbf{R}})_{N_1L_1,N_2L_2}^{M_1,M_2} = \sum_l \langle L_2M_2|L_1M_1|lm\rangle \times \sum_{n=\min(n)}^{\max(n)} T_{n,l}^{N_1L_1N_2L_2}\Lambda_{n,l}^{m}(\mathbf{R}), \quad (4.4a)$$

$$\min(n) = \max(l+1, |N_1-N_2|-1),$$
$$\max(n) = N_1+N_2+1, \quad m = M_2-M_1. \quad (4.4b)$$

The important conditions Eq. (4.4b) and Eq. (4.4c) are very similar to the triangular condition for the Gaunt coefficient[25] $\langle L_2M_2|L_1M_1|lm\rangle$ which limits the l summation to the range

$$|L_1-L_2| \leqslant l \leqslant L_1+L_2. \quad (4.4d)$$

The range of the summation index n in Eq. (4.4a) is given by $\max(n) - \min(n)$. This range, i.e., the number of $\Lambda_{n,l}^{m}$ functions that occur in the summation, is strictly limited. The lower limit of n increases if the upper limit increases. In many cases, therefore, the number of terms in the summation will not increase considerably if the indices N_1,L_1,N_2,L_2 are raised. As can be seen from examples discussed in Sec. 5B, this fact improves considerably the practical applicability of the formulas.

The coefficients $T_{n,l}^{N_1L_1N_2L_2}$ in the expansion Eq. (4.4a) fulfill the following useful symmetry relations:

$$T_{n,l}^{N_1L_1N_2N_2} = (-1)^l T_{n,l}^{N_2L_2N_1L_1}, \quad (4.5a)$$
$$T_{n,l}^{N_1L_1N_2L_2} = T_{N_2L_2}^{N_1L_1nl}, \quad (4.5b)$$
$$T_{n,l}^{N_1L_1N_2L_2} = (-1)^{L_2} T_{N_1L_1}^{nlN_2L_2}. \quad (4.5c)$$

For the coefficients $T_{n,l}^{N_1L_1N_2L_2}$ we first give integral representations. The coefficients $T_{n,l}^{N_1L_1N_2L_2}$ can be represented by a double convolution integral

$$\langle L_2M_2|L_1M_1|L_3M_3\rangle T_{N_3L_3}^{N_1L_1N_2L_2} = \int d\mathbf{r}\, \Lambda_{N_1L_1}^{M_1\dagger}(\mathbf{r}) \int d\mathbf{R}\, \Lambda_{N_3L_3}^{M_3\dagger}(\mathbf{R})\Lambda_{N_2L_2}^{M_2}(\mathbf{r}-\mathbf{R}). \quad (4.6)$$

It is also possible to obtain the coefficients as a one-dimensional integral over the product of three Jacobi polynomials[26] $P_N^{(\alpha,\beta)}(x)$:

$$T_{N_3L_3}^{N_1L_1N_2L_2} = (-1)^{\Delta L_1} c(N_1,L_1)c(N_2,L_2)c(N_3,L_3) \times \int_{-1}^{1} dx\,(1+x)^{\sigma+1/2}(1-x)^{\sigma+7/2} P_{N_1-L_1-1}^{(L_1+3/2,L_1+1/2)}(x) \times P_{N_2-L_2-1}^{(L_2+3/2,L_2+1/2)}(x) P_{N_3-L_3-1}^{(L_3+3/2,L_3+1/2)}(x) \quad (4.7a)$$

with

$$c(N,L) = 2^{N-L-1/2} \times [(N+L+1)!(N-L-1)!]^{1/2}/(2N-1)!!, \quad (4.7b)$$
$$\sigma = (L_1+L_2+L_3)/2, \quad \Delta L_1 = (L_2+L_3-L_1)/2. \quad (4.7c)$$

The last integral can be transformed into a simple linear combination of integrals over three Gegenbauer polynomials[27] $C_N^{(\nu)}(x)$:

$$T_{N_3L_3}^{N_1L_1N_2L_2} = 2^{L_1+L_2+L_3}(L_1!L_2!L_3!)(-1)^{\Delta L_1} \times \sum_{i,j,k=0}^{1} \gamma_i(N_1,L_1)\gamma_j(N_2,L_2) \times \gamma_k(N_3,L_3) \int_{-1}^{1} dx\,(1-x^2)^{\sigma+1/2} \times C_{N_1-L_1-i}^{(L_1+1)}(x) C_{N_2-L_2-j}^{(L_2+1)}(x) C_{N_3-L_3-k}^{(L_3+1)}(x), \quad (4.8a)$$

with

$$\gamma_i(N,L) = \begin{cases} (N+L+1)\mathcal{N}(N,L) & \text{for } i=1 \\ (N-L)\mathcal{N}(N,L) & \text{for } i=0 \end{cases}. \quad (4.8b)$$

These formulas are related to Eq. (2.11). For some special values of the parameters, the formulas Eqs. (4.7) and (4.8) reduce to integrals for which closed form expressions can be found in standard tables. For the general case, however, these integrals seem to be unknown.

Secondly we have found explicit formulas which express the general coefficients $T_{nl}^{N_1L_1N_2L_2}$ that occur in the series expansion Eq. (4.4), as finite sums. An explicit formula for the coefficients is the following finite triple sum:

$$T_{N_3L_3}^{N_1L_1N_2L_2} = 4\pi(-1)^{\Delta L_1}(2\sigma+1)!! \Big\{ \sum_{t_1t_2t_3} b_{t_1}^{N_1L_1} b_{t_2}^{N_2L_2} b_{t_3}^{N_3L_3}$$

$$\times \frac{(2t_1+2t_2+2t_3-2\sigma+1)!!}{(2t_1+2t_2+2t_3+4)!!}\Bigg\}, \tag{4.9a}$$

$$L_i+1 \leqslant t_i \leqslant N_i, \quad i=1,2,3. \tag{4.9b}$$

This expression exhibits the symmetry properties as stated in Eqs. (4.5a)–(4.5c). It may be applied to the numerical calculation of the coefficients if the "quantum numbers" N_1, N_2, N_3 are not too high.

If the "quantum numbers" N_1, N_2, N_3 are arranged according to the condition

$$N_3 \geqslant N_1, N_2 \tag{4.10a}$$

which is always possible with the help of Eqs. (4.5a)–(4.5c), then it is possible to express the nonvanishing coefficients $T^{N_1L_1N_2L_2}_{N_3L_3}$ by the following formula

$$\begin{aligned} T^{N_1L_1N_2L_2}_{N_3L_3} = {} & \pi(-1)^{L_1+N_3-1}\mathcal{N}^{-1}(N_3,L_3)2^{-N_3+3} \\ & \times \sum_{t_1=\min(t_1)}^{\Delta N_3}\Bigg\{\sum_{t_2=\min(t_2)}^{t_1} b^{N_1L_1}_{t_2-t_1+N_1} \\ & \times \frac{(t_2+N_3-L_3-1)!(2t_2+2N_3+1)!!}{(t_2+2N_3+1)!(2t_2)!!}\Bigg\} \\ & \times \Bigg\{\sum_{p=\min(p)}^{\max(p)}(-1)^p\binom{\Delta L_3}{p} b^{N_2L_2}_{t_1-\Delta N_3+N_2+p}\Bigg\}, \end{aligned} \tag{4.10b}$$

where we have put

$$\Delta N_3 = N_1+N_2+N_3+1, \tag{4.10c}$$

$$\Delta L_3 = (L_1+L_2-L_3)/2. \tag{4.10d}$$

The summation limits are given by

$$\min(t_1) = \max(0, \Delta N_3 - N_2 + L_2 + 1 - \Delta L_3) \tag{4.10e}$$

$$\min(t_2) = \max(0, L_1 + 1 + t_1 - N_1), \tag{4.10f}$$

$$\min(p) = \max(0, \Delta N_3 - N_2 + L_2 + 1 - t_1), \tag{4.10g}$$

$$\max(p) = \min(\Delta L_3, \Delta N_3 - t_1). \tag{4.10h}$$

The coefficients b^{NL}_t are defined by Eq. (3.18b). The summations on the r.h.s. of Eq. (4.10b) contain only a very limited number of terms due to the values of the upper and lower limits. Therefore, this expression Eq. (4.10) is well suited for numerical calculations.

B. Derivations

The formulas given above can be derived as follows. In the first place, we consider the integral

$$\begin{aligned} I^{L_1L_2}_{N_1N_2}(\mathbf{R}) = {} & \int d\mathbf{r} \Lambda^{M\dagger}_{N,L_1}(\mathbf{r}) e^{-|\mathbf{r}-\mathbf{R}|} \\ & \times p_{N_2}(|\mathbf{r}-\mathbf{R}|)\mathcal{Y}^{M_2}_{L_2}(\mathbf{r}-\mathbf{R}), \end{aligned} \tag{4.11}$$

where p_{N_2} stands for an arbitrary polynomial of degree N_2. For the evaluation of this integral, we make use of the convolution theorem of B functions which we have derived recently[28]:

$$\begin{aligned} & \int d\mathbf{r}\, B^{M\dagger}_{N,L_1}(\mathbf{r}) B^{M_2}_{N_2L_2}(\mathbf{r}-\mathbf{R}) \\ & = 4\pi \sum_l \langle L_2M_2|L_1M_1|lm\rangle (-1)^{L_2} \\ & \times \sum_t (-1)^t \binom{\Delta l}{t} B^m_{N_1+N_2+L_1+L_2-t-l+1,l}(\mathbf{R}). \end{aligned} \tag{4.12}$$

The $B^M_{N,L}$ functions as defined by Eq. (3.3) consist of a product of a polynomial of degree $N-1$, an exponential, and a solid spherical harmonic. Therefore, it must be possible to express the integral Eq. (4.11) as a linear combination of convolution products of B functions. Inspection of the degree of the powers of R, which occur in the convolution theorem Eq. (4.12), leads to the conclusion that the following finite series expansion for the integral $I^{L_1L_2}_{N_1N_2}(\mathbf{R})$ of Eq. (4.11) must be valid:

$$\begin{aligned} I^{L_1L_2}_{N_1N_2}(\mathbf{R}) = {} & \sum_l \langle L_2M_2|L_1M_1|lm\rangle \\ & \times \sum_{\nu=\min(\nu)}^{\max(\nu)} \xi^{N_1L_1N_2L_2}_{\nu,l} \Lambda^m_{\nu,l}(\mathbf{R}), \end{aligned} \tag{4.13a}$$

with

$$\min(\nu) = l+1, \quad \max(\nu) = N_1+N_2+L_2+2. \tag{4.13b}$$

Because the Λ functions form an orthonormal set, the coefficients ξ in Eq. (4.13) are given by the integrals

$$\begin{aligned} & \langle L_2M_2|L_1M_1|lm\rangle \xi^{N_1L_1N_2L_2}_{\nu,l} \\ & = \int d\mathbf{R}\, \Lambda^{m*}_{\nu,l}(\mathbf{R}) I^{L_1L_2}_{N_1N_2}(\mathbf{R}) \\ & = (-1)^{L_2}\int d\mathbf{r}\, \Lambda^{M\dagger}_{N_1L_1}(\mathbf{r}) I^{lL_2}_{\nu N_2}(r). \end{aligned} \tag{4.14}$$

If $I^{lL_2}_{\nu N_2}(\mathbf{r})$ in Eq. (4.14) is expressed according to Eq. (4.13), we obtain

$$\begin{aligned} & \langle L_2M_2|L_1M_1|lm\rangle \xi^{N_1L_1N_2L_2}_{\nu,l} \\ & = \sum_\lambda \langle L_2M_2|lm|\lambda\mu\rangle (-1)^{L_2} \\ & \times \sum_{n=0}^{\nu+N_2+1} \xi^{\nu lN_2L_2}_{n\lambda}\delta_{n,N_1}\delta_{\lambda,L_1} \end{aligned} \tag{4.15}$$

if we make use again of the orthogonality relation of the Λ functions. It can be seen from the expression Eq. (4.15) that the expansion coefficients have the following two interesting properties: In the first place, $\xi^{N_1L_1N_2L_2}_{\nu,l}$ must vanish for all ν which are smaller than

$$\min(\nu) = \max(l+1, N_1-N_2-L_2-2). \tag{4.16}$$

Therefore, the lower limit of the summation index ν in Eq. (4.13a), originally assumed to be $l+1$ as given by Eq. (4.13b), is in fact given by $\min(\nu)$ as defined by Eq. (4.16), which reduces the number of terms in the series expansion significantly. In the second place, the following symmetry relation must be valid:

$$\xi^{N_1L_1N_2L_2}_{\nu l} = (-1)^{L_2}\xi^{\nu lN_2L_2}_{N_1L_1}. \tag{4.17}$$

The relationships derived above are correct for arbitrary polynomials $p_N(r)$ as they are introduced in Eq. (4.11). If we now choose polynomials to be appropriate Laguerre polynomials,

$$p_{N_2-L_2-1}(r) = \mathcal{N}(N_2,L_2)L^{(2L_2+2)}_{N_2-L_2-1}(2r)2^{L_2}, \tag{4.18}$$

the integrals defined by Eq. (4.11) become identical with the matrix elements of the translation operator, i.e.,

$$(\mathcal{T}_{\mathbf{R}})^{M_1,M_2}_{N_1L_1,N_2L_2} = I^{L_1L_2}_{N_1,N_2-L_2-1}(\mathbf{R}). \tag{4.19}$$

Therefore, the results given by Eqs. (4.4)–(4.6) are special

cases of the formulas derived in this Sec. 4B.

Having considered some general properties of the finite series expansion Eq. (4.4), we are now in the position to derive explicit expressions for the summation coefficients T which occur in Eq. (4.4). In this derivation, the integral over Λ functions as given by Eq. (4.6), which represents the T coefficient, will be evaluated with the help of the convolution theorem of B functions as given by Eq. (4.12). After expressing the Λ functions in terms of B functions by utilizing the relationship Eq. (3.18) and performing the integration over $\mathbf{r}$ with the help of the convolution theorem Eq. (4.12), we obtain

$$\begin{aligned}\langle L_2M_2|L_1M_1|L_3M_3\rangle\ T^{N_1L_1N_2L_2}_{N_3L_3}&\\ = 4\pi \sum_{t_1t_2t_3} b^{N_1L_1}_{t_1} b^{N_2L_2}_{t_2} b^{N_3L_3}_{t_3}&\\ \times(-1)^{L_2}\sum_l \langle L_2M_2|L_1M_1|lm\rangle&\\ \times \sum_l (-1)^l \binom{\Delta l}{t} \int d\mathbf{R} B^{M\dagger}_{t_3-L_3,L_3}(\mathbf{R})&\\ \times B^{m}_{t_1+t_2-l-t,l}(\mathbf{R}).&\end{aligned} \tag{4.20}$$

The remaining integration over $\mathbf{R}$ can be performed with the help of the formula

$$\begin{aligned}\int d\mathbf{R}\, B^{M\dagger}_{N_1L_1}(\mathbf{R}) B^{M_2}_{N_2L_2}(\mathbf{R})&\\ = \delta_{L_1,L_2}\frac{(2L_1+1)!!(2N_1+2N_2+2L_1-1)!!}{(2N_1+2N_2+4L_1+2)!!}\delta_{M_1,M_2}&\end{aligned} \tag{4.21}$$

which we have derived recently.[29] If this value of the integral over the product of the two B functions is inserted into Eq. (4.20), it turns out that the summation over t can be expressed by the hypergeometric function ${}_2F_1$ with unit argument which is given by Gauss' formula[30]:

$${}_2F_1(a,b,c;1) = \Gamma(c)\Gamma(c-a-b)/[\Gamma(c-a)\Gamma(c-b)]. \tag{4.22}$$

For the T coefficients introduced in Eq. (4.4) we thus finally obtain the following expression:

$$\begin{aligned}T^{N_1L_1N_2L_2}_{N_3L_3} = 4(-1)^{\Delta L}\Gamma[\sigma+3/2]&\\ \times \sum_{t_1t_2t_3} b^{N_1L_1}_{t_1} b^{N_2L_2}_{t_2} b^{N_3L_3}_{t_3}&\\ \times \frac{\Gamma[t_1+t_2+t_3-\sigma+3/2]}{\Gamma(t_1+t_2+t_3+3)}&\end{aligned} \tag{4.23a}$$

with

$$L_i+1 \leqslant t_i \leqslant N_i, \quad i=1,2,3. \tag{4.23b}$$

This expression is identical with Eq. (4.9b). The number of terms in the summation is proportional to $(N_1-L_1-1)(N_2-L_2-1)(N_3-L_3-1)$. Therefore, this formula [Eq. (4.23) for the T coefficients] should preferably be used numerically for relatively small numbers (N_i-L_i-1).

The relationship Eq. (4.23) offers the possibility to derive one-dimensional integral representations for the T coefficient, because the gamma functions combine to a beta function[31] which may be represented by the following integral:

$$\begin{aligned}\frac{\Gamma[\sigma+3/2]\Gamma[t_1+t_2+t_3-\sigma+3/2]}{\Gamma(t_1+t_2+t_3+3)}&\\ = \int_0^1 dy\, y\left(\frac{1-y}{y}\right)^{\sigma+1/2} y^{t_1+t_2+t_3}.&\end{aligned} \tag{4.24}$$

If this expression is inserted into Eq. (4.23), under the integral sign the triple summation factorizes into the product of three sums, representing polynomials of degree N_1, N_2, and N_3, respectively:

$$\begin{aligned}T^{N_1L_1N_2L_2}_{N_3L_3} = 4\int_0^1 dy\, y\left(\frac{1-y}{y}\right)^{\sigma+1/2}\left(\sum_{t_1} b^{N_1L_1}_{t_1} y^{t_1}\right)&\\ \times\left(\sum_{t_2} b^{N_2L_2}_{t_2} y^{t_2}\right)\left(\sum_{t_3} b^{N_3L_3}_{t_3} y^{t_3}\right).&\end{aligned} \tag{4.25}$$

By inspection of the coefficients b^{NL}_t as they are given by Eq. (3.18b) it turns out that the polynomials can be written as Jacobi polynomials according to

$$\begin{aligned}2^{-N-1/2}(2N-1)!!\sum_t b^{NL}_t y^t&\\ = [(N+L+1)!(N-L-1)!\,]^{1/2} y^{L+1}&\\ \times P^{(L+3/2,L+1/2)}_{N-L-1}(1-2y).&\end{aligned} \tag{4.26}$$

If we substitute $x = 1-2y$ in Eq. (4.25) the integral representation given by Eq. (4.7) is obtained. This integral representation is closely connected to Eq. (2.11). The third integral representation as given by Eq. (4.8) can be derived at once from the last result by utilizing the relationship[32]:

$$\begin{aligned}(1-y)P^{(l+3/2,l+1/2)}_{n-l-1}(y)&\\ = \frac{(2l)!!(2n-1)!!}{2^{n-l-1}(n+l)!}&\\ \times\left[C^{(l+1)}_{n-l-1}(y) - \frac{(n-l)}{(n+l+1)}C^{(l+1)}_{n-l}(y)\right].&\end{aligned} \tag{4.27}$$

For the numerical calculation of the T coefficients it is advantageous to use another explicit expression which we are going to derive now. Again we start from Eq. (4.6). However, we do not express all Λ functions, which occur in the integrand, by B functions as we did before, but we write only $\Lambda^{M_1}_{N_1L_1}$ as a linear combination of B functions according to Eq. (3.18), obtaining

$$\begin{aligned}\langle L_2M_2|L_1M_1|L_3M_3\rangle\ T^{N_1L_1N_2L_2}_{N_3L_3}&\\ = \sum_{t=L_1+1}^{N_1} b^{N_1L_1}_t \int d\mathbf{R}\,\Lambda^{M\dagger}_{N_3L_3}(\mathbf{R}) J^{L_1L_2}_{tN_2}(\mathbf{R}),&\end{aligned} \tag{4.28}$$

where J is given by the following convolution integral:

$$J^{L_1L_2}_{tN_2}(\mathbf{R}) = \int d\mathbf{r}\, B^{M\dagger}_{t-N_1,L_1}(\mathbf{r})\Lambda^{M_2}_{N_2L_2}(\mathbf{r}-\mathbf{R}). \tag{4.29}$$

According to the discussion which led us to Eqs. (4.13) and (4.16), the integral J can be expressed as a linear combination of Λ functions,

$$\begin{aligned}J^{L_1L\dagger}_{tN_2}(-\mathbf{R}) = \sum_l \langle L_1M_1|L_2M_2|lm\rangle&\\ \times \sum_{\nu=\min(\nu)}^{\max(\nu)} a^{N_2L_2tL_1}_{\nu,l}\Lambda^{m}_{\nu,l}(\mathbf{R})&\end{aligned} \tag{4.30a}$$

with

$$\min(\nu) = \max(l+1, N_2-t-1), \tag{4.30b}$$

$$\max(\nu) = N_2 + t + 1. \tag{4.30c}$$

On the other hand, if in Eq. (4.29) the Λ function is expanded in terms of B functions with the help of Eq. (3.18), the integral J can be written in terms of B functions as

$$J^{L_1L_2\dagger}_{tN_2}(-\mathbf{R}) = 4\pi(-1)^{L_1}\sum_l \langle L_1M_1|L_2M_2|lm\rangle \times \sum_{t_2}\sum_p (-1)^p b^{N_2L_2}_{t_2}\binom{\Delta l}{p} \times B^m_{t+t_2+1-l-p,l}(\mathbf{R}). \tag{4.31}$$

The unknown expansion coefficients $a^{N_2L_2tL_1}_{\nu l}$ in Eq. (4.30a) can now be obtained as projections of the function J onto $\Lambda^m_{\nu,l}$. The integral which represents the projection can be evaluated analytically with the help of Eq. (4.31) and the relationship

$$\begin{aligned} J^{L_1L_2}_{tN_2}(0) &= \int d\mathbf{R}\, B^{M\dagger}_{t-L_1,L_1}(\mathbf{R}) \Lambda^{M_2}_{N_2L_2}(\mathbf{R}) \\ &= \delta_{L_1,L_2}\mathcal{N}^{-1}(N_2,L_2)2^{-t+1}(-1)^{N_2-L_1-1} \\ &\quad \times (t-L_1-1)!(2t+1)!! \\ &\quad \times [(t-N_2)!(t+N_2+1)!]^{-1} \\ &\qquad \text{for } t \geqslant N_2, \end{aligned} \tag{4.32a}$$

whereas

$$J^{L_1L_2}_{tN_2}(0) = 0 \text{ for } t < N_2. \tag{4.32b}$$

This can be found from the transformation formula Eq. (3.14). The explicit formula for the coefficients $a^{N_2L_2tL_1}_{\nu l}$ then reads

$$\begin{aligned} a^{N_2L_2tL_1}_{\nu l} &= 4\pi(-1)^{L_1-l+\nu-1}\mathcal{N}^{-1}(\nu,l) \\ &\quad \times \sum_{t_2,p}(-2)^p\binom{\Delta l}{p} b^{N_2L_2}_{t_2} 2^{-t-t_2} \\ &\quad \times \frac{(t_2+t-l-p)!(2t_2+2t-2p+3)!!}{(t_2+t-p+1-\nu)!(t_2+t-p+\nu+2)!} \end{aligned} \tag{4.33a}$$

with the summation limits

$$0 \leqslant p \leqslant \Delta l, \quad \nu + p - t - 1 \leqslant t_2 \leqslant N_2. \tag{4.33b}$$

If now Eq. (4.30) is used for replacing J in Eq. (4.28), the expression

$$T^{N_1L_1N_2L_2}_{N_3L_3} = \sum_{t=L_1+1}^{N_1} b^{N_1L_1}_t a^{N_2L_2tL_1}_{N_3L_3}(-1)^{L_3} \tag{4.34}$$

is obtained. From Eqs. (4.30a), (4.30b), and (4.30c) it follows that only terms with $t \geqslant |N_3 - N_2| - 1$ have to be taken into account. Then, after a few manipulations, Eq. (4.34) becomes identical with Eq. (4.9), Q.E.D.

5. SOME APPLICATIONS: ONE-CENTER EXPANSIONS AND MULTICENTER INTEGRALS

A. General aspects

The matrix representation of the translation operator makes it possible to derive normconvergent series expansion which represent new addition theorems—or one-center expansions—of three-dimensional functions $f(\mathbf{r})$. Such expansions can be used, for instance, for the evaluation of multicenter integrals. These examples for applications of the method presented in this article, which may be of general interest, shall be discussed in this section. They make it also possible to derive some useful relationships and to test the results numerically.

A relationship

$$f(\mathbf{r}-\mathbf{R}) = \sum_{n_1}\sum_{n_2} c_{n_1,n_2} g_{n_1}(\mathbf{r}) h_{n_2}(\mathbf{R}) \tag{5.1}$$

is called an addition theorem, as with the help of this formula, the two variables in the argument of the function f can be separated. If this formula, as usual, is understood as a relationship among functions which holds for any values of $\mathbf{r}$ and $\mathbf{R}$, respectively, the series expansion Eq. (5.1) is pointwise convergent. General aspects and methods for the derivation of such addition theorems for some special functions were discussed in earlier papers.[33] If the functions have certain physical meanings, there are many possible applications of such addition theorems in theoretical physics and chemistry, like, for instance, in the theory of molecular interactions, in thermodynamics,[34] and in the theory of the electronic structure of molecules and solids.

Addition theorems are especially valuable for the evaluation of generalized convolution integrals of the kind

$$\begin{aligned} &\Phi(\mathbf{R}_1,\ldots,\mathbf{R}_n) \\ &\quad = \int d\mathbf{r}\, g(\mathbf{r}) f_1(\mathbf{r}-\mathbf{R}_1) f_2(\mathbf{r}-\mathbf{R}_2)\cdots f_n(\mathbf{r}-\mathbf{R}_n). \end{aligned} \tag{5.2}$$

Integrals of this kind and of related types necessarily occur in electronic structure calculations which make use of variational principles in connection with the LCAO (linear combination of atomic oribital) method. They are called molecular multicenter integrals. The separation of the integration variable $\mathbf{r}$ with the help of an addition theorem of the type as given by Eq. (5.1) makes it possible to represent the complicated integral Eq. (5.2) by a series of simpler integrals. If applied to integration purposes, however, it is often not even necessary that the series expansion Eq. (5.1), which represents the addition theorem, is pointwise convergent. Rather it is sufficient to have an equivalent expansion in a suitable Hilbert space.

For any function f which is an element of $L_2(\mathbb{R}^3)$, the unitary translation operator $\mathcal{T}_{\mathbf{R}}$ as defined by Eq. (2.1b) causes the transformation $\mathcal{T}_{\mathbf{R}}|f\rangle = |F\rangle$ with $F(\mathbf{r}) = f(\mathbf{r}-\mathbf{R})$. Therefore, the expansion formula, which is equivalent to Eq. (5.1), reads

$$\mathcal{T}_{\mathbf{R}}|f\rangle = \sum_{n_1}\sum_{n_2} c_{n_1,n_2} h_{n_2}(\mathbf{R})|g_{n_1}\rangle. \tag{5.3}$$

Again, we have used Dirac's bra-ket notation in order to indicate that the series on the right-hand side of Eq. (5.3) is convergent in the mean, i.e.,

$$\left|\left| f(\mathbf{r}-\mathbf{R}) - \sum_{n_1}^{N_1}\sum_{n_2}^{N_2} c_{n_1,n_2} h_{n_2}(\mathbf{R}) g_{n_1}(\mathbf{r}) \right|\right| \to 0 \quad \text{for } N_1, N_2 \to \infty. \tag{5.4}$$

If the translation operator as it is given by Eq. (4.1) is applied to a function $f \in L_2(\mathbb{R}^3)$, one immediately obtains an

expansion of the type given by Eq. (5.3) in terms of the complete orthonormal system of Λ functions:

$$\mathcal{T}_{\mathbf{R}}|f\rangle = \sum_{n_1 l_1 m_1} |\Lambda^{m_1}_{n_1 l_1}\rangle \sum_{n_2 l_2 m_2} (\mathcal{T}_{\mathbf{R}})^{m_1,m_2}_{n_1 l_1, n_2 l_2} \langle \Lambda^{m_2}_{n_2 l_2}|f\rangle. \quad (5.5)$$

As the matrix elements $(\mathcal{T}_{\mathbf{R}})^{m_1,m_2}_{n_1 l_1, n_2 l_2}$ are already given in Sec. 4, we only need to evaluate the scalar products $\langle \Lambda^{m}_{nl}|f\rangle$ which are the expansion coefficients of the function f in terms of the functions Λ^{m}_{nl} according to

$$|f\rangle = \sum_{nlm} |\Lambda^{m}_{nl}\rangle \langle \Lambda^{m}_{nl}|f\rangle. \quad (5.6)$$

Hence, if these expansion coefficients are evaluated, one immediately has not only the series expansion Eq. (5.6) of the function $f \in L_2(\mathbb{R}^3)$ in terms of Λ^{m}_{nl}, but also the addition theorem Eq. (5.5) of the function f which is at least convergent in the mean.

Because of the representation of the unit operator as given by Eq. (3.6), the addition theorem Eq. (5.6) can also be written in the form

$$\mathcal{T}_{\mathbf{R}}|f\rangle = \sum_{nlm} |\Lambda^{m}_{nl}\rangle \langle \Lambda^{m}_{nl}|\mathcal{T}_{\mathbf{R}}|f\rangle. \quad (5.7)$$

If in this expansion the convolution integrals

$$\langle \Lambda^{m}_{nl}|\mathcal{T}_{\mathbf{R}}|f\rangle = \int d\mathbf{r}\, \Lambda^{m*}_{nl}(\mathbf{r}) f(\mathbf{r}-\mathbf{R}) \quad (5.8)$$

can be evaluated directly, the expansion in Eq. (5.7) has a simpler form than the expansion in Eq. (5.5). Of course, the expansion Eq. (5.5) would also result from the addition theorem Eq. (5.7) by expanding the functions f in terms of functions Λ.

B. Explicit one-center expansions for some exponential-type functions

In the case that in Eq. (5.7) the (so far unspecified) function f stands for a Λ function or a B function we have already obtained closed form expressions for the respective convolution integrals Eq. (5.8) in Sec. 4.

For $f \equiv \Lambda^{M}_{N,L}$, the convolution integrals Eq. (5.8) are just the matrix elements of the translation operator as given by Eq. (4.4). Therefore, it follows immediately from Eq. (4.1) that

$$\mathcal{T}_{\mathbf{R}}|\Lambda^{M}_{NL}\rangle = \sum_{nlm} (\mathcal{T}_{\mathbf{R}})^{m,M}_{nl,NL}|\Lambda^{m}_{nl}\rangle. \quad (5.9)$$

For $f \equiv B^{M}_{N,L}$ the convolution integral defined by Eq. (5.8) is the same as the integral given by Eq. (4.29). Therefore, we have

$$\mathcal{T}_{\mathbf{R}}|B^{M}_{N,L}\rangle = \sum_{nlm} J^{L,l*}_{N+L,n}(-\mathbf{R})|\Lambda^{m}_{nl}\rangle. \quad (5.10)$$

The last two identities, Eq. (5.9) and (5.10), are expansions which are valid with respect to the metric of the space $L_2(\mathbb{R}^3)$. The question arises what happens if the Dirac-kets are formally substituted by the values of the functions. It is clear that the resulting series are at least pointwise convergent for all points $\mathbf{r}$ apart from a set of measure zero. We shall show in the Appendix that the series expansions which are obtained in this way are pointwise convergent for *all* $\mathbf{R}, \mathbf{r}$ and, therefore, also represent new addition theorems in the classical sense.

From Eqs. (5.9) and (4.4) we thus obtain the formula

$$\Lambda^{M}_{N,L}(\mathbf{r}-\mathbf{R}) = \sum_{l_1 l_2} \sum_{m_1} \langle LM|l_1 m_1|l_2 m_2\rangle \times \sum_{n_1 n_2} T^{n_2 l_2 NL}_{n_1 l_1} \Lambda^{m_1}_{n_1 l_1}(\mathbf{r}) \Lambda^{m_2}_{n_2 l_2}(\mathbf{R}), \quad (5.11a)$$

with the summation limits

$$0 \leqslant l_1 < \infty, \quad |l_1 - L| \leqslant l_2 \leqslant l_1 + L, \quad m_2 = M - m_1, \quad (5.11b)$$

$$l + 1 \leqslant n_1 < \infty, \quad (5.11c)$$

$$\max(l_2 + 1, |n_1 - N| - 1) \leqslant n_2 \leqslant n_1 + N + 1. \quad (5.11d)$$

Explicit expressions for the expansion coefficients $T^{n_2 l_2 NL}_{n_1 l_1}$ are given by Eqs. (4.9) and (4.10).

From Eqs. (5.10) and (4.30) the following expansion is obtained:

$$B^{M}_{N,L}(\mathbf{r}-\mathbf{R}) = \sum_{l_1 l_2} \sum_{m_1} \langle LM|l_1 m_1|l_2 m_2\rangle \times \sum_{n_1 n_2} a^{n_2 l_2 N+L,L}_{n_1 l_1} \Lambda^{m_1}_{n_1 l_1}(\mathbf{r}) \Lambda^{m_2}_{n_2 l_2}(\mathbf{R}), \quad (5.12)$$

with the same summation limits as given in Eq. (5.11b). An explicit expression for the coefficients $a^{n_2 l_2 NL}_{n_1 l_1}$ is given by Eq. (4.33).

The appropriate addition theorem for Slater-type functions can be obtained from Eq. (5.11a) or (5.12) with the help of the transformation formula Eq. (3.7) or (3.10), respectively.

It should be noted that the computation of the coefficients $T^{n_2 l_2 NL}_{n_1 l_1}$ or $a^{n_2 l_2 NL}_{n_1 l_1}$, which occur in Eqs. (5.11a) and (5.12), respectively, is rather simple if Eq. (4.10) and Eq. (4.33) are used. The number of terms in the summations over l_2 and n_2 is completely determined by the fixed indices N and L. It is given by $(2L+1)(2N+2)$ for arbitrary order l_1 and n_1, and, therefore, the number of terms in the series does not increase with the order of the terms. The radial part of the Λ functions can easily be calculated by upward recursion with the help of well-known recurrence relations.[18]

The new addition theorems as given in the present paper differ from results given earlier in the literature as they exhibit a completely different representation of the radial functions in the series. Other authors who dealt with the problem of obtaining explicit addition theorems for ETF's usually considered Slater functions only.[35–37] The series expansions given by these authors are very involved and complicated in structure. This is partly due to the fact that most of these formulas exhibit a two-range form in a similar way as is well-known from the Laplace expansion of the Coulomb potential.

We can discuss the main differences between the two kinds of representation of addition theorems by comparing the new addition theorem of B functions as given by Eq. (5.12) with a pointwise convergent two-range formula which we obtained earlier by another method.[38] This latter addition theorem reads

$$B^{0}_{N,0}(\mathbf{r}-\mathbf{R}) = \sqrt{4\pi}(-1)^N [(2N)!!]^{-1} (rR)^{N-1/2} \times$$

$$\times \sum_{l=0}^{\infty} \sum_{m} \sum_{n=l}^{l+2N}{}^{(2)}(2n-2N+1)$$
$$\times T_{l,n}^{2N-1} I_{n-N+1/2}(r_<) K_{n-N+1/2}(r_>)$$
$$\times Y_l^{m*}(\Omega_r) Y_l^m(\Omega_R), \tag{5.13a}$$

where

$$r_< = \min(r,R), \quad r_> = \max(r,R). \tag{5.13b}$$

The symbol $\Sigma^{(2)}$ stands for summation in steps of 2. The formula Eq. (5.13) also represents an orthogonal expansion with respect to the angular variables. The radial dependence for any fixed value of l is given as a linear combination of products of modified Bessel functions I_ν, K_ν, and shows the characteristic two-range form. In this form, it is necessary to distinguish between $r_<$ and $r_>$. The *explicit* addition theorems for Slater functions given so far in the literature[35,36] are also complicated "two-range formulas"; in their structure, they resemble Eq. (5.13). In contrast, the new type of addition theorems as given by Eqs. (5.11) and (5.12) has the great advantage that neither complicated special functions nor a two-range behavior of the radial variables do occur, as is the case in the two-range addition theorems.

C. On the evaluation of multicenter integrals

Addition theorems are valuable tools for the evaluation of (generalized) convolution integrals or multicenter integrals, i.e., integrals which resemble those defined by Eq. (5.2), because the addition theorems make it possible to separate the variables in the integrands. However, the newly derived addition theorems, which have a uniform mathematical representation over the whole range of variables, are often of greater advantage for the the evaluation of integrals than the addition theorems which exhibit a two-range form, containing certain special functions of argument $r_<$ and $r_>$. If addition theorems of the two-range form are used to transform the integrand, the integration range is to be divided into different subregions due to the occurrence of $r_<$ and $r_>$. This finally requires the evaluation of certain indefinite integrals over special functions. It is a fact, however, that only a few indefinite integrals over special functions are known, whereas a lot of integrals over the entire space, i.e., integrals from 0 to ∞, are available from standard integration tables.

In the addition theorem as given by Eqs. (5.10) and (5.12) the radial dependence is given as an infinite series in terms of rather simple functions. As this series expansion holds for the whole region (between 0 and ∞) of r and R, one has no longer to distinguish between $r_>$ and $r_<$. Therefore, if we use this expansion in order to express a (shifted) function in the integrand of a generalized convolution integral we obtain an infinite series of integrals over the entire space. Usually these integrals can be evaluated much more easily than the integrals over subregions which occur if a two-range addition theorem is used. Therefore, with this method it will often be possible to evaluate generalized convolution integrals as a series expansion even if the evaluation of these integrals with the help of the first method (based upon a two-range addition theorem) is too complicated.

As an example and as a test for the numerical applicability of the new series expansions, we want to use the appropriate relationships for the evaluation of the following rather difficult multicenter integral:

$$\Delta(\mathbf{R}_1,\mathbf{R}_2) = \int d\mathbf{r}_1 \int d\mathbf{r}_2\, e^{-r_1} e^{-r_2} \times |\mathbf{r}_1 - \mathbf{r}_1|^{-1} e^{-|\mathbf{r}_1 - \mathbf{R}_1|} e^{-|\mathbf{r}_2 - \mathbf{R}_2|}. \tag{5.14}$$

In molecular physics integrals of this kind are called three-center exchange integrals. They occur, for instance, in calculations of properties of molecules with three or more atoms. Because of the identity

$$e^{-r} = \pi^{1/2} \Lambda_{1,0}^{0}(\mathbf{r}), \tag{5.15}$$

it is possible to use Eqs. (5.9) and (5.11) in order to represent the two shifted exponential functions in that coordinate system which is used for the integration. We thus get at once the following series expansion for the integral

$$\Delta(R_1,R_2) = \sum_{n_1 l_1 m_1} \sum_{n_2 l_2 m_2} c_{n_1 n_2}^{l_1 l_2} (\mathcal{T}_{\mathbf{R}_1})_{n_1 l_1, 1,0}^{m_1,0} (\mathcal{T}_{\mathbf{R}_2})_{n_2 l_2, 1,0}^{m_2,0}, \tag{5.16}$$

with

$$c_{n_1 n_2}^{l_1 l_2} = \int d\mathbf{r}_1 \int d\mathbf{r}_2\, e^{-r_1} e^{-r_2} |\mathbf{r}_1 - \mathbf{r}_2|^{-1} \times \Lambda_{n_1 l_1}^{m_1}(\mathbf{r}_1) \Lambda_{n_2 l_2}^{m_2}(\mathbf{r}_2) \pi. \tag{5.17}$$

This one-center integral can be evaluated analytically with the help of the Laplace expansion of the Coulomb operator. The coefficients then read

$$c_{n_1 n_2}^{l_1 l_2} = \delta_{l_1,l_2} 2^{-n_1-n_2-1} (n_1+n_2-2)! \times [n_1+n_2-(n_1-n_2)^2+(2l_1+2)(2l_1+4)] \times [(n_1+l_1+1)!(n_2+l_2+1)!(n_1-l_1-1)! \times (n_2-l_2-1)!]^{-1/2}. \tag{5.18}$$

Having substituted the matrix elements of the translation operator in Eq. (5.16), we obtain the three-center exchange integral as a double infinite series of Λ functions and simple coefficients.

A numerical analysis for the region $0 \leqslant R_1, R_2 < 3$ has led us to the conclusion that an accuracy of 7–8 significant figures can be obtained if the series expansion Eq. (5.16) is summed up to order $n_1, n_2 = 60$. For comparison purposes, tabulated results for the special case $\mathbf{R}_1 = \mathbf{R}_2$, for which it is possible to solve the integral by other methods,[39] have been used. Because of the simple structure of the formulas for the matrix elements of the translation operator it causes no difficulty to include higher order terms to get even more precise results if they are needed.

6. SUMMARY

In Eqs. (4.1) and (4.4) of this article, we have given a matrix representation of the translation operator which can be used to describe the shifting of physical fields, if these fields are mathematically represented by functions of the Hilbert space $L_2(\mathbb{R}^3)$. All elements of the matrix representing the translation operator are given analytically as a linear combination of simple functions of the displacement vector $\mathbf{R}$. If the translation operator in the realization that is given by Eqs. (4.1) and (4.4) is applied to a function which describes the original field, one immediately obtains the new function which describes the transformed, i.e. shifted, field, in form of

a series expansion with respect to the complete and orthonormal system of Λ functions; cf. Eq. (5.5). All that remains to do, is to expand the function describing the untransformed, i.e. original, field, into a series of Λ functions, as stated in Eq. (5.6). The Λ functions, which are closely related to Laguerre functions, are defined by Eq. (3.4).

Using the matrix representation of the translation operator, we have derived new addition theorems in Sec. 5B, and presented a method for the evaluation of generalized convolution integrals in Sec. 5C.

APPENDIX: POINTWISE CONVERGENCE OF THE ADDITION THEOREMS

The expansion as given in Eq. (5.12), which is at least convergent in the mean, can be written in the form

$$B^M_{N,L}(\mathbf{r}-\mathbf{R}) = \sum_{l_1=0}^{\infty} \sum_{m_1=-l_1}^{l_1} \bar{R}_{l_1}(r) Y^{m_1}_{l_1}(\Omega_r), \tag{A1}$$

with

$$\bar{R}_{l_1}(r) = \sum_{n_1=l_1+1}^{\infty} c^{l_1}_{n_1} L^{(2l_1+2)}_{n_1-l_1-1}(2r) e^{-r} r^{l_1}. \tag{A2}$$

Here we consider B as a function of $\mathbf{r}$; the vector $\mathbf{R}$ is an arbitrary but fixed parameter which shall be suppressed in the following equations.

On the other hand, a pointwise convergent addition theorem for the B function also does exist.[40] It can be written as

$$B^M_{N,L}(\mathbf{r}-\mathbf{R}) = \sum_{l_1=0}^{\infty} \sum_{m_1=-l_1}^{l_1} R_{l_1}(r) Y^{m_1}_{l_1}(\Omega_r). \tag{A3}$$

The radial function $R_{l_1}(r)$ is a continuous function for $0<r<\infty$ and has the asymptotic behavior

$$R_{l_1}(r) \to r^{N+L-1} e^{-r} \quad \text{for } r\to\infty, \tag{A4a}$$

$$R_{l_1}(r) \to r^{l_1} \quad \text{for } r\to 0. \tag{A4b}$$

These properties can be found in the following manner. Multiply both sides of Eq. (5.13) with a regular solid spherical harmonic $\mathscr{Y}^M_L(\mathbf{r}-\mathbf{R})$ and use the addition theorem for this function[41] to separate the variables $\mathbf{r}$ and $\mathbf{R}$. With the help of the well-known relationship

$$Y^m_l(\Omega) Y^M_L(\Omega) = \sum_{\lambda} \langle \lambda\mu | lm | LM \rangle \, Y^{\mu}_{\lambda}(\Omega), \tag{A5}$$

the products of spherical harmonics can be expressed as single harmonics. Now use the asymptotic properties of the modified Bessel functions[42]

$$K_\nu(r) \to r^{-1/2} e^{-r} \quad \text{for } r\to\infty, \tag{A6a}$$

$$I_\nu(r) \to r^{\nu} \quad \text{for } r\to 0 \tag{A6b}$$

to obtain Eqs. (A4a) and (A4b). Since the series on the r.h.s. of Eq. (A1) is at least convergent in the mean, it follows that the function $\bar{R}_{l_1}(r)$ must be the formal Laguerre expansion of the function $R_{l_1}(r)$, i.e.

$$R_{l_1}(r) \sim \sum_{n_1=l_1+1}^{\infty} c^{l_1}_{n_1} L^{(2l_1+2)}_{n_1-l_1-1}(2r) e^{-r} r^{l_1}. \tag{A7}$$

Now we can apply Szegö's equiconvergence theorem[43] which states under which conditions a formal Laguerre series of a given function does also converge point by point.

Due to Szegö, the limit relation

$$\lim_{N\to\infty} \left[e^{x/2} x^{-\lambda} g^{(x)}_N - \int_{x^{1/2}-\delta}^{x^{1/2}+\delta} dt\, g(t^2) t^{-2\lambda} e^{t^2/2} \phi_N(x^{1/2}-t) \right] = 0 \tag{A8}$$

with

$$\phi_N(y) = (\pi y)^{-1} \cdot \sin(2N^{1/2} y) \tag{A9}$$

holds for $x>0$, if the following prepositions are fulfilled:

The partial sums $g_N(x)$ are defined by

$$g_N = \sum_{n=0}^{N} a_n x^{\lambda} e^{-x/2} L^{(\alpha)}_n(x), \tag{A10}$$

where the coefficients a_N are obtained by the orthogonality relation of the Laguerre polynomials,

$$a_N = \left[\Gamma(\alpha+1) \binom{n+\alpha}{n} \right]^{-1} \times \int_0^{\infty} dx\, x^{\alpha-\lambda} e^{-x/2} L^{(\alpha)}_n(x) g(x). \tag{A11}$$

Furthermore, the function $g(x)$ must have the following integral properties: First,

$$e^{x/2} x^{-\lambda} g(x) \tag{A12}$$

is Lebesgue measurable for $x\in[0,\infty]$. Second, the integrals

$$\int_0^1 dx\, x^{\alpha-\lambda} e^{x/2} |g(x)|, \quad \int_0^1 dx\, x^{\alpha/2-\lambda-1/4} |g(x)| e^{x/2} \tag{A13}$$

exist. Third, the following asymptotic relationship is valid:

$$\int_n^{\infty} dx\, x^{\alpha/2-\lambda-1/4} |g(x)| = O(n^{-1/2}) \quad \text{for } n\to\infty. \tag{A14}$$

We notice the validity of the distributional relationship

$$\lim_{N\to\infty} \phi_N(y) = \delta(y), \tag{A15}$$

if the function $g(s)$ has the property that

$$\int_0^{\infty} dx'\, g(x') \delta(x-x') = g(x) \tag{A16}$$

is well defined for all $x>0$. Then it follows from Szegö's theorem Eqs. (A8)–(A14) that the sequence of the partial sums $g_N(x)$ does converge in a pointwise sense, which means that

$$\lim_{N\to\infty} g_N(x) = g(x) \tag{A17}$$

holds for any $x>0$.

In order to apply Szegö's considerations to our problem, we have to define the function $g(x)$ and the parameters λ and α adequately. If we set

$$g(x) = 2^{-l} R_l(x/2), \quad \lambda = l, \quad \alpha = 2l+2, \tag{A18}$$

and if we use Eqs. (A6a) and (A6b), then it is easy to prove that the conditions given in Eqs. (A8)–(A16) are fulfilled. Therefore, it must be true that the expansion Eq. (A7) is pointwise convergent for all $r>0$ and arbitrary R. Taking into account the symmetry of the series expansion as given in

Eq. (5.12) with respect to an interchange of **r** and **R**, we are led to the conclusion that the expansion is pointwise convergent for the whole region of the argument vectors. Since the Laguerre and Slater functions are given as simple linear combinations of B functions according to Sec. 2, the appropriate new addition theorems for these functions must also be pointwise convergent, Q.E.D.

[1]N. J. Vilenkin, *Special Functions and the Theory of Group Representations* (Amer. Math. Soc., Providence, 1968), Sec. 6.
[2]N. I. Achieser and I. M. Glasmann, *Theorie der linearen Operatoren im Hilbert–Raum* (Akademie, Berlin, 1968), Sec. 42.
[3]J. C. Browne, Adv. At. Mol. Phys. **7**, 47 (1971).
[4]K. G. Kay and H. J. Silverstone, J. Chem. Phys. **53**, 4269 (1970), and references therein.
[5]See Ref. 2, Secs. 29 and 40. For the matrix representation of unbounded operators, see also G. Epifanio and C. Trapani, J. Math. Phys. **20**, 148 (1979).
[6]M. Tinkham, *Group Theory and Quantum Mechanics* (McGraw–Hill, New York, 1964); M. E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1967).
[7]E. Filter and E. O. Steinborn, J. Math. Phys. **19**, 79 (1978).
[8]See, e.g., Ref. 1, Secs. 2 and 3.
[9]J. Arsac, *Fourier Transforms and the Theory of Distributions* (Prentice Hall, New York, 1966) Sec. 3.4.
[10]R. Askey, *Orthogonal Polynomials and Special Functions* (SIAM, Philadelphia, 1975), Lect. 5.
[11]E. U. Condon and G. H. Shortley, *Theory of Atomic Spectra* (Cambridge U.P., London, 1953), p. 52.
[12]H. A. Bethe and E. E. Salpeter, *Encyclopedia of Physics* XXXV (Springer, Berlin, 1957), pp. 88 ff.
[13]B. Klahn and W. A. Bingel, Theor. Chim. Acta (Berlin) **44**, 9, 27 (1977).
[14]E. Filter and E. O. Steinborn, Phys. Rev. A **18**, 1 (1978).
[15]In the present paper the definitions of Ref. 14 are used, which are slightly different from the definitions of Ref. 7.
[16]G. N. Watson, *Theory of Bessel Functions* (Cambridge U.P., London, 1966), p. 78.
[17]E. O. Steinborn and E. J. Weniger, Int. J. Quantum Chem. Symp. **11**, 509 (1977); **12**, 103 (1978).
[18]W. Magnus, E. Oberhettinger, and R. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Springer, New York, 1966), p. 240.
[19]P. O. Löwdin and H. Shull, Phys. Rev. **101**, 1730 (1956).
[20]See Ref. 18, p. 249.
[21]See Ref. 7, Eqs. (2.11) and (2.12).
[22]See Ref. 18, p. 241.
[23]E. J. Weniger (unpublished).
[24]For brevity, functions $A^{m}_{n,l}$, $B^{m}_{n,l}$, etc. will be written as A^{m}_{nl}, B^{m}_{nl}, ..., respectively, if there is no possibility for an ambiguity. Equivalently, the indices of the coefficients $T^{N_1L_1N_2L_2}_{N_3L_3}$ as defined by Eq. (4.4), will be separated by commas only if necessary, as is the case when an index is given by a sum of numbers.
[25]See Ref. 7, Eq. (2.13).
[26]See Ref. 18, p. 211.
[27]See Ref. 18, p. 218.
[28]See Ref. 14, Eqs. (4.2), (4.3), and (2.5).
[29]See Ref. 14, Eq. (4.5).
[30]See Ref. 18, p. 40.
[31]See Ref. 18, p. 7.
[32]See Ref. 18, p. 213.
[33]E. O. Steinborn and E. Filter, Theor. Chim. Acta (Berlin) **38**, 247 261, 281 (1975).
[34]J. O. Hirschfelder, *Intermolecular Forces* Vol. 12 of *Advances in Chemical Physics* edited by I. Prigogine and S. Rice (Interscience, New York, 1967).
[35]H. J. Silverstone, J. Chem. Phys. 47, 537 (1967); K. Ruedenberg, Theor. Chim. Acta (Berlin) **7**, 359 (1967).
[36]R. R. Sharma, Phys. Rev. A **13**, 517 (1976).
[37]I. I. Guseinov, J. Chem. Phys. **69**, 4990 (1978).
[38]See Ref. 33, p. 276, Eq. (3.4).
[39]H. Preuss, *Integraltafeln zur Quantenchemie* (Springer, Berlin, 1956–1960), Vols. I–V.
[40]E. Filter and E. O. Steinborn (unpublished).
[41]E. O. Steinborn and K. Ruedenberg, Adv. Quantum Chem. **7**, 1, 64 (1973).
[42]See Ref. 18, p. 139.
[43]G. Szegö, *Orthogonal Polynomials* (Amer. Math. Soc., Colloquium, Providence, R.I., 1975), Vol. 13, p. 246, theorem (9.1.5).

Infinities of polynomial conserved densities for nonlinear evolution equations

Mark J. McGuinness
Department of Mathematical Physics, University College Dublin, Dublin, Republic of Ireland

(Received 3 May 1979; accepted for publication 9 November 1979)

The infinities of polynomial conserved densities for several nonlinear evolution equations are investigated using Noether's theorem, and are identified as energy or momentum densities of higher-order enveloping equations. A recursive operator formula is derived for the densities.

1. INTRODUCTION

In a recent paper,[1] the infinity of conserved densities for the Korteweg–de Vries (KdV) equation was investigated using Noether's theorem. In that paper, these densities were identified via Noether's theorem as *energy* densities, not of the KdV equation but of higher-order enveloping KdV equations. By "enveloping" it is meant that the solution set of the KdV equation is contained by the solution set of each higher order KdV equation.

This paper presents an extension of that result to the modified KdV equations, the sine-Gordon equation, the classical nonlinear shallow-water (CNSW) equations, and the nonlinear Schrödinger (NLS) equations. For these equations (and the KdV equation) it has been found that the polynomial conserved densities may be identified either as canonical *energy* densities or as canonical *momentum* densities of the appropriate higher-order equations. This identification has been made by using Noether's theorem in both its conventional and its generalized form.[1]

For each of these nonlinear equations, the higher-order enveloping equations have taken the form of an integrodifferential operator operating n times on the original equation. They are integrodifferential, and are not known to be of any physical significance or interest in themselves. Their main property of interest in this paper is that their solution sets contain that of the original nonlinear evolution equation being considered.

This work differs from other techniques which use invariance groups and symmetry groups[2] in the following manner: These other techniques use an infinity of symmetries on the nonlinear evolution equation, whereas this work uses *one* symmetry (time or space translation invariance) on an infinity of nonlinear equations. The main advantage of this approach is felt to be the ease of interpretation of the results. The two approaches may well be parallel ones since, for example, the integrodifferential operators used in this paper to generate the infinity of nonlinear equations for the sine-Gordon, the KdV, and the modified KdV equations are the *same* operators as the "recursion operators" used by Olver.[3]

The integrodifferential operators used are closely related to the operator[4]

$$L^{+}=\frac{1}{2i}\begin{pmatrix} d_x-2r\int q, & 2r\int r \\ -2q\int q, & -d_x+2q\int r \end{pmatrix} \tag{1.1}$$

used by Ablowitz *et al.*[5] The operator L^{+} appears in their derivation of a general set of nonlinear equations which are solvable by the inverse scattering transform method,

$$\begin{pmatrix} r_t \\ -q_t \end{pmatrix}+2A_0(L^{+})\begin{pmatrix} r \\ q \end{pmatrix}=0, \tag{1.2}$$

where A_0 is a ratio of entire functions. The results of this paper will be extended to the general set of equations (1.2) in a future paper.

One interesting and novel feature in all cases has been the splitting of the problem into two parts. One is an *integrodifferential* part involving integrals over space of polynomials in the field variables and their derivatives, and the other is a *partial differential* part, involving polynomials in the field variables and their spatial derivatives (in all of the cases considered, there is one space dimension, x).

The integrodifferential part is proved inductively by using the generalized Noether's theorem,[1] and amounts to proving a type of anticommutation principle for the appropriate operator. The partial differential part is proved by deriving a Lagrangian density, using the work of Atherton and Homsy[6] on the inverse problem of the calculus of variations. An inductive approach is used to prove the Lagrangian density exists in each case, and the conventional Noether's theorem completes the proof of this part. *As a bonus, this approach gives each of the infinite number of conserved densities explicitly in terms of the integrodifferential operator.* With the exception of the KdV equation,[7] the author is not aware that such a form for the conserved densities has been previously derived.

Section 2 of this paper outlines the general method used in all five cases. The steps involved in each case are quite similar. Section 3 gives more details for each of the five equations mentioned above.

2. A GENERAL OUTLINE OF THE METHOD USED

It has been possible to identify the infinite sets of conserved densities arising in certain nonlinear systems, as energy or momentum densities of higher-order enveloping nonlinear systems. Let the original nonlinear system have the equation of motion

$$F\overset{\circ}{=}0, \tag{2.1}$$

(where $\overset{\circ}{=}$ means "equals for solutions," to be distinguished from "equals for all values of the field variables," the usual $=$), and the infinite number of conservation laws (we are dealing with only one spatial dimension)

 0022-2488/80/122737-06$1.00

$$d_t T_n + d_x X_n \overset{\circ}{=} 0, \quad n = 0,1,2,\cdots, \tag{2.2}$$

where T_n is a polynomial in the field variables and their derivatives. Then the higher-order enveloping equations are written

$$K^n(F) \overset{\circ}{=} 0, \quad n = 1,2,\cdots, \tag{2.3}$$

where K is a nonlinear integrodifferential operator, and the superscript n is a power.

The relationship which has been proved for five different systems is that

$$\phi_t [K^n(F)] = d_t T_{n+1} + d_x X_{n+1}, \quad n = 0,1,\cdots, \tag{2.4}$$

where

$$\phi_x \equiv d_x \phi \equiv \frac{d\phi}{dx},$$

where ϕ is the field variable, here written as a scalar for simplicity.

For these systems it has also been found that

$$\phi_x [K^n(F)] = d_t T_n + d_x X'_n, \quad n = 0,1,\cdots. \tag{2.5}$$

By the generalized Noether's theorem,[1] Eq. (2.5) identifies T_n as a *momentum* density[8] for Eq. (2.3), and Eq. (2.4) identifies T_{n+1} as an *energy* density for Eq. (2.3). Note that the solution sets of Eqs. (2.3) contain that of the original equation (2.1), hence *solutions to the original equation* (2.1) *must obey the momentum and energy conservation laws for Eqs.* (2.3). With the exception of the sine-Gordon equation, all of the nonlinear evolution equations considered can be written in the form

$$F = \phi_{at} + K(\phi_{a+1}) \overset{\circ}{=} 0, \quad a = 0 \text{ or } 1, \tag{2.6}$$

where

$$\phi_a \equiv \frac{d^a \phi}{dx^a}.$$

The field variable is here written as a scalar for simplicity; the results for vector fields are quite analogous, as will be seen in Sec. 3. Despite the fact that the sine-Gordon equation cannot be written in the form (2.6), the proof of Eq. (2.4) for the sine-Gordon equation is similar to the outline presented here.

In the cases where $a = 1$, the evolution equation has been of the form

$$F = \frac{d}{dx}[\phi_t + L(\phi_2)] \overset{\circ}{=} 0. \tag{2.7}$$

This property ensures that a polynomial conserved density may be required to contain only x-derivatives of the field variables, since all t-derivatives can be replaced by x-derivatives within a trivially conserved density, using Eq. (2.7).

The proof of Eq. (2.4) will be outlined here, since the proof of Eq. (2.5) follows in all cases considered. It is required to prove that

$$\phi_t K^n [\phi_{at} + K(\phi_{a+1})] = d_t T_{n+1} + d_x X_{n+1}, \tag{2.8}$$

where

$$a = 0 \text{ or } 1, \; n = 0,1,2,\cdots.$$

It is at this stage that the problem splits into two parts, the integrodifferential part

$$\phi_t K^n(\phi_{at}) = d_x X'_{n+1} \tag{2.9}$$

and the partial differential part

$$\phi_t K^{n+1}(\phi_{a+1}) = d_t T_{n+1} + d_x (X_{n+1} - X'_{n+1}). \tag{2.10}$$

This division is motivated by the fact that if the existence of a polynomial conserved density T_n depending only on x-derivatives of ϕ is assumed, the lhs of Eq. (2.9) cannot contribute any terms to the density T_n. Note that the equality in Eq. (2.8) must hold for general field variables ϕ, not just for solutions to the evolution equation.

If the operator K is integrodifferential in x, the lhs of Eq. (2.9) must also be integrodifferential in x. The lhs of Eq. (2.10) has, in all cases considered, been partial differential (i.e., polynomial in ϕ and its derivatives). As will be seen at the end of this section, this feature is closely related to the existence of an infinity of *polynomial* conserved densities.

Equation (2.9) is proved by deriving an anticommutation relation for the operator K,

$$fK^n(g_a) = -gK^n(f_a) + d_x R(g,f), \tag{2.11}$$

where f and g are test functions of ϕ and its derivatives. The proof of Eq. (2.11) has in all cases been straightforward and inductive. $R(g,f)$ contains integral terms, which in each case need to be shown to be *acceptable* flux terms when $g = f = \phi_t$. Integral terms are acceptable flux terms if they are equal to polynomial terms within a trivial flux term.

Equation (2.10) is proved by deriving a Lagrangian density for the equation

$$K^i(\phi_{a+1}) = 0, \quad i = 1,2,\cdots. \tag{2.12}$$

The first step in the derivation is to show that the lhs of Eq. (2.12) is a polynomial in ϕ and its x-derivatives. This is done by induction, using the anticommutation relation (2.11). The work of Atherton and Homsy[6] gives the Lagrangian density for Eq. (2.12) as

$$\mathcal{L}_i = \phi \int_0^1 \chi_i(\lambda\phi)\, d\lambda, \tag{2.13}$$

where

$$\chi_i(\phi) = K^i(\phi_{a+1}), \tag{2.14}$$

and χ_i is a function, not an operator. This Lagrangian density exists if and only if the Frechet derivative of the lhs of Eq. (2.12) is symmetric, i.e.,

$$\psi(\chi_i)'_\phi(\sigma) = \sigma(\chi_i)'_\phi(\psi) + d_x V, \tag{2.15}$$

for arbitrary test functions V, ψ, and σ of ϕ, where the Frechet differential in the direction σ is[6]

$$(\chi_i)'_\phi(\sigma) \equiv \lim_{\epsilon \to 0} [\chi_i(\phi + \epsilon\sigma) - \chi_i(\phi)]/\epsilon. \tag{2.16}$$

In each case, Eq. (2.15) is proved by induction. Note that V must contain only acceptable flux terms.

Once the Lagranian density (2.13) has been proved to exist, the conventional Noether's theorem[1] can be applied as follows: Since $\mathcal{L}_i$ has no explicit time dependence, energy is conserved in the corresponding equation of motion (2.12), and the energy density is given by

$$T_i = -\frac{\partial \mathcal{L}_i}{\partial \phi_t}\phi_t + \mathcal{L}_i, \tag{2.17}$$

that is,

$$T_i = \mathscr{L}_i = \phi \int_0^1 \chi_i(\lambda\phi)\, d\lambda\,. \tag{2.18}$$

Equation (2.10) follows immediately [it is simply Noether's relation, Eq. (2.5) in Ref. 1], with T_i as above. Equation (2.18) gives each of the infinity of conserved densities explicitly in terms of the integrodifferential operator K, as mentioned at the end of Sec. 1. If

$$K^i(\phi_{a+1}) = \chi_i(\phi) \tag{2.19}$$

is a polynomial, Eq. (2.18) implies that T_i is polynomial also.

3. PARTICULAR CASES

A. The KdV equation

For completeness, a brief outline of the results already published for the KdV equation[1] is given here. The form of the KdV equation used is

$$\phi_{1t} + \phi_1\phi_2 + \phi_4 \stackrel{\circ}{=} 0\,, \tag{3.1}$$

where

$$\phi_1 \equiv \frac{\partial\phi}{\partial x}, \quad \phi_2 \equiv \frac{\partial^2\phi}{\partial x^2}, \quad \phi_t \equiv \frac{\partial\phi}{\partial t}, \quad \text{etc.}$$

(A useful survey of recent work on the KdV equation has been published by Miura[9].) The integrodifferential operator for Eq. (3.1) is

$$N = d_x^2 + \tfrac{2}{3}\phi_1 + \tfrac{1}{3}\phi_2 \int^x dx\,. \tag{3.2}$$

The lower limit of the integral is chosen such that ϕ and its derivatives vanish there (e.g., $-\infty$). The limits on all integrals in this paper, unless stated otherwise, will be as above. Equation (3.1) may be written

$$\phi_{1t} + H(\phi_2) \stackrel{\circ}{=} 0\,. \tag{3.3}$$

The relation proved in a previous paper[1] is

$$\phi_t H^n[\phi_{1t} + H(\phi_2)] = d_t T_{n+1} + d_x X_{n+1}, \quad n = 0,1,2,\cdots, \tag{3.4}$$

where the densities T_k are polynomials in ϕ and its x-derivatives, that is, the density T_{n+1} is an energy density of the equation

$$H^n[\phi_{1t} + H(\phi_2)] \stackrel{\circ}{=} 0\,. \tag{3.5}$$

It is a straightforward corollary that T_n is a momentum density of Eq. (3.4) and is conserved, i.e.,

$$\phi_1 H^n[\phi_{1t} + H(\phi_2)] = d_t T_n + d_x X_n'\,, \tag{3.6}$$

as follows: The integrodifferential part is

$$\phi_1 H^n(\phi_{1t}) = d_x R_n - \phi_t H^n\phi_2\,, \tag{3.7}$$

where

$$\begin{aligned} R_n = & \sum_{i=0}^{n-1}\left[\left(\int H^i\phi_2\right)\left(\int H^{n-i-1}\phi_{1t}\right)\right. \\ & - (H^i\phi_2)(H^{n-i-1}\phi_{1t}) \\ & - \tfrac{1}{3}\left(\int \phi_1 H^i\phi_2\right)\left(\int H^{n-i-1}\phi_{1t}\right) \\ & \left. - \tfrac{1}{3}\left(\int \phi_1 H^i\phi_2\right)\left(\int \phi_1 H^{n-i-1}\phi_{1t}\right)\right] + \phi_t \int H^n\phi_2\,. \end{aligned} \tag{3.8}$$

In Ref. 1 R_n is proved to contain only acceptable flux terms. The last term in Eq. (3.7) can be written[1]

$$-\phi_t H^n(\phi_2) = d_t T_n + d_x X_n\,, \tag{3.9}$$

so that the integrodifferential part is

$$\phi_1 H^n(\phi_{1t}) = d_t T_n + d_x(X_n + R_n)\,. \tag{3.10}$$

A similar approach to that of Eq. (3.7) gives the partial differential part as

$$\phi_1 H^{n+1}(\phi_2) = d_x Q_n\,, \tag{3.11}$$

where Q_n is a polynomial in ϕ and its x-derivatives. This result also follows from the recursion formula[10]

$$H d_x A_n = d_x A_{n+1}\,, \quad A_0 = \phi_1\,, \tag{3.12}$$

where A_n is a polynomial conserved density for the KdV equation.

B. The modified KdV equation

The form of the modified KdV equation used is

$$\phi_{1t} + \phi_1^2\phi_2 + \phi_4 \stackrel{\circ}{=} 0\,. \tag{3.13}$$

The operator is

$$M = d_x^2 + \tfrac{2}{3}\phi_1^2 + \tfrac{2}{3}\phi_2 \int \phi_1\,, \tag{3.14}$$

and Eq. (3.13) may be written

$$\phi_{1t} + M\phi_2 \stackrel{\circ}{=} 0\,. \tag{3.15}$$

The proof of the integrodifferential part of Eq. (2.5),

$$\phi_t M^n\phi_{1t} = d_x X'_{n+1}\,, \tag{3.16}$$

is accomplished by first proving the anticommutation relation

$$fM^n g_1 = -gM^n f_1 + d_x R_n(f,g)\,, \tag{3.17}$$

where

$$\begin{aligned} R_n(f,g) \equiv \sum_{i=0}^{n-1}\Bigg[& \left(\int M^i f_1\right)\left(\int M^{n-i-1}g_1\right) \\ & - (M^i f_1)(M^{n-i-1}g_1) - \tfrac{2}{3}\left(\int \phi_1 M^i f_1\right) \\ & \times\left(\int \phi_1 M^{n-i-1}g_1\right) + g\int M^n f_1\Bigg]\,. \end{aligned} \tag{3.18}$$

The proof of Eq. (3.17) is omitted as it is quite straightforward. R_n contains integral terms, which must be shown to be acceptable flux terms for $f = g = \phi_t$. Note that

$$M^k(\phi_{1t}) = -M^{k+1}(\phi_2) + M^k(\phi_{1t} + M\phi_2)\,, \tag{3.19}$$

where the last term in Eq. (3.19) is a trivial flux term since it is zero for solutions. Hence if

$$\phi_1 M^{n+1}(\phi_2) = d_x P_{n+1}\,, \quad n = 0,1,2,\cdots\,, \tag{3.20}$$

where P_k is a polynomial in ϕ and its derivatives, then the integral

$$\int \phi_1 M^n\phi_{1t}\,, \quad n = 0,1,2,\cdots\,, \tag{3.21}$$

will be equivalent to the polynomial P_n, and will be an acceptable flux term. Also, since

$$M^n\phi_{1t} = d_x\left(d_x + \tfrac{2}{3}\phi_1\int\phi_1\right)M^{n-1}\phi_{1t}\,, \tag{3.22}$$

the term

$$\int M^k\phi_{1t} = \left(d_x + \tfrac{2}{3}\phi_1 \int \phi_1\right) M^{k-1}\phi_{1t}, \quad k = 1,2,\cdots, \tag{3.23}$$

will be an acceptable flux term.

Equation (3.20) may be proved by induction, using result (3.18). The proof is straightforward and is omitted.

The partial differential part of Eq. (2.4) is

$$\phi_t M^{n+1}\phi_2 = d_t T_{n+1} + d_x(X_{n+1} - X'_{n+1}). \tag{3.24}$$

The lhs is a polynomial by virtue of Eq. (3.20). Existence of a Lagrangian for the equation

$$M^n\phi_2 \overset{\circ}{=} 0 \tag{3.25}$$

is assured by the symmetry of the Frechet derivative,[6]

$$\psi(M^n\phi_2)'_\phi(\sigma) \simeq \sigma(M^n\phi_2)'_\phi(\psi), \tag{3.26}$$

where $\simeq$ means "equals within an acceptable x-derivative." Equation (3.26) is proved by induction, as outlined here.

Assume Eq. (3.26) holds for $n = 1,2,\ldots,k$. Then

$$\psi(M^{k+1}\phi_2)'_\phi(\sigma) = \psi M(M^k\phi_2)'_\phi(\sigma) + \psi(M)'_\phi(\sigma)(M^k\phi_2), \tag{3.27}$$

where

$$(M)'_\phi(\sigma) = \tfrac{4}{3}\phi_1\sigma_1 + \tfrac{2}{3}\phi_2\int\sigma_1 + \tfrac{2}{3}\sigma_2\int\phi_1. \tag{3.28}$$

Using the explicit form of M, Eq. (3.27) can be written

$$\begin{aligned}&\psi(M^{k+1}\phi_2)'_\phi(\sigma)\\ &\simeq\left(\int M\psi_1\right)(M^k\phi_2)'_\phi(\sigma) + \tfrac{2}{3}(\phi_1M^k\phi_2)\int\phi_1\psi_1\\ &\quad - \tfrac{2}{3}\psi_1\sigma_1\left(\int\phi_1M^k\phi_2\right).\end{aligned} \tag{3.29}$$

The inductive assumption gives

$$\left(\int M\psi_1\right)(M^k\phi_2)'_\phi(\sigma) \simeq \sigma(M^k\phi_2)'_\phi\left(\int M\psi_1\right), \tag{3.30}$$

and applying Eq. (3.29) to the rhs of Eq. (3.30) gives

$$\begin{aligned}&\psi(M^{k+1}\phi_2)'_\phi(\sigma)\\ &\simeq\left(\int M\sigma_1\right)(M^{k-1}\phi_2)'_\phi\left(\int M\psi_1\right) - \tfrac{2}{3}\psi_1\sigma_1\left(\int\phi_1M^k\phi_2\right)\\ &\quad + \tfrac{2}{3}(\sigma_1M^k\phi_2)\int\phi_1\psi_1 + \tfrac{2}{3}(M\psi_1)(M^{k-1}\phi_2)\int\phi_1\sigma_1\\ &\quad + \tfrac{2}{3}\sigma_1(M\psi_1)\left(\int\phi_1M^{k-1}\phi_2\right).\end{aligned} \tag{3.31}$$

The first term on the rhs of Eq. (3.31) is symmetric by the inductive assumption, the second term is symmetric by inspection, and the remaining terms may be shown to be so within an x-derivative by expanding them.

As explained in Sec. 2, the Lagrangian density for Eq. (3.25) is

$$\mathscr{L}_n = \phi\int_0^1 \mathscr{M}_n(\lambda\phi)\,d\lambda, \tag{3.32}$$

where

$$\mathscr{M}_n(\phi) = M^n\phi_2. \tag{3.33}$$

The Noether relation expressing conservation of energy for Eq. (3.25) is

$$\phi_t M^n\phi_2 = d_t T_n + d_x(X_n - X'_n), \tag{3.34}$$

i.e., Eq. (3.24) is proved with

$$T_{n+1} = \mathscr{L}_{n+1}, \quad n = 0,1,2,\cdots. \tag{3.35}$$

Note that T_{n+1} is a polynomial, and that Eq. (3.34) holds for all ϕ, not merely for solutions to Eq. (3.25).

The uniqueness of the infinity of polynomial conserved densities for the KdV equation (Kruskal *et al.*[11]), together with the transformation due to Miura[12] from solutions of the modified KdV equation to solutions of the KdV equation, implies that the infinity of polynomial conserved densities for the modified KdV equation is also unique. Hence the set of the energy densities T_n is equivalent to the set exhibited by Miura *et al.*[13]

The density T_n may be alternatively identified (within a trivial sign) as the momentum density of the equation

$$M^n(\phi_{1t} + M\phi_2) \overset{\circ}{=} 0, \tag{3.36}$$

by proving the relation

$$\phi_1M^n(\phi_{1t} + M\phi_2) = d_t(-T_n) + d_x\bar{X}_n, \quad n = 0,1,2,\cdots. \tag{3.37}$$

The partial differential part of Eq. (3.36) has been proved at Eq. (3.20). The integrodifferential part is proved by applying Eq. (3.17) with

$$f = \phi_1, \quad g = \phi_t,$$

to get

$$\phi_1M^n\phi_{1t} = -\phi_tM^n\phi_2 + d_xR_n(\phi_1,\phi_t), \tag{3.38}$$

where R_n contains acceptable flux terms. Equation (3.24) may be applied to the first term on the rhs of Eq. (3.38) to complete the integrodifferential part,

$$\phi_1M^n\phi_{1t} = d_t(-T_n) + d_x(X'_n - X_n + R_n). \tag{3.39}$$

C. The sine-Gordon equation

The form of the sine-Gordon equation used is

$$\phi_{1t} + \sin\phi \overset{\circ}{=} 0. \tag{3.40}$$

The operator is

$$S = d_x^2 + \phi_1^2 + \phi_2\int\phi_1, \tag{3.41}$$

and noting that

$$S(\sin\phi) = \phi_2[\cos\phi]_{\phi=0} = \phi_2, \tag{3.42}$$

the higher-order sine-Gordon equations may be written

$$S^n(\phi_{1t} + S^{-1}\phi_2) \overset{\circ}{=} 0, \quad n > 0. \tag{3.43}$$

Since the only difference between the operator S and the operator M for the modified KdV equation is a factor of $\frac{2}{3}$ in the nonlinear terms, and since the higher-order modified KdV equations are

$$M^n(\phi_{1t} + M\phi_2) \overset{\circ}{=} 0, \tag{3.44}$$

the proof of the energy relation

$$\phi_tS^n(\phi_{1t} + S^{-1}\phi_2) = d_tT_{n+1} + d_xX_{n+1}, \quad n = 0,1,\cdots \tag{3.45}$$

is identical to that for the modified KdV equation, and is

omitted here. The energy density of the nth higher-order sine-Gordon equation is given by

$$T_{n+1} = \phi \int_0^1 \mathscr{S}_{n-1}(\lambda\phi)\, d\lambda\,, \tag{3.46}$$

where the function $\mathscr{S}_{n-1}$ is given by

$$\mathscr{S}_{n-1}(\phi) = S^{n-1}\phi_2\,. \tag{3.47}$$

Similarly, the proof of the momentum relation

$$\phi_x S^n(\phi_{1t} + S^{-1}\phi_2) = d_t T_n + d_x X_n, \quad n = 0,1,\cdots \tag{3.48}$$

is identical to that for the modified KdV equation.

The similarity between the operators S and M indicates a similarity between the conserved densities of the sine-Gordon and the modified KdV equation.[14] The transformation $\phi \to \pm\sqrt{\tfrac{1}{2}}\phi$ takes operator M to operator S, and takes the infinite set of conserved densities for the modified KdV equation to that for the sine-Gordon equation.

The first three densities T_n have been found to be equivalent to the first three of the set of densities derived for the sine-Gordon equation by Lamb,[15] and Sanuki and Konno.[16]

D. The classical nonlinear shallow-water equations

The CNSW equations govern the irrotational motion of an inviscid homogenous fluid under gravity, in the long wave approximation.[17] They may be written in the form

$$\underline{\sigma}\underline{\phi}_{xt} + \underline{I}\underline{\sigma}\underline{\phi}_{xx} \overset{\circ}{=} 0\,, \tag{3.49}$$

where

$$\underline{\phi}_x \equiv \begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}_x \equiv \begin{pmatrix}u\\ h\end{pmatrix},$$

$$\underline{I} \equiv \begin{pmatrix} \tfrac{1}{2}\phi_{1x}\,, & \phi_{2x} + \tfrac{1}{2}\phi_{2xx}\int \\ 1\,, & \tfrac{1}{2}\phi_{1x} + \tfrac{1}{2}\phi_{1xx}\int \end{pmatrix} \tag{3.50}$$

and

$$\underline{\sigma} \equiv \begin{pmatrix}0, & 1\\ 1, & 0\end{pmatrix} \text{ (a Pauli spin matrix)},$$

and where $h(x,t)$ is the free surface height, $u(x,t)$ is the horizontal velocity component, and g is taken as 1. Note that the numeral subscripts here refer to vector components and not to x-derivatives.

The steps involved in proving the relation

$$\underline{\phi}_t \cdot [\underline{I}^n(\underline{\sigma}\underline{\phi}_{xt} + \underline{I}\underline{\sigma}\underline{\phi}_{xx})] = d_t T_{n+1} + d_x X_{n+1}\,, \quad n = 1,2,\cdots \tag{3.51}$$

(where " $\cdot$ " is the scalar product between vectors), are a generalization of those for the modified KdV equation to the vector case. Hence only a brief outline will be given here.

The relation analogous to relation (3.17) is

$$\underline{f}\cdot(\underline{I}^n\underline{\sigma}\underline{g}_x) = -\underline{g}\cdot(\underline{I}^n\underline{\sigma}\underline{f}_x) + d_x R_n(\underline{f},\underline{g})\,, \tag{3.52}$$

where

$$R_n(\underline{f},\underline{g}) = \sum_{i=0}^{n-1} \tfrac{1}{2}\left[0, \underline{\phi}_x\cdot\left(\int \underline{I}^i\underline{\sigma}\underline{f}_x\right)\right]\cdot\int (\underline{I}^{n-i-1}\underline{\sigma}\underline{g}_x)$$
$$- \tfrac{1}{2}\left[0, \int \underline{\phi}_x\cdot(\underline{I}^i\underline{\sigma}\underline{f}_x)\right]\cdot\int (\underline{I}^{n-i-1}\underline{\sigma}\underline{g}_x)$$
$$+ \underline{g}\cdot\int (\underline{I}^n\underline{\sigma}\underline{f}_x)\,. \tag{3.53}$$

The analogous result to Eq. (3.20) is

$$\underline{I}^n\underline{\sigma}\underline{\phi}_{xx} = d_x\underline{P}_n\,, \quad n = 0,1,2,\cdots, \tag{3.54}$$

where $\underline{P}_n$ is a vector whose elements are polynomials in $\underline{\phi}$ and its x-derivatives.

For the partial differential part, the equation analogous to Eq. (3.31) is

$$\underline{\rho}\cdot(\underline{I}^n\underline{\sigma}\underline{\phi}_{xx})'_{\underline{\phi}}(\underline{\psi}) \simeq \left(\int \underline{\sigma}\underline{I}\underline{\sigma}\underline{\rho}_x\right)\cdot(\underline{I}^{n-2}\underline{\sigma}\underline{\phi}_{xx})'_{\underline{\phi}}\left(\int \underline{\sigma}\underline{I}\underline{\sigma}\underline{\psi}_x\right)$$
$$+ \underline{\rho}\cdot(\underline{I})'_{\underline{\phi}}(\underline{\psi})(\underline{I}^{n-1}\underline{\sigma}\underline{\phi}_{xx})$$
$$+ \underline{\psi}\cdot(\underline{I})'_{\underline{\phi}}\left(\int \underline{\sigma}\underline{I}\underline{\sigma}\underline{\rho}_x\right)(\underline{I}^{n-2}\underline{\sigma}\underline{\phi}_{xx})\,. \tag{3.55}$$

The energy densities of the higher-order CNSW equations are given by

$$T_n = \underline{\phi}\cdot\int_0^1 \underline{\mathscr{I}}_n(\lambda\underline{\phi})\, d\lambda\,. \tag{3.56}$$

where $\underline{\mathscr{I}}_n$ is a function defined by

$$\underline{\mathscr{I}}_n(\underline{\phi}) = \underline{I}^n\underline{\sigma}\underline{\phi}_{xx}\,. \tag{3.57}$$

Benney[17] has derived an infinite number of conserved densities for the case of nonzero vorticity. These reduce to an infinite set for Eqs. (3.49) if the motion is required to be irrotational. The first four densities T_n in Eq. (3.56) have been found to be equivalent to the first four derived from Benney's work.

In the same manner as for the modified KdV equation, these densities may be alternatively identified as momentum densities of the higher-order CNSW equations, i.e., it may be proved that

$$\underline{\phi}_x\cdot[\underline{I}^n(\underline{\sigma}\underline{\phi}_{xt} + \underline{I}\underline{\sigma}\underline{\phi}_{xx})] = d_t T_n + d_x X_n\,. \tag{3.58}$$

E. The nonlinear Schrödinger equations

The nonlinear Schrödinger (NLS) equations may be written in the form

$$-i\underline{\tau}\underline{\phi}^*_t + \underline{N}\underline{\tau}\underline{\phi}^*_x \overset{\circ}{=} 0\,, \tag{3.59}$$

where

$$\underline{\phi} = \begin{pmatrix}\phi\\ \phi^*\end{pmatrix}, \underline{\phi}^* = \begin{pmatrix}\phi^*\\ \phi\end{pmatrix},$$

$$\underline{\tau} = \begin{pmatrix}1, & 0\\ 0, & -1\end{pmatrix} \text{ (a Pauli spin matrix)},$$

and

$$\underline{N} = \begin{pmatrix} d_x + \phi^*\int\phi, & -\phi^*\int\phi^* \\ \phi\int\phi, & -d_x - \phi\int\phi^* \end{pmatrix}. \tag{3.60}$$

Note that

$$\underline{N}\underline{g} = (d_x\underline{\tau} + \underline{\phi}^*\int\underline{\phi}\cdot\underline{\tau})\underline{g}\,. \tag{3.61}$$

The relation identifying an infinity of conserved densities as energy densities of an infinity of higher-order NLS equations,

$$\underline{\phi}_t \cdot [\underline{N}^n(-i\underline{\tau}\underline{\phi}_t^* + \underline{N}\underline{\tau}\underline{\phi}_x^*)] = d_t T_{n+1} + d_x X_{n+1}, \quad (3.62)$$

is proved in the same way as that for the modified KdV equation, generalized to the vector case.

In the integrodifferential part, the anticommutation relation analogous to Eq. (3.17) is

$$\underline{f} \cdot (\underline{N}^n \underline{\tau}\underline{g}^*) = -\underline{g} \cdot (\underline{N}^n \underline{\tau}\underline{f}^*) + d_x Y_n(\underline{f},\underline{g}), \quad (3.63)$$

where

$$Y_n(\underline{f},\underline{g}) = \sum_{i=0}^{n-1} (\underline{\tau}\underline{N}^{*i}\underline{\tau}\underline{f}) \cdot (\underline{\tau}\underline{N}^{n-i-1}\underline{\tau}\underline{g}) + \int (\underline{\tau}\underline{N}^i \underline{\tau}\underline{f}^*) \cdot \underline{\phi} \int (\underline{\tau}\underline{N}^{n-i-1}\underline{\tau}\underline{g}^*) \cdot \underline{\phi}. \quad (3.64)$$

The analogous result to Eq. (3.20) is

$$\underline{N}^n \underline{\tau}\underline{\phi}_x^* = P_n, \quad n = 0,1,\cdots. \quad (3.65)$$

In proving Eq. (3.65), it should be noted that

$$\underline{N}^k g = d_x^2(\underline{N}^{k-2} g) + 2\underline{\phi}^* [\underline{\phi} \cdot (\underline{N}^{k-2} g)] + (\underline{\tau}\underline{\phi}_x) \int \underline{\phi} \cdot (\underline{\tau}\underline{N}^{k-2} g) - \underline{\phi}^* \int \underline{\phi}_x \cdot (\underline{N}^{k-2} g). \quad (3.66)$$

The second-last term in Eq. (3.66) will be a polynomial by the inductive assumption, and the last term in that equation may be dealt with using relation (3.63).

In the partial differential part, the equation analogous to Eq. (3.31) is

$$\begin{aligned}
&\underline{\psi} \cdot (\underline{N}^n \underline{\tau}\underline{\phi}_x^*)'_{\underline{\phi}}(\underline{\rho}) \\
&\simeq \left(\underline{\tau}\underline{\rho}_x + \underline{\tau}\underline{\phi} \int \underline{\phi}^* \cdot \underline{\rho}\right)(\underline{N}^{n-2} \underline{\tau}\underline{\phi}_x^*)'_{\underline{\phi}} \left(\underline{\tau}\underline{\psi}_x + \underline{\tau}\underline{\phi} \int \underline{\phi}^* \cdot \underline{\psi}\right) \\
&\quad + \underline{\psi} \cdot \underline{\rho}^* \int (\underline{\tau}\underline{\phi}) \cdot (\underline{N}^{n-1} \underline{\tau}\underline{\phi}_x^*) \\
&\quad - (\underline{\tau}\underline{\rho}) \cdot (\underline{N}^{n-1} \underline{\tau}\underline{\phi}_x^*) \int \underline{\psi} \cdot \underline{\phi}^* \\
&\quad - \underline{\rho} \cdot \left(\underline{\tau}\underline{\psi}_x^* + \underline{\tau}\underline{\phi}^* \int \underline{\phi}^* \cdot \underline{\psi}\right) \int (\underline{\tau}\underline{\phi}) \cdot (\underline{N}^{n-2} \underline{\tau}\underline{\phi}_x^*) \\
&\quad + (\underline{\tau}\underline{\psi}_x + \underline{\tau}\underline{\phi} \int \underline{\phi}^* \cdot \underline{\psi}) \cdot (\underline{\tau}\underline{N}^{n-2} \underline{\tau}\underline{\phi}_x^*) \int \underline{\rho} \cdot \underline{\phi}^*. \quad (3.67)
\end{aligned}$$

The energy densities of the higher-order NLS equations

$$\underline{N}^n(-i\underline{\tau}\underline{\phi}_x^* + \underline{N}\underline{\tau}\underline{\phi}_x^*) \stackrel{\circ}{=} 0, \quad n = 1,2,\cdots \quad (3.68)$$

are

$$T_n = \underline{\phi} \cdot \int_0^1 \underline{\mathscr{N}}_n(\lambda \underline{\phi})\, d\lambda, \quad (3.69)$$

where the vector function $\underline{\mathscr{N}}_n$ is defined by

$$\underline{\mathscr{N}}_n(\underline{\phi}) \equiv \underline{N}^n \underline{\tau}\underline{\phi}_x^*, \quad n = 0,1,2,\cdots. \quad (3.70)$$

The first five of these densities have been found to be equivalent to the five densities presented by Zakharov and Shabat,[18] from the infinity of conserved densities they derive for the NLS equations.

As in the previous cases, it is straightforward to show that these polynomial conserved densities may be alternatively identified as momentum densities of Eq. (3.68), i.e.,

$$\underline{\phi}_x \cdot [\underline{N}^n(-i\underline{\tau}\underline{\phi}_x^* + \underline{N}\underline{\tau}\underline{\phi}_x^*)] = d_t T_n + d_x X'_n. \quad (3.71)$$

A further result for the NLS equation is that the densities T_n may also be associated with an infinitesimal gauge transformation of the first kind, by the relation

$$(i\underline{\tau}\underline{\phi}) \cdot [\underline{N}^n(-i\underline{\tau}\underline{\phi}_x^* + \underline{N}\underline{\tau}\underline{\phi}_x^*)] = d_t T_{n-1} + d_x X''_{n-1}. \quad (3.72)$$

A gauge transformation of the first kind is given by

$$\phi \to \phi e^{i\epsilon}, \quad \phi^* \to \phi^* e^{-i\epsilon}, \quad (3.73)$$

so that for infinitesimal ϵ,

$$\delta\phi = ie\phi, \quad \delta\phi^* = -i\epsilon\phi^*, \quad (3.74)$$

that is,

$$\delta\underline{\phi} = i\epsilon\underline{\tau}\underline{\phi}. \quad (3.75)$$

The integrodifferential part of Eq. (3.72) is proved by noting that

$$(i\underline{\tau}\underline{\phi}) \cdot [\underline{N}^n(-i\underline{\tau}\underline{\phi}_x^*)] \simeq -\underline{\phi}_t \cdot (\underline{N}^n \underline{\phi}^*), \quad (3.76)$$

using Eq. (3.63), and that

$$\underline{N}\underline{\phi}^* = \underline{\tau}\underline{\phi}_x^*, \quad (3.77)$$

and by using the partial differential part of Eq. (3.62).

The partial differential part of equation (3.72) is proved by using Eqs. (3.65) and (3.61). Note that in the case of the NLS equation, invariance (of the action integral) under the gauges transformation (3.75) implies conservation of the number of particles,

$$\int_{-\infty}^{\infty} \phi\phi^*\, dx.$$

ACKNOWLEDGMENTS

I am grateful to Professor W.L. Jones for his guidance in this work, to Dr. J.D. Gibbon for proofreading this paper, and to the Department of Education in the Republic of Ireland for the Research Fellowship that has enabled me to complete this paper.

[1] M.J. McGuinness, J. Math. Phys. **19**, 2285 (1978).
[2] See, for example, N.H. Ibraghimov, Lett. Math. Phys. **1**, 423 (1977); S. Kumei, J. Math. Phys. **16**, 2461 (1975).
[3] P.J. Olver, J. Math. Phys. **18**, 1212 (1977).
[4] All integrals in this paper, unless stated otherwise, are from some boundary on which the field variable and its derivatives vanish (e.g., $x = -\infty$), to x.
[5] M.J. Ablowitz *et al.*, Stud. Appl. Math. **53**, 249 (1974).
[6] R.W. Atherton and C.M. Homsy, Stud. Appl. Math. **54**, 31 (1975).
[7] C.S. Gardner *et al.*, Commun. Appl. Math. **27**, 132 (1974).
[8] Since $\bar{\delta}\phi \equiv \phi'(x) - \phi(x) = -\epsilon\phi_x$, i.e., $\delta x = \epsilon$, $\delta t = 0$, $\delta\phi \equiv \phi'(x') - \phi(x) = 0$, i.e., an infinitesimal x-translation.
[9] R.M. Miura, Siam. Rev. **18**, 412 (1976).
[10] C.S. Gardner *et al.*, Commun. Appl. Math. **27**, 97 (1974).
[11] M.D. Kruskal *et al.*, J. Math. Phys. **11**, 952 (1970).
[12] R.M. Miura, J. Math. Phys. **9**, 1202 (1968).
[13] R.M. Miura *et al.*, J. Math. Phys. **9**, 1204 (1968).
[14] Compare the work of P.J. Olver, "Symmetry Groups and Conservation Laws in the Formal Variational Calculus" (preprint), in which is noted the same correspondence between the "recursion operators" for the modified KdV and the sine-Gordon equations.
[15] G.L. Lamb, Jr., Rev. Mod. Phys. **43**, 99 (1971).
[16] H. Sanuki and K. Konno, Phys. Lett A **48**, 221 (1974).
[17] D.J. Benney, Stud. Appl. Math. **52**, 45 (1973).
[18] V.E. Zakharov and A.B. Shabat, JETP **34**, 62 (1972); [Zh. Eksp. Teor. Fiz. **61**, 118–34 (1971).]

An infinity of polynomial conserved densities for a class of nonlinear evolution equations

Mark J. McGuinness
Department of Mathematical Physics, University College, Belfield, Dublin 4, Ireland

(Received 21 September 1979; accepted for publication 15 November 1979)

Noether's theorem is applied to the infinity of polynomial conserved densities possessed by a general class of nonlinear evolution equations. The densities are identified on the solution sets of higher-order enveloping equations as canonical energy or momentum densities, and a new recursive formula is derived for these densities.

1. INTRODUCTION

An infinity of polynominal conserved densities (p.c.d.'s) $\{T_n\}$ is derived for the class of nonlinear evolution equations

$$\mathbf{F}_0 \equiv \boldsymbol{\sigma}\boldsymbol{\tau}\mathbf{r}_t + 2A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r} \doteq 0\,, \tag{1.1}$$

where A_0 is entire in its argument,

$$\mathsf{L} \equiv \frac{1}{2i}\begin{pmatrix} -d_x + 2q\int_{-\infty}^{x} r\,dx, & -2q\int_{-\infty}^{x} q\,dx \\ 2r\int_{-\infty}^{x} r\,dx, & d_x - 2r\int_{-\infty}^{x} q\,dx \end{pmatrix}, \tag{1.2}$$

and

$$\boldsymbol{\sigma} \equiv \begin{pmatrix} 0, & 1 \\ 1, & 0 \end{pmatrix}, \quad \mathbf{r} \equiv \begin{pmatrix} r \\ q \end{pmatrix},$$

$$\boldsymbol{\tau} \equiv \begin{pmatrix} 1, & 0 \\ 0, & -1 \end{pmatrix}, \quad d_x \equiv \frac{d}{dx},$$

$$\mathbf{r}_t \equiv \partial\mathbf{r}/\partial t\,, \quad \mathbf{r}_x \equiv \partial\mathbf{r}/\partial x\,,$$

and $\doteq$ means "equals for solutions", and where the field variable $\mathbf{r}$ and its derivatives are assumed to vanish on the lower boundary $[x = -\infty]$ of the integrals. The derivation extends to the more general case that A_0 is a ratio of entire functions, under certain assumptions, discussed at the end of Sec. 2.

The class of equations (1.1) was shown by Ablowitz *et al.*[1] to be solvable by the inverse scattering transform, and is already known to possess an infinity of p.c.d.'s[1,2] However, the derivation presented here uses the Lagrangian formalism and Noether's theorem, which has some advantages when dealing with conservation laws. What is perhaps the main point of this paper is interpretative: that each of the infinity of p.c.d.'s T_n is identified as a canonical *energy* or *momentum* density of the higher-order enveloping equation

$$\mathbf{F}_k \equiv \mathsf{L}^k\mathbf{F}_0 \doteq 0\,, \tag{1.3}$$

where $k = n - 1$ for a momentum density, and $k < n - 1$ for an energy density. The energy result is proved in Sec. 2, and the momentum result follows in Sec. 3.

The solution sets of equations (1.3), for $k > 0$, contain the solution set of equations (1.1), so that the T_n are identified as energy or momentum densities on *enveloping* solution sets. The enveloping equations (1.3) are integrodifferential, and are of no known interest in themselves.

The other advantage of the Lagrangian approach used here is the derivation of a recursion formula for the T_n, which to the best of the author's knowledge is original:

$$T_n = \mathbf{r}\cdot\int_0^1 \mathbf{L}_n(\lambda\mathbf{r})\,d\lambda\,, \quad n = 0, 1, \cdots, \tag{1.4}$$

where

$$\mathbf{L}_n(\mathbf{r}) = \mathsf{L}\mathbf{L}_{n-1}(\mathbf{r})\,,$$

and

$$\mathbf{L}_0(\mathbf{r}) = 2A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r} \quad \text{(for energy densities)} \tag{1.5}$$

or

$$\mathbf{L}_0(\mathbf{r}) = \boldsymbol{\sigma}\mathbf{r} \quad \text{(for momentum densities)}\,. \tag{1.6}$$

Note that the integral over λ merely introduces a different constant factor for each term in the polynomial $\mathbf{L}_n$. The formula (1.4) differs from the algebraic formula of Konno *et al.*,[2] which has no integral, and expresses T_n in terms of all previous T_i, $i < n$. They derive two sets of conserved densities $\phi_1^{(i)}$ and $\phi_2^{(i)}$ and for $i \leqslant 4$ this author finds that $\phi_1^{(i)}$ is equivalent to $\phi_2^{(i)}$, and is also equivalent to the T_i obtained from Eqs. (1.4) and (1.6). Hence, the two formulas are likely to be equivalent.

The set of equations (1.1) contains the Korteweg–de Vries, the modified Korteweg–de Vries, the Sine–Gordon, and the nonlinear Schrödinger equations. Hence, the results of this paper contain many of the results of a previous paper,[3] excluding the case of the classical nonlinear shallow-water equations. Appendix A contains a short note relating the operator L to the operators used in that paper. Appendix B applies the analysis to linear equations, with the result that any existing conservation law for the equation

$$F \doteq 0 \tag{1.7}$$

gives rise to an infinity of conservation laws for that equation, each of the infinity being identified via Noether's theorem on the enveloping solution set of the equation

$$d_x^{2n}(F) \doteq 0\,, \quad n = 1, 2, \ldots \tag{1.8}$$

in the same way as the original density is identified on the original solution set [that of Eq. (1.7)].

Appendix C shows that if $r = \pm q^*$ in the nonlinear equations (1.1), an infinite set of conserved densities corresponding to gauge invariance of the first kind may also be derived. This set coincides with the set of energy or momentum densities (1.4), so that if $r = \pm q^*$ we have an alternative identification of that set. This leads to the speculation that perhaps the same may be true for the nonlinear equations (1.1) as a proved in Appendix B for linear equations,

 0022-2488/80/122743-06$1.00

that any existing conserved density leads to an infinite set of conserved densities.

2. THE ENERGY DENSITIES OF THE ENVELOPING EQUATIONS

An infinity of energy densities will be derived, one for each of the enveloping equations

$$\mathbf{G}_n \equiv \mathsf{L}^n[\sigma\tau\mathbf{r}_t + \mathsf{L}^m\sigma\mathbf{r}] \doteq , \quad n = 0, 1, \ldots , \tag{2.1}$$

where m is some positive integer. This infinite set of energy densities will then constitute an infinity of conserved densities for the equation $\mathbf{G}_0 \doteq 0$, since the solution set of that equation is contained by the solution sets of the equations $\mathbf{G}_n \doteq 0$, $n > 0$.

The energy relation identifying T_n as an energy density of the nth equation (2.1) is

$$\mathbf{r}_t \cdot \mathbf{G}_n = d_t T_n + d_x X_n , \tag{2.2}$$

since this associates T_n with invariance of the action integral (if it exists) under the infinitesimal time translation $\bar{\delta}\mathbf{r} = -\epsilon\,\mathbf{r}_t$, where ϵ is an infinitesimal parameter (Ref. 5, Sec. 2). Equation (2.2) will be proved to hold for all positive integers m, n, and a recursive formula for T_n will be obtained from the Lagrangian formalism used in the proof.

The proof of Eq. (2.2) is attempted in two parts: the *partial differential* part

$$\mathbf{r}_t \cdot (\mathsf{L}^{n+m}\sigma\mathbf{r}) = d_t T_n + d_x X_n' , \tag{2.3}$$

and the *integrodifferential* part

$$\mathbf{r}_t \cdot (\mathsf{L}^n\sigma\tau\mathbf{r}_t) = d_x (X_n - X_n') . \tag{2.4}$$

The following anticommutation relation will be useful in both parts:

Lemma: For arbitrary vector functions $\mathbf{f}$ and $\mathbf{g}$ of $\mathbf{r}$ and its derivatives,

$$\mathbf{f}\cdot(\mathsf{L}^n\tau\sigma\mathbf{g}) = -\mathbf{g}\cdot(\mathsf{L}^n\tau\sigma\mathbf{f}) + \frac{d}{dx} W_n(\mathbf{f}, \mathbf{g}) , \tag{2.5}$$

where

$$W_n(\mathbf{f}, \mathbf{g}) \equiv \sum_{i=0}^{n-1} \{(\sigma\tau\mathsf{L}^i\tau\sigma\mathbf{f})\cdot(\tau\mathsf{L}^{n-i-1}\tau\sigma\mathbf{g}) + 2\left[\int_{-\infty}^{x} \mathbf{r}\cdot(\tau\mathsf{L}^i\tau\sigma\mathbf{f})\, dx\right] \times \left[\int_{-\infty}^{x} \mathbf{r}\cdot(\tau\mathsf{L}^{n-i-1}\tau\sigma\mathbf{g})\, dx\right]\} . \tag{2.6}$$

Proof: The proof is inductive, straightforward, and omitted.

A. The partial differential part

Application of lemma (2.5) implies that the lhs of Eq. (2.3) is a polynomial in $\mathbf{r}$ and its x derivatives (hence the name "partial differential part") as follows:

The operator L may be written

$$2i\mathsf{L} = -\tau d_x + 2(\sigma\mathbf{r})\int_{-\infty}^{x} \mathbf{r}\cdot\tau \tag{2.1.1}$$

when it operates on some vector. Hence, $\mathsf{L}^k\sigma\mathbf{r}$ is partial differential if

$$\mathbf{r}\cdot(\tau\mathsf{L}^{k-1}\sigma\mathbf{r}) = \frac{d}{dx} P_{k-1} , \tag{2.1.2}$$

where P_{k-1} is partial differential.

Assume Eq. (2.1.2) holds for all positive $k \leqslant n$. Then

$$\mathbf{r}\cdot(\tau\mathsf{L}^n\sigma\mathbf{r}) = -(\tau\mathbf{r})\cdot[\mathsf{L}^n\tau\sigma(\tau\mathbf{r})] , \tag{2.1.3}$$

$$= (\tau\mathbf{r})\cdot[\mathsf{L}^n\tau\sigma(\tau\mathbf{r})] - d_x W_n(\tau\mathbf{r}, \tau\mathbf{r}) , \tag{2.1.4}$$

using the properties

$$\mathbf{a}\cdot(\tau\mathbf{b}) = (\tau\mathbf{a})\cdot\mathbf{b} , \quad \mathbf{a}\cdot(\sigma\mathbf{b}) = (\sigma\mathbf{a})\cdot\mathbf{b} , \quad \sigma\sigma = \tau\tau = \text{identity} , \quad \sigma\tau = -\tau\sigma . \tag{2.1.5}$$

Hence,

$$\mathbf{r}\cdot(\tau\mathsf{L}^n\sigma\mathbf{r}) = -\tfrac{1}{2}\frac{d}{dx} W_n(\tau\mathbf{r}, \tau\mathbf{r}) , \tag{2.1.6}$$

where W_n is a polynomial by the inductive assumption. Since

$$\mathbf{r}\cdot(\tau\mathsf{L}\sigma\mathbf{r}) = \frac{d}{dx}(-rq) , \tag{2.1.7}$$

the induction is started and Eq. (2.1.6) is proved for all $n > 0$. Note that if $\mathsf{L}^k\sigma\mathbf{r}$ is partial differential for some $k < 0$, then $\mathsf{L}^n\sigma\mathbf{r}$ is partial differential for all $n \geqslant k$.

Since there are no integral terms on the lhs of Eq. (2.3), a Lagrangian density will be derived for the equation

$$\mathsf{L}^k\sigma\mathbf{r} = 0 , \quad k = m, m+1, \ldots , \tag{2.1.8}$$

and conventional Noether's theorem will be applied to that Lagrangian density to obtain Eq. (2.3).

According to Atherton and Homsy,[5] a Lagrangian density exists for Eq. (2.1.8) if the Frechet derivative of the lhs is symmetric, i.e., if for arbitrary vector functions ψ and ρ,

$$\psi\cdot(\mathbf{L}_k)_{\mathbf{r}}'(\rho) \simeq \rho\cdot(\mathbf{L}_k)_{\mathbf{r}}'(\psi) , \tag{2.1.9}$$

where

$$\mathbf{L}_k = \mathsf{L}\mathbf{L}_{k-1}$$

and

$$\mathbf{L}_0 = \mathsf{L}^m\sigma\mathbf{r} , \tag{2.1.10}$$

where $\simeq$ means equals within the x derivative of a function of $\mathbf{r}$ and its derivatives. This function must vanish when evaluated on some boundary at which its arguments vanish. The Frechet differential in the direction ρ of a vector function $\mathbf{F}$ of the vector $\mathbf{u}$ is

$$(\mathbf{F})_{\mathbf{u}}'(\rho) = \lim_{\epsilon\to 0} \frac{\mathbf{F}(\mathbf{u} + \epsilon\rho) - F(\mathbf{u})}{\epsilon} , \tag{2.1.11}$$

and if G is a polynomial in ϕ and its derivatives,

$$G_{\phi}'(\rho) = \sum_{i=0}^{\infty} \frac{\partial G}{\partial \phi_i}\left(\frac{d}{dx}\right)^i \rho , \quad \phi_i \equiv \frac{d^i\phi}{dx^i} . \tag{2.1.12}$$

Equation (2.1.9) will be proved by induction, assuming it holds for all positive $k \leqslant n$. Then

$$\psi\cdot(\mathbf{L}_{n+1})_{\mathbf{r}}'(\rho) = \psi\cdot(\mathsf{L})_{\mathbf{r}}'(\rho)(\mathbf{L}_n) + \psi\cdot[\mathsf{L}(\mathbf{L}_n)_{\mathbf{r}}'(\rho)] , \tag{2.1.13}$$

where

$$2i(\mathsf{L})_{\mathbf{r}}'(\rho) = 2\sigma\rho\int \mathbf{r}\cdot\tau + 2\sigma\mathbf{r}\int \rho\cdot\tau . \tag{2.1.14}$$

With some manipulation and application of the inductive assumption, Eq. (2.1.14) becomes

$$\psi\cdot(\mathbf{L}_{n+1})'_t(\rho) \simeq (\sigma\tau\mathsf{L}\tau\sigma\rho)\cdot(\mathbf{L}_{n-1})'_t(\sigma\tau\mathsf{L}\tau\sigma\psi) + [\rho\cdot(\sigma\psi) + \psi\cdot(\sigma\rho)]\int \mathbf{r}\cdot(\tau\mathbf{L}_n)$$
$$-2\rho\cdot(\tau\mathbf{L}_n)\int \psi\cdot(\sigma\tau) + 2\rho\cdot(\tau\mathsf{L}\tau\sigma\psi)\int \mathbf{r}\cdot(\tau\mathbf{L}_{n-1}) - 2(\sigma\tau\mathsf{L}\tau\sigma\psi)\cdot(\tau\mathbf{L}_{n-1})\int \rho\cdot(\sigma\tau)\,. \tag{2.1.15}$$

The first term on the rhs is symmetric by the inductive assumption, the second term is clearly symmetric, and the remaining terms may be explicitly shown to be symmetric.

Since for $m = 0$,

$$2i\psi\cdot(\mathbf{L})'_t(\rho) \simeq \tfrac{1}{2}\psi\cdot(\sigma\tau\rho_x) + \tfrac{1}{2}\rho\cdot(\sigma\tau\psi_x) \tag{2.1.16}$$

and

$$(2i)^2\psi\cdot(\mathbf{L}_2)'_t(\rho) \simeq -\psi_x\cdot(\sigma\rho_x) - 4qr\psi\cdot(\sigma\rho) - 2q^2\rho^{(1)}\psi^{(1)} - 2r^2\rho^{(2)}\psi^{(2)}\,, \tag{2.1.17}$$

where

$$\rho = \begin{pmatrix}\rho^{(1)}\\ \rho^{(2)}\end{pmatrix}, \qquad \psi = \begin{pmatrix}\psi^{(1)}\\ \psi^{(2)}\end{pmatrix},$$

the induction is started and Eq. (2.1.9) is proved for all $k, m \geqslant 0$.

Hence a Lagrangian density exists for equation (2.1.8) and is given by[5]

$$\mathscr{L}_k = \mathbf{r}\cdot\int_0^1 \mathbf{L}_k(\lambda\mathbf{r})\,d\lambda\,. \tag{2.1.18}$$

Since the Lagrangian density has no explicit time dependence, energy is conserved in Eq. (2.1.8), and the conventional Noether's theorem[4] gives the energy relation for that equation as

$$\mathbf{r}_t\cdot(\mathsf{L}^{k+m}\sigma\mathbf{r}) = d_t T_k + d_x X'_k\,, \quad k = 0, 1, \ldots\,. \tag{2.1.19}$$

where

$$T_k = \mathscr{L}_k = \mathbf{r}\cdot\int_0^1 \mathbf{L}_k(\lambda\mathbf{r})\,d\lambda\,. \tag{2.1.20}$$

Note that Noether's relation (2.1.19) holds for all values of the field variable $\mathbf{r}$, so that we are not restricted to solutions of Eq. (2.1.8).

B. The integrodifferential part

The application of Lemma (2.5) to the lhs of Eq. (2.4) gives (using $\tau\sigma = -\sigma\tau$)

$$\mathbf{r}_t\cdot(\mathsf{L}^n\sigma\tau\mathbf{r}_t) = -\tfrac{1}{2}\frac{d}{dx}W_n(\mathbf{r}_t, \mathbf{r}_t)\,, \tag{2.2.1}$$

where

$$W_n(\mathbf{r}_t, \mathbf{r}_t) = \sum_{i=0}^{n-1}\{(\sigma\tau\mathsf{L}^i\tau\sigma\mathbf{r}_t)\cdot(\tau\mathsf{L}^{n-i-1}\tau\sigma\mathbf{r}_t) + 2\left[\int \mathbf{r}\cdot(\tau\mathsf{L}^i\tau\sigma\mathbf{r}_t)\right]\left[\int \mathbf{r}\cdot(\tau\mathsf{L}^{n-i-1}\tau\sigma\mathbf{r}_t)\right]\}\,. \tag{2.2.2}$$

The integral terms in W_n are not in general acceptable as flux terms in a conservation equation, since they do not vanish when evaluated on a boundary at which $\mathbf{r}$ and its derivatives are assumed to vanish, e.g., $x = -\infty$. However, for all $i > 0$,

$$\int \mathbf{r}\cdot(\tau\mathsf{L}^i\tau\sigma\mathbf{r}_t) = -\int \mathbf{r}\cdot(\tau\mathbf{G}_i) + \int \mathbf{r}\cdot(\tau\mathsf{L}^{i+m}\sigma\mathbf{r})\,, \tag{2.2.3}$$

$$= -\int \mathbf{r}\cdot(\tau\mathbf{G}_i) - \tfrac{1}{2}W_i(\tau\mathbf{r}, \tau\mathbf{r})\,, \tag{2.2.4}$$

where $W_i(\tau\mathbf{r}, \tau\mathbf{r})$ is partial differential, using Eq. (2.1.6). Hence, the integral terms in $W_n(\mathbf{r}_t, \mathbf{r}_t)$ are equal to polynomial terms, within the integral of an expression which vanishes for solutions to the original equation $\mathbf{G}_0 \doteq 0$. However, the integral does not necessarily vanish for all solutions to the equation $\mathbf{G}_n \doteq 0$, so that it will in general lead to nonconservation of the energy density for the equation $\mathbf{G}_n \doteq 0$.

The sum of the results (2.3) and (2.4) gives the result that the density

$$T_n = \mathbf{r}\cdot\int_0^1 \mathbf{L}_n(\lambda\mathbf{r})\,d\lambda\,, \tag{2.2.5}$$

where

$$\mathbf{L}_n(\mathbf{r}) = \mathsf{L}\mathbf{L}_{k-1}(\mathbf{r}) \tag{2.2.6}$$

and

$$\mathbf{L}_0(r) = \mathsf{L}^m\sigma\mathbf{r}\,, \tag{2.2.7}$$

is an energy density for the equation

$$\mathbf{G}_n \equiv \mathsf{L}^n(\sigma\tau\mathbf{r}_t + \mathsf{L}^m\sigma\mathbf{r}) = 0\,. \tag{2.2.8}$$

The energy density T_n is not in general conserved for all solutions to Eq. (2.2.8). In fact,

$$d_t\int_{-\infty}^{\infty} T_n\,dx = \sum_{i=1}^{n-1}\left[\int_{-\infty}^{\infty}(\mathbf{r}\cdot\tau\mathbf{G}_i)\,dx\right]\left[\int_{-\infty}^{\infty}(\mathbf{r}\cdot\tau\mathbf{G}_{n-i-1})\,dx\right]. \tag{2.2.9}$$

The rhs of Eq. (2.2.9) is nonzero in general for solutions to the equations $\mathbf{G}_k \doteq 0$, $k > n/2$. However, for the subset of the set of solutions to Eq. (2.2.8) which is the solution set of $\mathbf{G}_0 \doteq 0$, the source/sink terms on the rhs of Eq. (2.2.9) vanish, and T_n is conserved. Since this is true for any m, $n \geqslant 0$, the densities $\{T_n; n = 0, 1, \ldots\}$ constitute an infinity of p.c.d.'s for the equations $\mathbf{G}_0 \doteq 0$.

The derivation in this section clearly extends to the general set of equations (1.1) if A_0 is entire in its argument. If A_0 is the ratio of entire functions, then the derivation applies provided that the original equations (1.1) are partial differential, and that a Lagrangian density exists for $A_0(\mathsf{L})\sigma\mathbf{r}$ and for $\mathsf{L}A_0(\mathsf{L})\sigma\mathbf{r}$. To see this, let

$$2A_0(\mathsf{L})\sigma\mathbf{r} = \frac{\Sigma a_n\mathsf{L}^n}{\Sigma_{k=0}^{M} b_k L^k}\sigma\mathbf{r} = \mathbf{P}\,, \tag{2.2.10}$$

where a_i, b_i, are (possibly zero) constants, b_M is nonzero, and $\mathbf{P}$ is a polynomial in $\mathbf{r}$ and its x derivatives. Then

$$\left[\sum_{k=0}^{M} b_k\mathsf{L}^k\right][2A_0(\mathsf{L})\sigma\mathbf{r}] = \sum_n a_n\mathsf{L}^n\sigma\mathbf{r}\,, \tag{2.2.11}$$

$$= \sum_n a_n\mathbf{P}_N\,, \tag{2.2.12}$$

where $\mathbf{P}_n$ is partial differential by the results of Sec. 2.A.

Hence, since L^j is linearly independent of L^j for $i \neq j$, and since $\mathsf{L}\mathbf{v}$ cannot be partial differential if $\mathbf{v}$ contains an integral $\int_{-\infty}^{x}$, $\mathsf{L}^i[2A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r}]$ is partial differential for all $i \leqslant M$. Further, since for all $j \geqslant 0$,

$$\mathsf{L}^j \sum_{k=0}^{M} b_k \mathsf{L}^k [2A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r}] = \sum_n a_n \mathsf{L}^{n+j}\boldsymbol{\sigma}\mathbf{r}\,, \tag{2.2.13}$$

$\mathsf{L}^i[2A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r}]$ is partial differential for all i.

Hence, the Frechet derivative of $\mathsf{L}^i[2A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r}]$ exists, and the analyses of Sec. 2.A and 2.B apply under the above assumptions.

3. THE MOMENTUM DENSITIES OF THE ENVELOPING EQUATIONS

The densities T_k derived in Sec. 2 may be alternatively identified within a trivial sign as *momentum* densities of the enveloping equations (1.3) by proving the momentum relation

$$\mathbf{r}_x \cdot [\mathsf{L}^n(\boldsymbol{\sigma}\mathbf{r}_t + \mathsf{L}^m\boldsymbol{\sigma}\mathbf{r})] = d_t(-T_{N+1}) + d_x(\bar{X}_{n+1})\,, \tag{3.1}$$

since this is the Noether relation associating T_{n+1} with the field variation $\bar{\delta}\mathbf{r} = -\epsilon\mathbf{r}_x$, i.e., with an infinitesimal x translation. The proof of Eq. (3.1) is straightforward, using the results of Sec. 2 and noting that

$$2i\mathsf{L}\boldsymbol{\sigma}\mathbf{r} = \boldsymbol{\sigma}\mathbf{r}_x\,. \tag{3.2}$$

The integrodifferential part is, using Lemma (2.5),

$$\mathbf{r}_x \cdot (\mathsf{L}^n\boldsymbol{\sigma}\mathbf{r}_t) \simeq -\mathbf{r}_t \cdot (\mathsf{L}^n\boldsymbol{\sigma}\mathbf{r}_x)\,, \tag{3.3}$$

$$\simeq -\mathbf{r}_t \cdot (\mathsf{L}_{n+1}\boldsymbol{\sigma}\mathbf{r})\,, \tag{3.4}$$

which using Eq. (2.1.19) gives

$$\mathbf{r}_x \cdot (\mathsf{L}^n\boldsymbol{\sigma}\mathbf{r}_t) \simeq d_t(-T_{n+1})\,, \tag{3.5}$$

where

$$T_{n+1} = \mathbf{r} \cdot \int_0^1 \mathbf{L}_{n+1}(\lambda\mathbf{r})\, d\lambda\,, \tag{3.6}$$

with

$$\mathbf{L}_{n+1}(\mathbf{r}) = \mathsf{L}\mathbf{L}_n(\mathbf{r}) \tag{3.7}$$

and

$$\mathbf{L}_0(\mathbf{r}) = \boldsymbol{\sigma}\mathbf{r}\,. \tag{3.8}$$

Note that the momentum densities start with $\mathbf{L}_0 = \boldsymbol{\sigma}\mathbf{r}$, whereas the energy densities start with $\mathbf{L}_0 = 2A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r}$.

The partial differential part is

$$\mathbf{r}_x \cdot (\mathsf{L}^{n+m}\boldsymbol{\sigma}\mathbf{r}) = \mathbf{r}_x \cdot (\mathsf{L}^{n+m-1}\boldsymbol{\sigma}\mathbf{r}_x)\,, \tag{3.9}$$

and application of lemma (2.5) to the rhs of this gives

$$\mathbf{r}_x \cdot (\mathsf{L}^{n+m}\boldsymbol{\sigma}\mathbf{r}) \simeq 0\,. \tag{3.10}$$

4. COROLLARIES

An immediate corollary to Sec. 3 is that the set of equations

$$\boldsymbol{\sigma}\mathbf{r}_t + \mathsf{L}^n A_0(\mathsf{L})\boldsymbol{\sigma}\mathbf{r} = 0\,, \quad n = 0, 1, \ldots \tag{4.1}$$

(where A_0 is polynomial in its argument) has the *same* infinite set of p.c.d.'s $\{T_n\}$. The set of equations (4.1) is soluble by the same inverse scattering method, since it belongs to the AKNS class.[1] In the particular case of the KdV equations $[r = 1/\sqrt{6}, q = \phi_x/\sqrt{6}, A_0(q) = q^3]$ this set has become known as the Lax set.

The results of Sec. 2 may be used to investigate the p.c.d.'s of the equations

$$\mathsf{L}^m\boldsymbol{\sigma}\mathbf{r}_t + \boldsymbol{\sigma}\mathbf{r} \doteq 0\,, \quad m = 1, 2, \ldots\,. \tag{4.2}$$

In particular, it may be shown that for the first m energy densities T_k of the enveloping equations, T_k is *not* conserved for Eqs. (4.2), since for $k < m$,

$$d_t T_k + d_x X_k \doteq -d_x\left[\sum_{i=k}^{m-1} \int_{-\infty}^{x} \mathbf{r}\cdot(\boldsymbol{\tau}\mathsf{L}^i\boldsymbol{\tau}\boldsymbol{\sigma}\mathbf{r}_t)dx \times \int_{-\infty}^{x} \mathbf{r}\cdot(\boldsymbol{\tau}\mathsf{L}^{k+m-i-1}\boldsymbol{\tau}\boldsymbol{\sigma}\mathbf{r}_t)dx\right], \tag{4.3}$$

where X_k contains acceptable flux terms, but the terms on the rhs are not acceptable flux terms for Eq. (4.2). Integration of Eq. (4.3) over x, assuming the field variables and their derivatives vanish at $x = \pm\infty$, gives

$$d_t\left(\int_{-\infty}^{\infty} T_k dx\right) \doteq -\sum_{i=k}^{M-1} \int_{-\infty}^{\infty} (\mathbf{r}\cdot\boldsymbol{\tau}\mathsf{L}^i\boldsymbol{\tau}\boldsymbol{\sigma}\mathbf{r}_T)dx \times \int_{-\infty}^{\infty} (\mathbf{r}\cdot\boldsymbol{\tau}\mathsf{L}^{k+m-i-}{}^{+}\boldsymbol{\tau}\boldsymbol{\sigma}\mathbf{r}_t)dx\,, \tag{4.4}$$

i.e., the quantity $\int_{-\infty}^{\infty} T_k dx$ is not constant in time, since the source/sink terms on the rhs of Eq. (4.4) do not vanish for solutions to Eq. (4.2).

If $k \geqslant m$, T_k is conserved for solutions to Eqs. (4.2) so that Eqs. (4.2) have an infinite set of p.c.d.'s $\{T_k, k \geqslant m\}$ that diminishes with increasing m.

ACKNOWLEDGMENT

I am greatful to the Department of Education in the Republic of Ireland for the Research Fellowship that has made this paper possible.

APPENDIX A

The Korteweg–de Vries equation

In previous papers[3,4] the operator which generated higher-order equations for the KdV equation

$$\phi_{xt} + \phi_1\phi_2 + \phi_4 \doteq 0 \tag{A1}$$

was given as

$$H = d_x^2 + \tfrac{2}{3}\phi_1 + \tfrac{1}{3}\phi_2 \int_{-\infty}^{x}\,. \tag{A2}$$

In this case[1], $r = 1/\sqrt{6}$ and $q = \phi_x/\sqrt{6}$, so that

$$2i\mathsf{L} = \begin{pmatrix} -d_x - \tfrac{1}{3}\phi_x \int, & -\tfrac{1}{3}\phi_x \int \phi_x \\ \tfrac{1}{3}\int, & d_x + \tfrac{1}{3}\int \phi_x \end{pmatrix} \tag{A3}$$

and

$$(2i)^2\mathsf{L}^2 = \begin{pmatrix} H, & \tfrac{1}{3}\phi_{xx}\int\phi_x + \tfrac{1}{3}\phi_x\int\phi_{xx} \\ 0, & d_x^2 + \tfrac{2}{3}\phi_x - \tfrac{1}{3}\int\phi_{xx} \end{pmatrix}. \tag{A4}$$

Note that

$$(2i)^2\mathsf{L}^2\begin{pmatrix} v \\ 0 \end{pmatrix} = \begin{pmatrix} Hv \\ 0 \end{pmatrix}. \tag{A5}$$

The modified Korteweg–de Vries equation

In a previous paper[3], the operator generating higher-order equations for the modified KdV equation

$$\phi_{xt} + \phi_1^2\phi_2 + \phi_4 = 0 \tag{A6}$$

was given as

$$M = d_x^2 + \tfrac{2}{3}\phi_1^2 + \tfrac{2}{3}\phi_2\int_{-\infty}^{x}\phi_1 . \tag{A7}$$

In this case[1], $r = -q = \phi_x/\sqrt{6}$, so that

$$2i\mathsf{L} = \begin{pmatrix} -d_x - \tfrac{1}{3}\phi_1\int\phi_1, & -\tfrac{1}{3}\phi_1\int\phi_1 \\ \tfrac{1}{3}\phi_1\int\phi_1, & d_x + \tfrac{1}{3}\phi_1\int\phi_1 \end{pmatrix} \tag{A8}$$

and

$$(2i)^2\mathsf{L}^2 = \begin{pmatrix} a, & b \\ b, & a \end{pmatrix}, \tag{A9}$$

where

$$a \equiv d_x^2 + \tfrac{2}{3}\phi_1^2 + \tfrac{1}{3}\phi_2\int\phi_1 - \tfrac{1}{3}\phi_1\int\phi_2 ,$$

$$b \equiv \tfrac{1}{3}\phi_2\int\phi_1 + \tfrac{1}{3}\phi_1\int\phi_2 . \tag{A10}$$

Note that

$$(2i)^2\mathsf{L}^2\begin{pmatrix} v \\ v \end{pmatrix} = \begin{pmatrix} Mv \\ Mv \end{pmatrix}. \tag{A11}$$

The Sine–Gordon equation

The Sine–Gordon equation

$$\phi_{xt} + \sin\phi = 0 \tag{A12}$$

has the operator[3]

$$S = d_x^2 + \phi_1^2 + \phi_2\int\phi_1 . \tag{A13}$$

Since in this case[1] $r = -q = \phi_x$, and since operator S differs from operator M only by a scale transformation in ϕ, the results are identical within that scale transformation to those for operator M, and are omitted here.

The nonlinear Schrödinger equation

The higher-order equations for the nonlinear Schrödinger equation

$$i\phi_t + \phi_{xx} + \phi^2\phi^* = 0 \text{ and conjugate} \tag{A14}$$

are generated by the operator[3]

$$\mathsf{N} = \begin{pmatrix} d_x + \phi^*\int\phi, & -\phi^*\int\phi^* \\ \phi\int\phi, & -d_x - \phi\int\phi^* \end{pmatrix}. \tag{A15}$$

The substitutions[1] $q = -r^* = \phi/\sqrt{2}$ give

$$2i\mathsf{L} = \begin{pmatrix} -d_x - \phi\int\phi^*, & -\phi\int\phi \\ \phi^*\int\phi^*, & d_x + \phi^*\int\phi \end{pmatrix}, \tag{A16}$$

$$= \sigma\tau\mathsf{N}\tau\sigma , \tag{A17}$$

where σ are τ are defined at Eq. (1.2).

APPENDIX B

Infinities of conserved densities for linear equations

Assume that the linear equation

$$F(u) \doteq 0 \tag{B1}$$

has the conserved density $T_0(u)$ associated with the infinitesimal transformation $\bar{\delta}u$ via Noether's relation

$$\bar{\delta}uF(u) = d_tT_0(u) + d_xX_0(u) . \tag{B2}$$

Replace u by $v_n = d_x^n v$, where n is a positive integer, in Eq. (B2), to get

$$\bar{\delta}v_nF(v_n) = d_tT_0(v_n) + d_xX_0(v_n) . \tag{B3}$$

Since F is linear,

$$F(v_n) = d_x^nF(v) , \tag{B4}$$

and assuming that $\bar{\delta}u$ is linear in u (i.e., T_0 is quadratic), Eq. (B3) becomes

$$\bar{\delta}vd_x^{2n}F(v) = d_tT_n + d_xX_n , \tag{B5}$$

where

$$T_n(v) = T_0(v_n)$$

and

$$X_n(v) = X_0(v_n) - \sum_{i=1}^{n} d_x^{n-i}(\bar{\delta}v)d_x^{n-i-1}F . \tag{B6}$$

Hence, $T_n(u)$ is conserved for solutions to Eq. (B1), and is associated via Noether's theorem with the variation $\bar{\delta}u$ on the higher-order enveloping equation

$$d_x^{2n}F(u) = 0 , \tag{B7}$$

i.e., T_n is identified on the enveloping solution set of Eq. (B6) in the same way as T_0 is identified on the solution set of Eq. (B1).

Hence, if energy is conserved in Eq. (B1), there is an infinity of conserved densities for that equation, each identified as an energy density on the solution set of each Eq. (B6) ($n = 1, 2, \ldots$).

APPENDIX C

The densities T_k derived in Sec. 2 may be alternatively identified as being associated with an infinitesimal gauge transformation of the first kind[6] if $r = \pm q^*$, by proving the relation (which holds even if $r \neq \pm q^*$)

$$(\tau\mathbf{r})\cdot[\mathsf{L}^n(\tau\sigma\mathbf{r}_t + \mathsf{L}^m\sigma\mathbf{r})] = d_tT_n + d_x\bar{X}_n . \tag{C1}$$

The transformation $\bar{\delta}\mathbf{r} = \tau\mathbf{r}$ is a gauge transformation of the first kind if $r = \pm q^*$. Equation (C1) is easily proved by using the results of Sec. 2. The integrodifferential part is, using Lemma (2.5),

$$(\tau\mathbf{r})\cdot(\mathsf{L}^n\tau\sigma\mathbf{r}_t) \simeq \mathbf{r}_t\cdot(\mathsf{L}^n\sigma\mathbf{r}) , \tag{C2}$$

and result (2.1.19) may be applied to the rhs of Eq. (C2) to get

$$(\tau\mathbf{r})\cdot(\mathsf{L}^n\tau\sigma\mathbf{r}_t) \simeq d_tT_n , \tag{C3}$$

where

$$T_n = \mathbf{r}\cdot\int_0^1 \mathbf{L}_n(\lambda\mathbf{r})d\lambda\ , \tag{C4}$$

$$\mathbf{L}_n = \mathsf{L}\mathbf{L}_{n-1}\ ,$$

and

$$\mathbf{L}_0(\mathbf{r}) = \boldsymbol{\sigma}\mathbf{r}\ . \tag{C5}$$

The partial differential part is, using Lemma (2.5),

$$(\boldsymbol{\tau}\mathbf{r})\cdot(\mathsf{L}^{N+m}\boldsymbol{\sigma}\mathbf{r}) \simeq 0\ . \tag{C6}$$

Note that infinitesimal gauge transformations of the first kind may be associated where appropriate, with conservation of the number of particles (e.g., in the nonlinear Schrödinger equation), of wave action, or of charge.[6]

[1] M.J. Ablowitz, D.J. Kaup, A.C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[2] K. Konno, H. Sanuki, and Y.H. Ichikawa, Prog. Theor. Phys. **52**, 886 (1974).
[3] M.J. McGuinness, J. Math. Phys. **21**, 2737 (1980).
[4] M.J. McGuinness, J. Math. Phys. **19**, 2285 (1978).
[5] R.W. Atherton and G.M. Homsy, Stud. Appl. Math. **54**, 31 (1975).
[6] E.J. Saletan and A.H. Cromer, *Theoretical Mechanics* (Wiley, New York, 1971), p. 298.

The nonabelian Toda lattice—Discrete analogue of the matrix Schrödinger spectral problem

M. Bruschi, S. V. Manakov[a], O. Ragnisco, and D. Levi[b]
Istituto di Fisica dell'Universitá, 00185 Roma, Italy
and
Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Roma, Italy

(Received 9 May 1980; accepted for publication 20 June 1980)

We investigate the discrete analog of the matrix Schrödinger spectral problem and derive the simplest nonlinear differential-difference equation associated to such problem solvable by the inverse spectral transform. We also display the one and two soliton solution for this equation and tersely discuss their main features.

1. INTRODUCTION

Within the class of nonlinear differential-difference equations which so far have been integrated by the inverse spectral transform (IST), the Toda lattice, historically the first in this list,[1-3] is considered to be the most interesting from the physical point of view. Actually, this system provides an integrable model of one-dimensional classical crystals and moreover it is now established that its quantum version is also solvable.[4] In the present paper we investigate the nonabelian (matrix) generalization of this model, namely we consider the following system of differential-difference equations:

$$\frac{\partial}{\partial t}[\dot{G}(n)G^{-1}(n)] = G(n+1)G^{-1}(n) - G(n)G^{-1}(n-1), \tag{1.1}$$

where $G(n)$ is an arbitrary (in general, complex valued) nondegenerate $N \times N$ matrix, depending on the integer variable n and on the continuous real variable t, $\dot{G}(n)$ being its t-derivative.

As a discrete version of the principal chiral field model

$$\frac{\partial}{\partial t}(\dot{g}g^{-1}) = \frac{\partial}{\partial x}(g'g^{-1}) \quad \left(g' = \frac{\partial g}{\partial x}\right)$$

system (1.1) was first introduced by Poliakov,[5] who also discovered for it an infinite sequence of conserved quantities.

It is easy to show that system (1.1) can be cast in the Lax form:

$$\dot{L} = [L, M]. \tag{1.2}$$

Namely, introducing the new fields:

$$A(n) = G^{-1}(n)G(n+1); \quad B(n) = G^{-1}(n)\dot{G}(n); \tag{1.3}$$

Eq. (1.1) can be identically rewritten as:

$$\dot{A}(n) = A(n)B(n+1) - B(n)A(n),$$
$$\dot{B}(n) = A(n) - A(n-1). \tag{1.4}$$

The Lax representation (1.2) is achieved with the help of the two operators:

$$L = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & I & B(n) & A(n) & \cdot & \cdot \\ \cdot & 0 & 0 & I & B(n+1) & A(n+1) & 0 \\ \cdot & 0 & 0 & 0 & I & B(n+2) & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix};$$

$$M = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & 0 & -A(n) & 0 & 0 & 0 \\ \cdot & 0 & 0 & 0 & -A(n+1) & 0 & 0 \\ \cdot & 0 & 0 & 0 & 0 & -A(n+2) & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}. \tag{1.5}$$

In the following section we shall study the direct and inverse spectral problem:

$$L\psi = \lambda\psi \tag{1.6}$$

for the operator L defined here above, in the general case,

[a]Permanent address: Landau Institute for theoretical Physics of the Academy of Sciences, Moscow, USSR.

[b]Presently at Rockefeller University, New York 10021. The research reported in this paper has been supported in part by C. N. R. grant n. 78.00919.02

 0022-2488/80/122749-05$1.00

when $G(n)$ belongs to GL(N, C). In Sec. 3 we give the time-dependence of the spectral data of the operator L which corresponds to the dynamics given by Eq. (1.1). In Sec. 4 we first display the one-soliton solution for our problem, and discuss its behavior: we shall see that in the general complex case such solution is not well behaved, since a bounded initial datum can evolve in a solution diverging in a finite time; to prevent singularities we are thus forced to require that $G(n)$ be real valued ($G(n) \in$ GL(N, R)). In the same section we also discuss the two-soliton solution, which as in the abelian case, exhibits the typical phase-shift phenomenon and, for special choices of the spectral parameters, has the characteristic "breather" features.

2. DIRECT AND INVERSE SPECTRAL PROBLEM

The spectral problem (1.6) will be treated under the most natural boundary conditions, i.e.:

$$G(n) \to G_{\pm}, \quad \dot{G}(n) \to 0 \quad (n \to \pm \infty), \tag{2.1a}$$

and hence

$$A(n) \to I, \quad B(n) \to 0 \quad (n \to \pm \infty). \tag{2.1b}$$

Due to such boundary conditions, we have a twice-degenerate continuous spectrum, which can be parametrized by setting $\lambda = z + z^{-1}$, where z belongs to the unit circle in the complex plane; consequently, Eq. (1.6) can be rewritten in the form:

$$\psi(n-1, z) + B(n)\psi(n, z) + A(n)\psi(n+1, z) = (z + z^{-1})\psi(n, z), \tag{2.2}$$

$\psi(n, z)$ being a fundamental matrix solution. We define the Jost matrix solutions $\psi_{\pm}(n, z)$, $\varphi_{\pm}(n, z)$ by the following asymptotic behavior:

$$\lim_{n \to +\infty} \psi_{\pm}(n, z) z^{\mp n} = I \qquad (|z| = 1) \qquad \lim_{n \to -\infty} \varphi_{\pm}(n, z) z^{\mp n} = I. \tag{2.3}$$

From Eqs. (2.3) it obviously follows that

$$\psi_+(n, z) = \psi_-(n, z^{-1}) \qquad \varphi_+(n, z) = \varphi_-(n, z^{-1}). \qquad (|z|) = 1 \tag{2.4}$$

The monodromy matrix $\hat{M} = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ for $|z| = 1$ can be introduced in the standard way:

$$\varphi_-(n, z) = \psi_-(n, z)a(z) + \psi_+(n, z)b(z), \tag{2.5a}$$

$$\varphi_+(n, z) = \psi_-(n, z)c(z) + \psi_+(n, z)d(z), \tag{2.5b}$$

where, due to Eq. (2.4) the matrices $a(z)$, $b(z)$, $c(z)$, $d(z)$ are related by:

$$c(z) = b(z^{-1}); \quad d(z) = a(z^{-1}). \tag{2.5c}$$

It is easy to show that $\psi_+(n, z)$, $\varphi_-(n, z)$ are analytic inside the unit circle, and consequently $\psi_-(n, z)$, $\varphi_+(n, z)$ are analytic outside the unit circle. Let us consider, for example, ψ_+. By introducing

$$\chi(n, z) = z^{-n}\psi_+(n, z); \quad \chi(n, z) \underset{n \to +\infty}{\to} I, \tag{2.6a}$$

we get, from Eq. (2.2):

$$\chi(n-1, z) + (zB(n) - z^2 - 1)\chi(n, z) + z^2 A(n)\chi(n+1) = 0. \tag{2.6b}$$

Thus, as the solution exists for $|z| = 1$, it will exist *a fortiori* and be bounded also for $|z| < 1$; moreover we have

$$\lim_{z \to 0} \chi(n, z) = I, \tag{2.6c}$$

$$\dot{G}(n) = G(n) \frac{d}{dz} [z^{-n}\psi_+(n, z) - z^{-n+1}\psi_+(n-1, z)]_{z=0}. \tag{2.7a}$$

In a similar way it can be shown that $\varphi_-(n, z)$ has the same analytical properties, and that

$$G^{-1}(n) = \left[\lim_{z \to 0} \varphi_-(n, z) z^n\right] G_-^{-1}. \tag{2.7b}$$

In order to prove that $a(z)$ is also analytic inside the unit circle, we shall first assume that the "potentials" $A(n) - I$ and $B(n)$ are on compact support. In this case, there exists an integer N_0 such that

$$\varphi_-(n, z) z^n = a(z) + z^{2n} b(z) \quad \forall n > N_0, \quad |z| = 1. \tag{2.8}$$

On the other hand, for potentials on compact support, it is clear that both $a(z)$ and $b(z)$ can be analytically continued inside the unit circle (actually, they depend polynomially on z) so that Eq. (2.8) holds for $|z| < 1$ as well. But, if the potentials $A(n) - I$ and $B(n)$ vanish rapidly enough as $|n| \to \infty$, they can be uniformly approached by sequences of potentials on compact support, and thus it still holds true, for such potentials, that:

$$a(z) = \lim_{n \to +\infty} \varphi_-(n, z) z^n \tag{2.9a}$$

which means that $a(z)$ is analytic inside the unit circle. Furthermore, from formulas (2.7b), and (2.9) it follows that

$$a(0) = G_+^{-1} G_-. \tag{2.9b}$$

In order to treat the bound states, we shall assume $a(z)$ to be a nonsingular matrix on the unit circle, so that, for $|z| < 1$, its determinant can have at most a finite number of zeros z_j ($j = 1, \ldots, N$) which will be taken to be simple.

Hence, in a convenient neighborhood of z_j, we can write:

$$a(z) = C_j(I - \tilde{P}_j) + (z - z_j)a'(z_j) + O[(z - z_j)^2] \qquad \left(a'(z) = \frac{d}{dz} a(z)\right), \tag{2.10}$$

where C_j is a nonsingular matrix and $\tilde{P}_j$ is some one-dimensional projector, such that

$$\tilde{P}_j |c_j\rangle = |c_j\rangle, \quad \tilde{P}_j^2 = \tilde{P}_j, \tag{2.11a}$$

where, of course

$$a(z_j)|c_j\rangle = 0. \tag{2.11b}$$

From Eqs. (2.10) and (2.11) it follows that

$$a^{-1}(z) = \frac{\langle c_j | c_j \rangle}{\langle c_j | C_j^{-1} a'(z_j) | c_j \rangle} \tilde{P}_j C_j^{-1} (z - z_j)^{-1} + O(1). \tag{2.12}$$

Introducing the row vector $\langle d_j| = \langle c_j | C_j^{-1}$, such that $\langle d_j | a(z_j) = 0$, we can cast formula (2.12) in the simpler form:

$$a^{-1}(z) = -\rho_j z_j |c_j\rangle \langle d_j | (z - z_j)^{-1} + O(1), \tag{2.13a}$$

where

$$\rho_j = -(\langle d_j|z_j a'(z_j)|c_j\rangle)^{-1}. \tag{2.13b}$$

Finally, from formula (2.5a), taking into account (2.11b), it follows that there exists some vector $|\mathbf{c}_j\rangle$, such that

$$\varphi_-(n, z_j|c_j\rangle = \psi_+(n, z_j)|\mathbf{c}_j\rangle. \tag{2.14}$$

This formula, together with the asymptotic conditions (2.3), show that the solutions of Eq. (2.11b) provide the bound states of the operator L.

We now turn to the inverse problem, in order to reconstruct the potentials from the spectral data. To this aim, we define the matrix function

$$\Phi(n, z) = \begin{Bmatrix} \varphi_-(n, z)a^{-1}(z), & |z|<1 \\ \psi_-(n, z), & |z|>1 \end{Bmatrix}, \tag{2.15}$$

which is obviously analytic both outside and inside the unit circle, and has on the unit circle the jump [see Eq. (2.5a)]

$$\Delta\Phi(n, z) = \psi_+(n, z)b(z)a^{-1}(z). \tag{2.16}$$

Moreover, it satisfies the "normalization" condition:

$$\lim_{n\to+\infty} \Phi(n, z)z^n = I. \tag{2.17}$$

Hence, we can write for it the usual Cauchy formula

$$\Phi(n, z)z^n = I + \sum_{j=1}^{N} \frac{R_j(n)}{z - z_j} + (2\pi i)^{-1} \times \oint_{|z|=1} dz'\, z'^n \frac{\psi_+(n, z')b(z')a^{-1}(z')}{z' - z}, \tag{2.18}$$

where [see Eqs. (2.13) and (2.14)]

$$\begin{aligned} R_j(n) &= -\varphi_-(n, z_j)z_j^{n+1}\rho_j|c_j\rangle\langle d_j| \\ &= -\psi_+(n, z_j)z_j^{n+1}\rho_j|\mathbf{c}_j\rangle\langle d_j|. \end{aligned} \tag{2.19}$$

Taking into account (2.4) and (2.15), we can write

$$\psi_+(n, z_j) = z_j^n\Big(I - \sum_{k=1}^{N} \psi_+(n, z_k)z_k^{n+1}\rho_k|\mathbf{c}_k\rangle\langle d_k| \Big/ (z_j^{-1} - z_k)\Big) + (2\pi i)^{-1}\oint_{|z|=1} dz'\, z'^n \frac{\psi_+(n, z')b(z')a^{-1}(z')}{z' - z_j^{-1}} \quad (j = 1, ..., N) \tag{2.20}$$

which, together with formula (2.18), written in terms of ψ_+, defines uniquely $\psi_+(n, z)$ ($|z|\leqslant 1$). Consequently, from (2.18), we know $\Phi(n, z)$ in the whole complex z-plane, and hence $\varphi_-(n, z)a^{-1}(z)$ inside the unit circle, according to definition (2.15).

Therefore, we can assert that the spectral data

$$S \equiv \{b(z)a^{-1}(z),\ |z| = 1;\ z_j, \rho_j|\mathbf{c}_j\rangle\langle d_j| \quad (|z_j|<1, j = 1,...N)\} \tag{2.21}$$

together with the boundary conditions (2.1), enable us to reconstruct uniquely the potentials $G(n)$, $\dot{G}(n)$, taking into account formulas (2.7) and (2.9b).

To end this section, we notice that from the Cauchy formula (2.18) it is straightforward to obtain the proper discrete version of the Gel'fand–Levitan–Marchenko equation. Assuming that $\psi_+(n, z)$ admits the triangular representation:

$$\psi_+(n, z) = \sum_{m=n}^{\infty} K(n, m)z^m, \tag{2.22a}$$

where

$$K(n, n) = I;\quad \lim_{n\to+\infty} K(n, m) = \delta_{n,m}I \tag{2.22b}$$

and inserting this representation in formula (2.18), written in terms of ψ_+, we get the following "integral" equation for $K(n, l)$ $(l > n)$

$$K(n, l) + F(n + l) + \sum_{m=n+1}^{\infty} K(n, m)F(m + n) = 0, \tag{2.23a}$$

where

$$F(n) = \sum_{j=1}^{N} z_j^n\rho_j|\mathbf{c}_j\rangle\langle d_j| + (2\pi i)^{-1}\oint_{|z|=1} dz\, z^{n-1}b(z)a^{-1}(z). \tag{2.23b}$$

The potentials $A(n)$, $B(n)$ are easily obtained in terms of $K(n, m)$ by inserting formula (2.22a) into the eigenvalue equation (2.2) and requiring compatibility for the lowest powers of z. This yields

$$A(n) = I - K(n-1, n+1) + K(n, n+2) - K^2(n, n+1) + K(n-1, n)K(n, n+1), \tag{2.24a}$$

$$B(n) = K(n, n+1) - K(n-1, n); \tag{2.24b}$$

of course, from $A(n)$, $B(n)$, once given the boundary conditions (2.1a), one can recover uniquely $G(n)$, $\dot{G}(n)$ according to their definition (1.3).

3. TIME EVOLUTION OF THE SPECTRAL DATA

The Lax equation (1.2) implies the following time evolution for the eigenfunctions $\psi(n, \lambda)$

$$(L - \lambda I)[\dot{\psi}(n, \lambda) + M\psi(n, \lambda)] = 0, \tag{3.1}$$

so that we can assert that

$$\dot{\psi}(n, \lambda) + M\psi(n, \lambda) = \alpha(\lambda)\psi(n, \lambda), \tag{3.2}$$

the function α being determined by the boundary conditions.

In particular, it follows that the Jost solutions $\psi_\pm(n, z)$, $\varphi_\pm(n, z)$ obey the evolution equations

$$\dot{\psi}_\pm(n, z)[\dot{\varphi}_\pm(n, z)] + (M + (z^{\pm 1}/\lambda)L)\psi_\pm(n, z) \times [\varphi_\pm(n, z)] = 0. \tag{3.3}$$

Performing the time derivative of Eq. (2.5a) and inserting there formula (3.3), we get the following evolution equations for the elements of the monodromy matrix $a(z)$, $b(z)$

$$\dot{a}(z, t) = 0, \tag{3.4a}$$

$$\dot{b}(z, t) = (z - z^{-1})b(z, t). \tag{3.4b}$$

Hence, the reflection coefficient $R(z, t) = b(z, t)\,a(z, t)^{-1}$ evolves in time according to the formula:

$$R(z, t) = R(z, 0)\exp[(z - z^{-1})t]. \tag{3.5}$$

As for the bound-state spectral data, their time evolution obtains by the requirement that for the bound state eigenvector

$$|\psi(n, z_j)\rangle = \varphi_-(n, z_j)|c_j\rangle = \psi_+(n, z_j)|\mathbf{c}_j\rangle \tag{3.6}$$

the function α be zero. Thus we get:

$$|c_j(t)\rangle = |c_j(0)\rangle\exp(z_j^{-1}t), \tag{3.7a}$$

$$|\mathbf{c}_j(t)\rangle = |\mathbf{c}_j(0)\rangle \exp(z_j t). \tag{3.7b}$$

Recalling now formula (2.13a) and Eq. (3.4a), from (3.7a) we obtain

$$\rho_j(t)\langle d_j(t)| = \rho_j(0)\langle d_j(0)|\exp(-z_j^{-1}t) \tag{3.8}$$

whence it follows that

$$\rho_j(t)|\mathbf{c}_j(t)\rangle\langle d_j(t)| = \rho_j(0)|\mathbf{c}_j(0)\rangle\langle d_j(0)|\exp[(z_j - z_j^{-1})t]. \tag{3.9}$$

The above formula can be cast in a more convenient form through the following definitions

$$P_j = \frac{|\mathbf{c}_j\rangle\langle d_j|}{\langle d_j|\mathbf{c}_j\rangle}; \quad \sigma_j = \langle d_j|\mathbf{c}_j\rangle; \quad \nu_j = \rho_j\sigma_j, \tag{3.10}$$

which yield

$$P_j(t) = P_j(0), \tag{3.11a}$$

$$\nu_j(t) = \nu_j(0)\exp[(z_j - z_j^{-1})t]. \tag{3.11b}$$

4. ONE AND TWO SOLITON SOLUTIONS

As in the abelian case, the N-soliton solution can be evaluated by a purely algebraic procedure, starting from the Cauchy formula (2.18) and setting there $b(z) = 0$. Then Eq. (2.20) gives rise to a system of linear algebraic equations for the N unknown matrices $\psi_+(n, z_j)$. In particular, the one-soliton solution reads:

$$G(n) = G_+\{I - \sinh\zeta\exp(-\zeta) \times [1 - \tanh[\zeta(n - \tfrac{1}{2} - \xi)]]P_1\}, \tag{4.1}$$

$$\dot{G}(n) = G_+\sinh^2\zeta\exp(-\zeta)\operatorname{sech}^2[\zeta(n - \tfrac{1}{2} - \xi)]P_1, \tag{4.2}$$

where we have set

$$z_1 = \exp(-\zeta); \quad \xi = (2\zeta)^{-1}\ln\left[\frac{\nu_1}{2\sinh\zeta}\right]. \tag{4.3}$$

We notice that, due to Eqs. (3.12), the projector P_1 is constant in time, as it is, of course, z_1, while the parameter ξ evolves according to the formula

$$\xi(t) = \xi(0) - (\sinh\zeta/\zeta)t \tag{4.4}$$

which means that the (complex) position of the soliton moves with the (complex) speed $v_1 = -\sinh\zeta/\zeta$. In terms of the more familiar fields $A(n)$, $B(n)$ the one-soliton solution reads:

$$A(n) = I + \sinh^2\zeta\operatorname{sech}^2[\zeta(n + \tfrac{1}{2} - \xi)]P_1, \tag{4.5a}$$

$$B(n) = \sinh^2\zeta\operatorname{sech}[\zeta(n - \tfrac{1}{2} - \xi)]\operatorname{sech}[\zeta(n + \tfrac{1}{2} - \xi)]P_1. \tag{4.5b}$$

It is necessary to remark here that, if z_1 is not real, there exists always (i.e., for any initial condition) a finite time for which the one-soliton solution is unbounded around a certain point of the lattice. This is a characteristic feature of complex solutions; to prevent such singularities it is sufficient to restrict consideration to real valued matrices $G(n)$.

We now turn to describe very briefly the main features of the two-soliton solution, which can be obtained by the procedure outlined at the beginning of the present section. We give just the explicit expression of $[G(n)]^{-1}$, which is the easiest quantity to evaluate and, on the other hand, provides all relevant informations. It reads

$$[G(n)]^{-1}G_+ = I + \left(1 - \gamma\frac{\sinh\zeta_1\sinh\zeta_2}{4\sinh^2[(\zeta_1 + \zeta_2)/2]}\right)^{-1} \times \Big\{[\exp(\zeta_1)\sinh\zeta_1]\tau_1P_1 + [\exp(\zeta_2)\sinh\zeta_2]\tau_2P_2 - \left[\exp\left(\frac{\zeta_1 + \zeta_2}{2}\right)\frac{\sinh\zeta_1\sinh\zeta_2}{2\sinh[(\zeta_1 + \zeta_2)/2]}\right]\tau_1\tau_2(P_1P_2 + P_2P_1)\Big\}, \tag{4.6}$$

where

$$z_j = \exp(-\zeta_j); \quad \xi_j = (2\zeta_j)^{-1}\ln\left[\frac{\nu_j}{2\sinh\zeta_j}\right] \quad (j = 1, 2) \tag{4.7a}$$

$$P_1P_2P_1 = \gamma P_1; \quad P_2P_1P_2 = \gamma P_2, \tag{4.7b}$$

$$\tau_j = 1 - \tanh[\zeta_j(n - \tfrac{1}{2} - \xi_j)] \quad (j = 1, 2) \tag{4.7c}$$

the parameters ξ_j evolving linearly in time according to the formula

$$\xi_j(t) = \xi_j(0) - \frac{\sinh\zeta_j}{\zeta_j}t. \tag{4.8}$$

It is perhaps worthwhile to remark incidentally the striking similarity between solution (4.6) and the two-soliton solution associated to the matrix Schrödinger spectral problem.[6]

In order to prevent singularities, we have to require, as we did for the one-soliton solution, that the matrix $G(n)$ be real. But now, this "reality" requirement can be fulfilled in two different ways: either by assuming both the discrete eigenvalues and the corresponding polarizations to be real, or by letting them to be mutually complex conjugate.

In the first case (all parameters real) the situation is quite analogous to the abelian case. In particular, the two solitons are asymptotically separated, and moreover it can be shown that their interaction is such that, after the collision, the two solitons preserve the shape and the polarization they had before, just exhibiting a shift in their relative position. The easiest way to see this phenomenon is to choose a reference frame moving with one of the two solitons (for instance, the soliton 1) and to look at the asymptotic behavior of the solution in this frame, where we have of course,

$$\begin{aligned}\tau_1 &= 1;\\ \tau_2 &= 1 - \tanh[\zeta_2(\xi_2 - \xi_1)]\\ &= 1 - \tanh[\zeta_2(\xi_{\rm rel}(0) - v_{\rm rel}t)]\\ &\quad (\xi_{\rm rel} = \xi_2(0) - \xi_1(0);\ v_{\rm rel} = v_2 - v_1).\end{aligned} \tag{4.9}$$

Assuming, with no restriction, $v_{\rm rel} > 0$, it follows:

$$[G(n)]^{-1}G_+ \underset{t\to+\infty}{\sim} I + \exp(-\zeta_1)\sinh\zeta_1P_1 + \alpha x\beta_+Q, \tag{4.10a}$$

$$[G(n)]^{-1}G_+ \underset{t\to-\infty}{\sim} I + \left(1 - \gamma\frac{\sinh\zeta_1\sinh\zeta_2}{2\sinh^2[(\zeta_1 + \zeta_2)/2]}\right)^{-1} \times \Big\{\exp(-\zeta_1)\sinh\zeta_1P_1 + 2\exp(-\zeta_2)\sinh\zeta_2P_2 - 2\exp[(\zeta_1 + \zeta_2)/2]$$

$$\times\frac{\sinh\zeta_1\sinh\zeta_2}{\sinh[(\zeta_1+\zeta_2)/2]}(P_1P_2+P_2P_1)\Big\}+\alpha^{-1}x\beta_-Q, \tag{4.10b}$$

where $\alpha = 2\exp(-2\zeta_{2\mathrm{I}}\xi_{\mathrm{rel}}(0))$, $x = \exp(-2v_{\mathrm{rel}}|t|)$, $\beta_\pm$ are numerical coefficients defined by:

$$\beta_- = \sinh\zeta_2[\exp(\zeta_2) - \gamma\sinh\zeta_1];$$

$$\beta_+ = \beta_-\Big/\left(1-\frac{\sinh\zeta_1\sinh\zeta_2}{2\sinh^2[(\zeta_1+\zeta_2)/2]}\right)^2 \tag{4.11a}$$

and Q is the following projection matrix:

$$Q=\beta_-^{-1}\Big\{\frac{\exp(-\zeta_2)}{\sinh\zeta_2}P_2+\frac{\gamma}{4}\exp(\zeta_1)\frac{\sinh^2\zeta_1\sin\zeta_2}{\sinh^2[(\zeta_1+\zeta_2)/2]}P_1 - \exp[(\zeta_1+\zeta_2)/2]\frac{\sinh\zeta_1\sinh\zeta_2}{2\sinh[(\zeta_1+\zeta_2)/2]} \times(P_1P_2+P_2P_1)\Big\}. \tag{4.11b}$$

Formulas (4.10) clearly show that, but for an inessential constant matrix, the solution has the same structure both in the remote past and in the far future, the only difference consisting in the phase shift

$$\delta = \ln\left(\frac{\beta_+}{\beta_-}\right) = -2\ln\left|1-\frac{\sinh\zeta_1\sinh\zeta_2}{2\sinh^2[(\zeta_1+\zeta_2)/2]}\right|. \tag{4.12}$$

In the second case (complex conjugate parameters) the solution exhibits the typical breather behavior, since it is a matrix oscillating with the frequency

$$\omega = (2p)^{-1}|\psi\cos\psi\sinh p + p\cosh p\sin\psi|$$

$$\zeta_{1(2)} = p \pm i\psi \tag{4.13}$$

in the reference frame of the center of mass of the two solitons, defined, of course,

$$\bar{\xi} = (\zeta_1\xi_1+\zeta_2\xi_2)/(\zeta_1+\zeta_2), \tag{4.14}$$

which moves with the constant speed $\bar{v} = -\sinh p\cos\psi/p$.

5. CONCLUSIONS

We want just to remark that the results contained in the present paper can be useful for the study of the principal chiral field with zero moment at infinity (i.e. such that $\dot{g}$, $g'\to 0$, $|x|\to\infty$), where the standard Riemann problem technique[7] is not applicable.

[1]M. Toda, Progr. Theor. Phys. Suppl. **45**, 174 (1970).
[2]H. Flaschka, Progr. Theor. Phys. **51**, 703 (1974); Phys. Rev. B **9**, 1925 (1974).
[3]S. V. Manakov: Zh. Eksper. Teor. Fiz. **67**, 543 (1974).
[4]L. A. Thaktadjan, (private communication).
[5]Polyakov, (private communication).
[6]F. Calogero and A. Degasperis, Nuovo Cimento B **39**, 1 (1977).
[7]V. E. Zakharov and A. V. Mikhailov, Sov. Phys. JETP **47**, 1017 (1978).

An addition theorem for vector Helmholtz harmonics [a)]

F. Borghese, P. Denti, and G. Toscano
Universita di Messina, Istituto di Fisica, 98100 Messina, Italy

O. I. Sindoni
Chemical Systems Laboratory, Aberdeen Proving Ground, Maryland 21010

(Received 22 January 1980; accepted for publication 23 May 1980)

An addition theorem for the vector solutions of Helmholtz equations under translation of the coordinate axes is proposed and its results compared with those of a previous addition theorem for Hansen's M and N vectors. The resulting comparisons are also separated into their radial and transverse components.

In a problem of interaction of electromagnetic radiation with molecules we met the need of relating to each other the characteristic vector solutions of Helmholtz equations in two mutually translated systems of spherical coordinates. These vector functions, hereafter referred to as Vector Helmholtz Harmonics (VHH), can be written as

$$\mathbf{A}^M_{JL}(\mathbf{r}) = f_L(kr)\mathbf{T}^M_{JL}(\hat{\mathbf{r}}), \tag{1}$$

where f_L is a spherical Bessel or Hankel function and

$$\mathbf{T}^M_{JL}(\hat{\mathbf{r}}) = \sum_\mu C(1,L,J;-\mu,M+\mu)Y_{LM+\mu}(\hat{\mathbf{r}})\xi_{-\mu} \tag{2}$$

is an irreducible spherical tensor of rank-J.[1] The set of VHH's defined in Eqs. (1) and (2) is complete and orthogonal and diagonalizes simultaneously the operators J^2, J_z, L^2 and S^2 for vector fields.[2]

Our starting point is the addition theorem for Scalar Helmholtz Harmonics which we rewrite here in a form slightly different from that reported by Nozawa[3]:

$$f_L(kr)Y_{LM}(\hat{\mathbf{r}}) = \sum_{L'M'} G_{L'M',LM}(-\mathbf{R})g_{L'}(kr')Y_{L'M'}(\hat{\mathbf{r}}'), \tag{3}$$

where the quantities

$$G_{L'M',LM}(-\mathbf{R}) = 4\pi\sum_\lambda i^{L'-L-\lambda}I_\lambda(L'M';LM) \times\psi_\lambda(kR)Y^*_{\lambda M'-M}(\hat{\mathbf{R}}), \tag{4}$$

with $\mathbf{r}=\mathbf{R}+\mathbf{r}'$, are the matrix elements in the angular momentum representation of the free space propagator for spherical waves.[4] In Eqs. (3) and (4), where $f_L = j_L$, $\psi_\lambda = j_\lambda$ and $g_{L'} = j_{L'}$, but when $f_L = h_L$,

$$\psi_\lambda = h_\lambda, \quad g_{L'} = j_{L'} \quad :r'<R,$$

$$\psi_\lambda = j_\lambda, \quad g_{L'} = h_{L'} \quad :r'>R,$$

and the quantities

$$I_\lambda(L'M';LM) = \int Y^*_{L'M'}Y_{LM}Y_{\lambda M'-M}\,d\Omega = [(2L+1)(2\lambda+1)/4\pi(2L'+1)]^{1/2} C(L,\lambda,L';00)C(L,\lambda,L';M,M'-M) \tag{5}$$

are the well-known Gaunt integrals.[5] Substitution of Eq. (2) into Eq. (1) and application of Eq. (3) yields

$$\mathbf{A}^M_{JL}(\hat{\mathbf{r}}) = \sum_\mu C(1,L,J;-\mu,M+\mu)\sum_{L'M''} G_{L'M'',LM+\mu}(-\mathbf{R}) \times g_{L'}(kr')Y_{L'M''}(\hat{\mathbf{r}}')\xi_{-\mu}$$

which can be written as

$$\mathbf{A}^M_{JL}(\mathbf{r}) = \sum_\mu C(1,L,J;-\mu,M+\mu)\sum_{L'M''} G_{L'M'',LM+\mu}(-\mathbf{R}) \times\sum_{J'} C(1,L',J';\mu,M'')g_{L'}(kr')\mathbf{T}^{M''-\mu}_{J'L'}(\hat{\mathbf{r}}')$$

through the use of the inverse to Eq. (2). Now, putting $M' = M'' - \mu$,

$$\mathcal{G}^{M'M}_{J'L',JL} = \sum_\mu C(1,L'J';-\mu,M'+\mu) \times G_{L'M'+\mu,LM+\mu}(-\mathbf{R})C(1,L,J;-\mu,M+\mu),$$

we get

$$\mathbf{A}^M_{JL}(\mathbf{r}) = \sum_{L'}\sum_{J'M'}\mathcal{G}^{M'M}_{J'L',JL}\,g_{L'}(kr')\mathbf{T}^{M'}_{J'L'}(\hat{\mathbf{r}}'), \tag{6}$$

which is the required addition theorem.

Unlike the previous addition theorem of Stein[6] and Cruzan[7] for M and N vectors, the applicability of Eq. (6) is not restricted to solenoidal fields. Of course we could add an addition theorem for L simply by taking the gradient of both sides of Eq. (3) but the lack of orthogonality of L and N may be cumbersome. Anyway since

$$\mathbf{M}_{LM}(\mathbf{r}) = f_L(kr)\mathbf{X}_{LM}(\hat{\mathbf{r}}) = -f_L(kr)\mathbf{T}^M_{LL}(\hat{\mathbf{r}}), \tag{7a}$$

$$\mathbf{N}_{LM}(\mathbf{r}) = \frac{1}{k}\nabla\times\mathbf{M}_{LM} = i\left[\left(\frac{L+1}{2L+1}\right)^{1/2} f_{L-1}\,\mathbf{T}^M_{LL-1} - \left(\frac{L}{2L+1}\right)^{1/2} f_{L+1}\,\mathbf{T}^M_{LL+1}\right], \tag{7b}$$

the theorem of Cruzan can be easily related to our Eq. (6). Indeed for $\mathbf{M}_{LM}$ we have

$$\mathbf{M}_{LM}(\mathbf{r}) = -\sum_{L'M'}\{\mathcal{G}^{M'M}_{L'L';LL}\,g_{L'}(kr')\mathbf{T}^{M'}_{L'L'}(\hat{\mathbf{r}}') + \mathcal{G}^{M'M}_{L'L'-1;LL}\,g_{L'-1}(kr')\mathbf{T}^{M'}_{L'L'-1} + \mathcal{G}^{M'M}_{L'L'+1;LL}\,g_{L'+1}(kr')\mathbf{T}^{M'}_{L'L'+1}(\hat{\mathbf{r}}')\}, \tag{8}$$

where on account of the divergenceless characters of M_{LM}, the recursions relation follows

$$\left(\frac{L'+1}{2L'+1}\right)^{1/2}\mathcal{G}^{M'M}_{L'L'+1;LL} + \left(\frac{L'}{2L'+1}\right)^{1/2}\mathcal{G}^{M'M}_{L'L'-1;LL} = 0, \tag{9}$$

[a)]Based on work supported by the U.S. Army European Research Office through Grant DAERO 78-G-A06.

 0022-2488/80/122754-02$1.00

which can also be proved by direct calculation making use of the Clebsch–Gordan coefficients. With the help of Eq. (9), Eq. (8) can be put into the form by

$$\mathbf{M}_{LM}(\mathbf{r}) = \sum_{L'M'} \{A(L'M';LM)\bar{\mathbf{M}}_{L'M'}(\hat{\mathbf{r}}') + B(L'M';LM)\bar{\mathbf{N}}_{L'M'}(\hat{\mathbf{r}}')\},$$

where we put

$$A(L'M';LM) = \mathcal{G}^{M'M}_{L'L';LL};$$

$$B(L'M';LM) = -i\left(\frac{2L'+1}{L'}\right)^{1/2} \mathcal{G}^{M'M}_{L'L'+1;LL}$$

and $\bar{\mathbf{M}}$ and $\bar{\mathbf{N}}$ are identical to $\mathbf{M}$ and $\mathbf{N}$ but for the substitution of g_L to f_L. The A and B coefficients as defined here differ from those of Cruzan because of the different normalization chosen for $\mathbf{M}$ and $\mathbf{N}$. However, the properties of Cruzan's coefficients are a direct consequence of those of $\mathcal{G}$ which in turn follow from the symmetry and recursion properties of the G matrix elements[8] and of the Clebsch–Gordan coefficients.[1,2,9]

The last point we want to stress is that the right-hand side of Eq. (6) can be easily separated into radial and transverse components with respect to $\mathbf{r}'$ through the equations[10,11]

$$\mathbf{T}^M_{LL+1} = \left(\frac{L}{2L+1}\right)^{1/2}(-i)\hat{\mathbf{r}}\times\mathbf{X}_{LM} - \left(\frac{L+1}{2L+1}\right)^{1/2}\hat{\mathbf{r}}Y_{LM},$$

$$\mathbf{T}^M_{LL} = \mathbf{X}_{LM},$$

$$\mathbf{T}^M_{LL-1} = \left(\frac{L+1}{2L+1}\right)^{1/2}(-i)\hat{\mathbf{r}}\times\mathbf{X}_{LM} + \left(\frac{L}{2L+1}\right)^{1/2}\hat{\mathbf{r}}Y_{LM}.$$

The resulting equations can be very useful e.g. to impose boundary conditions on a spherical surface centered at $\mathbf{R}$.

[1] E.M. Rose, *Multipole Fields* (Wiley, New York, 1955); the notation and phase convention of this reference are followed as closely as possible.
[2] E.M. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957).
[3] R. Nozawa, J. Math Phys. **7**, 1861 (1966).
[4] K.H. Johnson, J. Chem. Phys. **45**, 3085 (1966).
[5] S. Stein, Quart. Appl. Math. **19**, 15 (1961).
[6] O.R. Cruzan, Quart. Appl. Math. **20**, 33 (1962).
[7] W.W. Hansen, Phys. Rev. **47**, 139 (1935).
[8] J.A. Stratton, *Electromagnetic Theory* (McGraw–Hill, New York, 1961).
[9] A.R. Edwards, *Angular Momentum in Quantum Mechanics* (Princeton U. P., Princeton, N.J., 1957).
[10] R.G. Newton, *Scattering Theory of Waves and Particles* (McGraw–Hill, New York, 1966).
[11] J.D. Jackson, *Classical Electromagnetism* (Wiley, New York, 1962).

Particle trajectories in $1/r$ fields

M. Arnow
Zenith Radio Corporation—Rauland Division, 2407 North Avenue, Melrose Park, Illinois 60160

(Received 10 June 1980, accepted for publication 25 July 1980)

The trajectory of a particle subjected to an attractive $1/r$ force is discussed. The general mathematical solution is given. Various analytical results are derived including the representations for the trajectory function.

I. INTRODUCTION

The trajectory of a particle subjected to an attractive central force varying as $1/r$, where r is the radial displacement is usually avoided in books on classical dynamics.[1] Although this force appears at first glance to be unnatural, it has been well approximated in devices such as electrostatic cylindrical spectrometers.[2] The general analysis of this problem is the subject of this paper with special attention given to the mathematical properties of the trajectory function.

II. THE DYNAMICAL PROBLEM

Assume that a particle of mass m experiences a force F, where

$$F = -A/r, \tag{1}$$

and where A is the force constant. Then the total energy E of the particle is given by

$$E = \tfrac{1}{2}mv^2 + A\ln(r) + \text{const}, \tag{2}$$

where v is the particle's velocity. For a particle with nonzero angular momentum and finite E, r must be bounded. Let the maximum and minimum radial displacements be $r_{\max}$ and $r_{\min}$ respectively.

Consider a particle approaching $r_{\max}$; let its radial displacement be r_a and velocity be v_a. After the particle passes through $r_{\max}$ and arrives again at the displacement r_a, its velocity must be v_a because of energy and angular momentum conservation. The trajectory is therefore symmetric about $r_{\max}$. A similarly constructed argument for a particle passing through $r_{\min}$ results in showing that the trajectory is symmetric about $r_{\min}$.

The angular displacement between consecutive $r_{\max}$ and $r_{\min}$ must remain constant due to symmetry and angular momentum conservation. The particle must therefore have a trajectory periodic in the angular displacement variable θ; i.e.,

$$r(\theta + 2P) = r(\theta), \tag{3}$$

where $2P$ is the period and P is the angular distance between $r_{\max}$ and $r_{\min}$; see Fig. 1.

III. THE MATHEMATICAL PROBLEM

The differential equation for the trajectory is given by[3]

$$d^2u/d\theta^2 + u = c^2/u, \tag{4}$$

where $u = r_0/r$, $r = r_0$ at $\theta = 0$ and

$$c^2 = mAr_0^2/l^2, \tag{5}$$

where l is the angular momentum. The boundary conditions are $r = r_0$ and $du/d\theta = -\tan(\phi)$ at $\theta = 0$, where ϕ is the angle between the tangent to the trajectory and the perpendicular to the displacement r_0; refer to Fig. 1.

Equation (4) has the trivial solution $u = c$; this is the circular orbit solution which is only valid for $c = 1$. It is convenient to label the kinetic energy of the circularly orbiting particle K_c, where

$$K_c = \tfrac{1}{2}A. \tag{6}$$

Equation (6) follows from Newton's Second Law and the condition for a circular orbit.

The solution of Eq. (4) is simplified if r_0 is chosen to be an extremum; e.g., $r_0 = r_{\min}$. Then $r = r_{\min}$ and $du/d\theta = 0$ at $\theta = 0$. The above arguments imply that u has the following properties:

Property I: $u_{\min} \leqslant u \leqslant 1$, where $u_{\min} = r_{\min}/r_{\max}$.
Property II: $u(\theta + 2P) = u(\theta)$.
Property III: $u(P) = u_{\min}$.

The parameter c^2 is now $c^2_{\min}$ where

$$c^2_{\min} = K_c/K_0, \tag{7}$$

and where K_0 is the kinetic energy of the particle at $\theta = 0$. Since $K_0 \geqslant K_c$, $c^2_{\min} \leqslant 1$.

Equation (4) may be integrated to obtain

$$du/d\theta = -[1 + c^2\ln(u^2) - u^2]^{1/2} \equiv -f(u)^{1/2}, \tag{8}$$

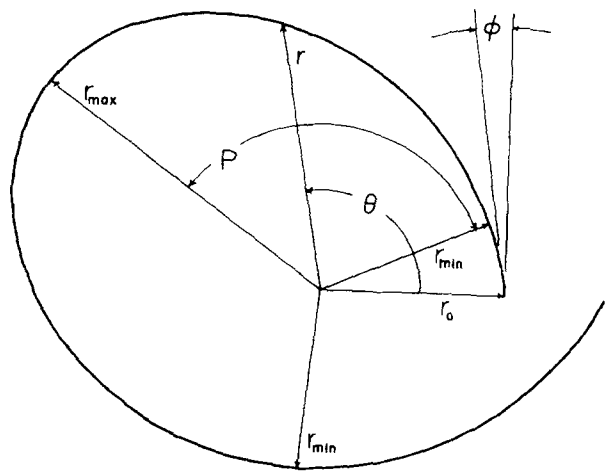


FIG. 1. The trajectory of a particle subjected to a force varying as $1/r$. The radial displacement is r; the angular displacement is θ. The trajectory's maximum and minimum are $r_{\max}$ and $r_{\min}$, respectively. The half period P is the angular distance between $r_{\max}$ and $r_{\min}$. The displacement r_0 is the radial displacement when $\theta = 0$, and ϕ is the angle between the tangent at $\theta = 0$ and the perpendicular to r_0.

 0022-2488/80/122756-04$1.00

for $0 \leqslant \theta \leqslant P$. The function $f(u)$ has roots at $u = 1$ and $u = u_{\min}$. The half-period P is obtained by integrating Eq. (8); i.e.,

$$P = -\int_1^{u_{\min}} f(t)^{-\frac{1}{2}} dt. \tag{9}$$

The value of P in the limit $c^2 \to 1$ is found by first noting that for $c^2 = 1 - \epsilon$, where $\epsilon \ll 1$,

$$u = 1 - \epsilon y_1 - \epsilon^2 y_2 - \cdots, \tag{10}$$

where $y_1, y_2, \ldots$, are functions of θ and of order unity. When Eq. (10) is substituted into Eq. (8) and the terms of order ϵ^3 and higher are discarded, $f(u)$ is approximated by

$$f(y_1) = 2\epsilon^2 y_1 - 2\epsilon^2 y_1^2. \tag{11}$$

Using Eq. (11) to find an expression for $u_{\min}$ results in

$$u_{\min} = 1 - \epsilon, \tag{12}$$

and the integral in Eq. (9) can be transformed into

$$P = \frac{1}{\sqrt{2}} \int_{-1}^{1-2\epsilon} [1 - t^2]^{-\frac{1}{2}} dt. \tag{13}$$

Finally,

$$\lim_{c^2 \to 1} P = \pi/\sqrt{2} \tag{14}$$

Determining P when $c^2 \to 0$ is more involved if rigor is required. The solution requires the facts that the integral in Eq. (9) is convergent and that $u_{\min} \to 0$. The latter follows from $r_{\max} \to \infty$ when $K_0 \to \infty$. Careful attention to the limiting process produces

$$\lim_{c^2 \to 0} P = \int_0^1 [1 - t^2]^{-\frac{1}{2}} dt, \quad = \tfrac{1}{2}\pi. \tag{15}$$

The period is therefore bounded in the interval $\pi < 2P \leqslant (\sqrt{2})\pi$.

IV. THE ANALYTICAL PROPERTIES OF u

To facilitate the analysis the behavior of u in the complex plane will be found useful. Borrowing a technique from elliptic function theory,[4,5] let

$$y = u(z); \tag{16}$$

then

$$z = u^{-1}(y). \tag{17}$$

The generalized u is defined through its inverse by the relation

$$z = -\int_1^y f(t)^{-\frac{1}{2}} dt, \tag{18}$$

with the parameter c defined to be real. Clearly Eqs. (4) and (8) follow from Eq. (18).

Let z_s be the point where u is infinite. Then

$$z_s = -\int_1^{\infty} f(t)^{-\frac{1}{2}} dt. \tag{19}$$

The latter integral diverges for all c; therefore, u has no infinities on the finite complex plane.

Let z_0 be the location of the nearest zero from the origin. Then

$$z_0 = -\int_1^0 f(t)^{-\frac{1}{2}} dt. \tag{20}$$

The latter integral is finite for all c. The existence of a zero implies that u has a branch point at z_0.

Noting that u is symmetric about $\theta = 0$, the power series representation for u on the real axis is given by

$$u = 1 + \sum_{n=1}^{\infty} a_n \theta^{2n}. \tag{21}$$

The radius of convergence R for the above series is the distance to the closest nonanalytical point in the complex plane, or

$$R = |z_0|. \tag{22}$$

Using arguments similar to those given for evaluating P, it can be shown that $R \geqslant P$ for $c \leqslant 1$ with $R \to P$ as $c \to 0$.

The recursion relation for the coefficients a_n in Eq. (21) is found by substituting the series in Eq. (21) into Eq. (4); the result being

$$a_n = [(-1)^n/(2n)!](1 - c^2)Q_n(c^2), \tag{23}$$

where

$$Q_1 = 1,$$

and,

$$Q_{n+1} = (1 + c^2)Q_n + (1 - c^2)(2n)! \times \sum_{i=1}^{n-1} \frac{1}{(2i)!} \frac{1}{(2n - 2i)!} (Q_i - Q_{i+1})Q_{n-i}. \tag{24}$$

Some of the Q_n polynomials have been computed and listed in the Appendix.

It is possible to sum the series in Eq. (21) when $c = 1 - \epsilon, \epsilon \ll 1$. The latter condition reduces Q_n in Eq. (24) to

$$Q_n = 2^{n-1} + O(\epsilon). \tag{25}$$

Then Eq. (21) sums to

$$u = 1 - \epsilon + \epsilon \cos(\sqrt{2}\theta) + O(\epsilon^2). \tag{26}$$

Equation (26) is a form of the solution of electron trajectories in electrostatic cylindrical spectrometers first given by Hughes and Rojansky in 1929.[6]

Had the analysis been chosen to have $r = r_{\max}$ at $\theta = 0$, nearly identical results would have been obtained. The differences being that now $c^2 = c^2_{\max}$, and the radius of convergence R for the series in Eq. (21) is changed to $R \geqslant P$ for a small range of c near unity, with $R \to 0$ as $c \to \infty$. The parameters $c_{\max}$ and $c_{\min}$ are related through the constraints of energy and angular momentum conservation; the relation being

$$c_{\max}^{-2} + \ln(c_{\max}^2) = c_{\min}^{-2} + \ln(c_{\min}^2). \tag{27}$$

The definition of u gives

$$u(\theta, c_{\max}) = u(\theta + P, c_{\min})/u_{\min}. \tag{28}$$

V. THE FOURIER SERIES REPRESENTATION

The function $u(\theta)$ is periodic and an even function of θ. A Fourier cosine series representation is therefore permit-

ted; i.e.,

$$u = \tfrac{1}{2}b_0 + \sum_{n=1}^{\infty} b_n \cos\left(\frac{n\pi\theta}{P}\right), \tag{29}$$

where,

$$b_n = \frac{2}{P}\int_0^P u(\theta)\cos\left(\frac{n\pi\theta}{P}\right) d\theta. \tag{30}$$

The power series for $u(\theta)$ may be used to evaluate the integral in Eq. (30), provided that the convergence criteria are met, but this leads to slowly converging infinite series for each b_n. An exceptional case occurs for $c^2 \approx 1$; the series for the b_n's can be put into a rapidly converging form by use of a perturbation expansion.

VI. A PERTURBATION SOLUTION FOR $c^2 \approx 1$

Consider the case where the trajectory is nearly circular, then c may be expressed as $c = 1 - \epsilon$, where $|\epsilon| \ll 1$. Let

$$u(\theta) = c(1 + w(\theta)), \tag{31}$$

where $|w(\theta)| \ll 1$. Then Eq. (4) can be expanded and written as

$$d^2w/d\theta^2 + 2w = w^2 - w^3 + w^4 - \cdots, \tag{32}$$

with

$$w(0) = \epsilon + \epsilon^2 + \epsilon^3 + \cdots, \tag{33}$$

and

$$dw/d\theta\,|_{\theta=0} = 0. \tag{34}$$

Using the method of Lindstedt–Poincaré,[7,8] the variable β is defined to be

$$\beta = \omega\theta, \tag{35}$$

where

$$\omega = 1 + \epsilon\omega_1 + \epsilon^2\omega_2 + \epsilon^3\omega_3 + \cdots. \tag{36}$$

The parameters $\omega_1, \omega_2, \cdots$ are to be determined. Now let w be expanded parametrically as

$$w = \epsilon w_1 + \epsilon^2 w_2 + \epsilon^3 w_3 + \cdots, \tag{37}$$

where $w_1, w_2, \ldots$ are also to be determined. Rewriting Eq. (32) in terms of the variable β gives

$$\omega^2 \frac{d^2w}{d\beta^2} + 2w = w^2 - w^3 + w^4 - \cdots. \tag{38}$$

Substituting Eqs. (36) and (37) into Eq. (38) gives

$$(1 + \epsilon\omega_1 + \cdots)^2\left(\frac{d^2}{d\beta^2} + 2\right)(\epsilon w_1 + \epsilon^2 w_2 + \cdots) = (\epsilon w_1 + \epsilon^2 w_2 + \cdots)^2 - \cdots. \tag{39}$$

When Eq. (39) is expanded and terms of like order in ϵ are equated the following sequence of equations is produced:

$$O(\epsilon): d^2w_1/d\beta^2 + 2w_1 = 0; \tag{40}$$

$$O(\epsilon): d^2w_2/d\beta^2 + 2w_2 = w_1^2 - 2\omega_1(d^2w_1/d\beta^2); \tag{41}$$

$$O(\epsilon): d^2w_3/d\beta^2 + 2w_3 = -2\omega_2(d^2w_1/d\beta^2) - 2w_1w_2 - w_1^3; \tag{42}$$

etc. The sequence may be extended as far as patience permits. The Eqs. (40), (41), (42), etc. are solved sequentially, and the parameters $\omega_1, \omega_2, \cdots$ are chosen to make resonant terms in the solution vanish; e.g., the solutions of Eqs. (40) and (41) result in

$$y_1 = \cos(\sqrt{2}\,\beta), \tag{43}$$

and

$$y_2 = \tfrac{1}{4} + \tfrac{1}{6}\cos(\sqrt{2}\,\beta) - \tfrac{1}{12}\cos(2\sqrt{2}\,\beta), \tag{44}$$

with the requirement that $\omega_1 = 0$. Thus, to the second order in ϵ, Eq. (4) is given by

$$u = 1 - \epsilon + \tfrac{1}{4}\epsilon^2 + (\epsilon - \tfrac{1}{6}\epsilon^2) \times \cos(\sqrt{2}\theta) + \tfrac{1}{12}\epsilon^2\cos(2\sqrt{2}\theta). \tag{45}$$

The agreement between Eqs. (45) and (26) is obvious. The above method when extended to the higher order terms in ϵ produces the results listed in the Appendix.

VII. THE REPRESENTATIONS FOR $r(\theta)$

The power series representation for $r(\theta)$ is readily found from Eq. (4) since

$$r/r_0 = (1/c^2)(d^2u/d\theta^2 + u); \tag{46}$$

implying that

$$\frac{r}{r_0} = 1 + \frac{(1-c^2)}{c^2}\sum_{n=1}^{\infty}\frac{(-1)^n}{(2n)!} \times (Q_n(c^2) - Q_{n+1}(c^2))\theta^{2n}. \tag{47}$$

The radius of convergence of the latter series is the same as that for the series in Eq. (21) with a corresponding parameter c^2.

If the Fourier series for $r(\theta)$ is given by

$$r(\theta) = \tfrac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(n\pi\theta/P), \tag{48}$$

where

$$a_n = \frac{2}{P}\int_0^P r(\theta)\cos(n\pi\theta/P)\, d\theta. \tag{49}$$

The coefficients a_n can be related to the coefficients b_n in Eq. (29) by an integration by parts. The relation is

$$b_n = (1/c^2)(1 - n^2\pi^2/P^2)a_n. \tag{50}$$

VIII. REMARKS

In working through the mathematics of this problem, one is struck by the similarities $u(\theta)$ has to the Jacobi elliptic functions. For example, if Eq. (32) were approximated by using terms to the third order and neglecting higher orders, a solution in closed form in terms of elliptic functions is possible.[9] Yet the similarities are not close enough to permit simplification of the representations for $u(\theta)$. For instance, there is no apparent algebraic addition formula which will express $u(\theta_1 + \theta_2)$ in terms of $u(\theta_1)$, $u(\theta_2)$, and their derivatives. Also, contour integrations to find a closed form for the Fourier coefficients do not appear promising.

Finally, the particle trajectories have the interesting property that depending on the parameter c the orbits may be open or closed. This property is a consequence of the period being a continuous function of c. When the period is a rational multiple of π the orbit is closed; when the period is an irrational multiple of π the orbit is open.

APPENDIX

The first seven $Q_n(c^2)$ polynomials computed from Eq. (24) are:

$$Q_1 = 1;$$
$$Q_2 = 1 + c^2;$$
$$Q_3 = 1 - 4c^2 + 7c^4;$$
$$Q_4 = 1 + 87c^2 - 207c^4 + 127c^6;$$
$$Q_5 = 1 - 2138c^2 + 8070c^4 - 10286c^6 + 4369c^8;$$
$$Q_6 = 1 + 79883c^2 - 432308c^4 + 863404c^6 - 754597c^8 + 243649c^{10};$$

and

$$Q_7 = 1 - 5266677c^2 + 30997509c^4 - 85021777c^6 + 116205843c^8 - 76951818c^{10} + 20036983c^{12}.$$

When the perturbation method in Sec. VI is taken to the fourth order in ϵ, where $\epsilon = c - 1$ for $c \approx 1$, the Fourier coefficients in Eq. (29) are:

$$\tfrac{1}{2}b_0 = 1 - \epsilon + (1/4)\epsilon^2 + (1/6)\epsilon^3 + (11/64)\epsilon^4;$$
$$b_1 = \epsilon - (1/6)\epsilon^2 - (19/144)\epsilon^3 - (607/4320)\epsilon^4;$$
$$b_2 = -(1/12)\epsilon^2 - (1/18)\epsilon^3 - (1/18)\epsilon^4;$$
$$b_3 = (1/48)\epsilon^3 + (1/32)\epsilon^4;$$

and

$$b_4 = -(61/8640)\epsilon^4;$$

With the aid of Eq. (46), the Fourier coefficients in Eq. (48) are found to be:

$$\tfrac{1}{2}a_0 = 1 + \epsilon + (5/4)\epsilon^2 + (5/3)\epsilon^3 + (433/192)\epsilon^4;$$
$$a_1 = -\epsilon - (11/6)\epsilon^2 - (413/144)\epsilon^3 - (18413/4320)\epsilon^4;$$
$$a_2 = (7/12)\epsilon^2 + (14/9)\epsilon^3 + (109/36)\epsilon^4;$$
$$a_3 = -(17/48)\epsilon^3 - (79/96)\epsilon^4;$$

and

$$a_4 = (1891/8640)\epsilon_4.$$

The expansion parameter ω is given by

$$\omega = \pi/(\sqrt{2})P = 1 + (1/12)\epsilon^2 + (5/36)\epsilon^3 + (111/576)\epsilon^4.$$

Then the half period P is

$$P = (\pi/\sqrt{2})(1 - (1/12)\epsilon^2 - (5/36)\epsilon^3 - (107/576)\epsilon^4).$$

[1] H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, MA, 1950), footnote p. 73.
[2] M. Arnow, J. Phys. E **9**, 372 (1976).
[3] H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, MA, 1950), p. 72.
[4] P. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953), pp. 432–3.
[5] H. Jeffreys and B. Jeffreys, *Methods of Mathematical Physics* (Cambridge, U.P., London, 1972) pp. 667–72.
[6] A. Ll. Hughes and V. Rojansky, Phys. Rev. **34**, 284 (1929).
[7] A. Nayfeh, *Perturbation Methods* (Wiley, New York, 1973), pp. 58–60.
[8] R. Struble, *Nonlinear Differential Equations* (McGraw-Hill, New York, 1962), pp. 70–1.
[9] H. Davis, *Introduction to Nonlinear Differential and Integral Equations* (Dover, New York, 1962), pp. 209–11.

A concise and accurate solution for Poiseuille flow in a plane channel

C. E. Siewert, R. D. M. Garcia,[a] and P. Grandjean
Service d'Etudes des Réacteurs et de Mathématiques Appliquées, Centre d'Etudes Nucléaires de Saclay, B.P. 2, 91190 Gif Sur Yvette, France
and
Nuclear Engineering Department, North Carolina State University, Raleigh, North Carolina 27650

(Received 10 November 1978; accepted for publication 6 November 1979)

The recently developed F_N method of solving problems in particle transport theory is used to establish a concise and accurate solution for the flow of a rarefied gas between two parallel plates. The Bhatnagar, Gross, and Krook model is used, and numerical results are given for a wide range of the Knudsen number.

I. INTRODUCTION

In two basic papers in the field of rarefied gas dynamics, Cercignani and Daneri,[1] and Cercignani[2] reported on two different methods of studying the flow of a rarefied gas between two parallel plates. In both papers[1,2] the BGK[3] model was used to describe the physical problem. Cercignani and Daneri[1] used the integral form of the particle transport equation and finite difference techniques to develop numerical results applicable to a wide range of the Knudsen number, and Cercignani[2] used the method of elementary solutions[4] to reduce the problem to one of solving a Fredholm equation for the required expansion coefficient. Additional numerical results have been obtained more recently by Boffi, De Socio, Gaffuri, and Pescatore,[5] and Loyalka, Petrellis, and Storvick.[6] Here we wish to describe the F_N method[7] of solving the same problem. The method utilizes aspects of the exact elementary solutions to establish an approximate solution that is particularly concise and very economical to use from the point of view of computer-time requirements.

As discussed by Cercignani,[2] the linearized BGK model for flow in the z direction between plates a distance d apart can be written as

$$\kappa c_z + c_x(\partial/\partial x)\, h(x,\mathbf{c}) = Lh(x,\mathbf{c}), \tag{1}$$

where $\mathbf{c}$ is the molecular velocity, $h(x,\mathbf{c})$ is the perturbation of the particle distribution function from the Maxwellian and κ is proportional to the pressure gradient that causes the flow. For the BGK model, Cercignani[2] uses the appropriate form of the collision operator L, and considers

$$Z(x,c_x) = \frac{1}{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(c_y^2+c_z^2)}\, c_z\, h(x,\mathbf{c})\, dc_y\, dc_z, \tag{2}$$

to be the basic unknown, and thus reduces the problem to one of solving

$$\tfrac{1}{2}\kappa\theta + \theta c_x(\partial/\partial x)\, Z(x,c_x) + Z(x,c_x) = \pi^{-1/2}\int_{-\infty}^{\infty} e^{-c_x^2} Z(x,c_x)\, dc_x, \tag{3}$$

subject to the boundary conditions

$$Z[-(d/2)\,\mathrm{sgn}c_x, c_x] = 0. \tag{4}$$

In Eq. (3) the mean-free-time is denoted by θ. In the next section we use the F_N method to deal with Eqs. (3) and (4) and thus to establish a concise result for the flow rate Q.

II. ANALYSIS

In regard to Eqs. (3) and (4), we prefer to let $\mu = c_x$, $\tau = x/\theta$, $\delta = d/\theta$ and thus to consider

$$\tfrac{1}{2}\kappa\theta + \mu(\partial/\partial\tau)\, Z(\tau,\mu) + Z(\tau,\mu) = \pi^{-1/2}\int_{-\infty}^{\infty} e^{-\mu^2} Z(\tau,\mu)\, d\mu, \tag{5}$$

and

$$Z[-(\delta/2)\,\mathrm{sgn}\mu,\mu] = 0, \quad \mu\in(-\infty,\infty). \tag{6}$$

If we substitute

$$Z(\tau,\mu) = \tfrac{1}{2}\kappa\theta\,[\tau^2 - 2\tau\mu + 2\mu^2 - (\delta^2/4) - 2Y(\tau,\mu)], \tag{7}$$

into Eqs. (5) and (6) then we see at once that $Y(\tau,\mu)$ is the solution of

$$\mu(\partial/\partial\tau)\, Y(\tau,\mu) + Y(\tau,\mu) = \pi^{-1/2}\int_{-\infty}^{\infty} e^{-\mu^2} Y(\tau,\mu)\, d\mu, \tag{8}$$

subject to

$$Y(-a,\mu) = Y(a,-\mu) = \mu^2 + a\mu, \quad \mu > 0, \tag{9}$$

where $2a = \delta$. In order to simplify the calculation of the flow rate Q, we first wish to note several useful relationships concerning some moments of $Y(\tau,\mu)$. If we let

$$Y_\alpha(\tau) = \pi^{-1/2}\int_{-\infty}^{\infty} e^{-\mu^2} Y(\tau,\mu)\,\mu^\alpha\, d\mu, \tag{10}$$

then we can multiply Eq. (8) by $\exp(-\mu^2)$ and integrate over all μ to deduce that $Y_1(\tau)$ is a constant, say $Y_1(a)$. Multiplying Eq. (8) by $\mu\exp(-\mu^2)$ and integrating over all μ, we find

$$(d/d\tau)\, Y_2(\tau) + Y_1(a) = 0, \tag{11}$$

which, after we multiply by τ and integrate over τ from $-a$ to a, yields

$$\int_{-a}^{a} Y_2(\tau)\, d\tau = 2aY_2(a). \tag{12}$$

If we multiply Eq. (8) by $\mu^2\exp(-\mu^2)$ and integrate over μ, we find

$$(d/d\tau)\, Y_3(\tau) + Y_2(\tau) = \tfrac{1}{2}Y_0(\tau), \tag{13}$$

[a] Permanent address: Instituto de Pesquisas Energeticas e Nucleares, Cidade Universitaria, São Paulo, Brasil

 0022-2488/80/122760-04$1.00

which we can integrate over τ to find, after using Eq. (12),

$$\int_{-a}^{a} Y_0(\tau)\,d\tau = 4Y_3(a) + 4aY_2(a)\,. \tag{14}$$

Now since the flow rate is

$$Q(a) = -\frac{1}{\kappa da}\int_{-a}^{a} q(\tau)\,d\tau\,, \tag{15}$$

where the macroscopic velocity is

$$q(\tau) = \pi^{-1/2}\int_{-\infty}^{\infty} e^{-\mu^2} Z(\tau,\mu)\,d\mu\,, \tag{16}$$

we can use Eqs. (7) and (15) to express the flow rate simply in terms of surface quantities, i.e.,

$$Q(a) = \frac{a}{3} - \frac{1}{2a} + \frac{2}{a^2}\left[Y_3(a) + aY_2(a)\right]. \tag{17}$$

If we use

$$Y_2(a) = \pi^{-1/2}\int_0^{\infty} \mu^2 e^{-\mu^2} Y(a,\mu)\,d\mu + \tfrac{3}{8} + \pi^{-1/2}\frac{a}{2}\,, \tag{18a}$$

and

$$Y_3(a) = \pi^{-1/2}\int_0^{\infty} \mu^3 e^{-\mu^2} Y(a,\mu)\,d\mu - \tfrac{3}{8}a - \pi^{-1/2}\,, \tag{18b}$$

in Eq. (17) we can write

$$Q(a) = \frac{a}{3} - \frac{1}{2a} + \pi^{-1/2}\left(1 - \frac{2}{a^2}\right) + \pi^{-1/2}\frac{2}{a^2} \times \int_0^{\infty} \mu^2 e^{-\mu^2} Y(a,\mu)(\mu + a)\,d\mu\,, \tag{19}$$

so that $Q(a)$ finally is expressed in terms only of $Y(a,\mu), \mu > 0$.

We now wish to consider the boundary-value problem defined by Eqs. (8) and (9). The desired symmetrical solution, $Y(\tau,\mu) = Y(-\tau,-\mu)$, can be expressed in terms of the elementary solutions[8] as

$$Y(\tau,\mu) = A\pi^{-1/2} + \int_0^{\infty} A(\nu)\left[\phi(\nu,\mu)\,e^{-\tau/\nu} + \phi(-\nu,\mu)\,e^{\tau/\nu}\right]d\nu\,, \tag{20}$$

where

$$\phi(\nu,\mu) = \pi^{-1/2}\,\nu P v\left(\frac{1}{\nu-\mu}\right) + \pi^{-1/2}\,p(\nu)\,\delta(\nu-\mu)\,. \tag{21}$$

Here

$$p(\nu) = \pi^{1/2}\left(e^{\nu^2} - 2\nu\int_0^{\nu} e^{x^2}\,dx\right), \tag{22}$$

and the expansion coefficients A and $A(\nu)$ are to be determined by the boundary condition, Eq. (9). To proceed with the method of elementary solutions we would substitute Eq. (20) into Eq. (9) and regularize the resulting singular integral equation to obtain ultimately a Fredholm-type integral equation for $A(\nu)$. Since we have expressed the desired flow rate $Q(a)$ in terms of $Y(a,\mu)$, [see Eq. (19)] we do not need $Y(\tau,\mu)$ for all τ, and thus we do not pursue the method of elementary solutions further. Instead, we pay special attention to establishing $Y(a,\mu)$. Since the functions $\phi(\nu,\mu)$ are orthogonal, in the sense that

$$\int_{-\infty}^{\infty} e^{-\mu^2}\phi(\nu,\mu)\,\phi(\nu',\mu)\,\mu\,d\mu = 0, \quad \nu \neq \nu'\,, \tag{23a}$$

and

$$\int_{-\infty}^{\infty} e^{-\mu^2}\phi(\nu,\mu)\,\mu\,d\mu = 0\,, \tag{23b}$$

we can multiply Eq.(20), evaluated at $\tau = \pm a$, by $\mu\exp(-\mu^2)\,\phi(-\nu,\mu)$ and integrate over all μ to find

$$\int_{-\infty}^{\infty} e^{-\mu^2}\phi(-\nu,\mu)\,Y(\mp a,\mu)\,\mu\,d\mu = A(\nu)\,N(-\nu)\,e^{\mp a/\nu}\,, \tag{24}$$

where $N(-\nu)$ is a normalization factor that can be eliminated between the two forms of Eq. (24) to yield

$$\int_{-\infty}^{\infty} e^{-\mu^2}\phi(-\nu,\mu)\,Y(-a,\mu)\,\mu\,d\mu = e^{-2a/\nu}\int_{-\infty}^{\infty} e^{-\mu^2}\phi(-\nu,\mu)\,Y(a,\mu)\,\mu\,d\mu\,, \tag{25}$$

or, after we use Eq. (9),

$$\int_0^{\infty} e^{-\mu^2}\phi(\nu,\mu)\,Y(a,\mu)\,\mu\,d\mu + e^{-2a/\nu} \times \int_0^{\infty} e^{-\mu^2}\phi(-\nu,\mu)\,Y(a,\mu)\,\mu\,d\mu = K(\nu)\,. \tag{26}$$

Here the known function $K(\nu)$ is given by

$$K(\nu) = \int_0^{\infty} e^{-\mu^2}\phi(-\nu,\mu)(\mu^2 + a\mu)\,\mu\,d\mu + e^{-2a/\nu} \times \int_0^{\infty} e^{-\mu^2}\phi(\nu,\mu)(\mu^2 + a\mu)\,\mu\,d\mu\,. \tag{27}$$

In a similar manner, we can multiply Eq. (20), evaluated at $\tau = a$, by $\mu\exp(-\mu^2)$, and integrate over all μ to find

$$\int_0^{\infty} e^{-\mu^2} Y(a,\mu)\,\mu\,d\mu = \int_0^{\infty} e^{-\mu^2}(\mu^2 + a\mu)\,\mu\,d\mu\,. \tag{28}$$

Equations (26) and (28) constitute a singular integral equation and a constraint to be solved to establish $Y(a,\mu)$. It is clear that the methods of Muskhelishvili[9] could be used to convert Eqs. (26) and (28) to a Fredholm-like integral equation for $Y(a,\mu)$. However, we prefer here to introduce the F_N method[7] and thus to substitute the approximation

$$Y(a,\mu) = \mu(\mu + a)\,\theta(a)\,e^{-2a/\mu} + \sum_{\alpha=0}^{N} a_\alpha\left[1 - (-1)^\alpha\,\theta(a)\,e^{-2a/\mu}\right]\mu^\alpha\,, \quad \mu > 0\,, \tag{29}$$

where the constants a_α are to be determined, into Eqs. (26) and (28) to obtain

$$\sum_{\alpha=0}^{N} a_\alpha\left[B_\alpha(\nu) - \theta(a)(-1)^\alpha D_\alpha(\nu) + e^{-2a/\nu}\left[A_\alpha(\nu) - \theta(a)(-1)^\alpha C_\alpha(\nu)\right]\right] = R(\nu), \tag{30}$$

and

$$\sum_{\alpha=0}^{N} a_\alpha \left[K_\alpha - \theta(a)(-1)^\alpha T_{\alpha+1}(2a)\right] = K_2 + aK_1 - \theta(a)[T_3(2a) + aT_2(2a)], \tag{31}$$

where a known function is

$$R(\nu) = A_2(\nu) + aA_1(\nu) - \theta(a)[D_2(\nu) + aD_1(\nu)] + e^{-2a/\nu}\{B_2(\nu) + aB_1(\nu) - \theta(a)[C_2(\nu) + aC_1(\nu)]\}. \tag{32}$$

Here we have used the definitions

$$\nu A_\alpha(\nu) = \pi^{1/2}\int_0^\infty e^{-\mu^2}\phi(-\nu,\mu)\mu^{\alpha+1}\,d\mu, \tag{33a}$$

$$\nu B_\alpha(\nu) = \pi^{1/2}\int_0^\infty e^{-\mu^2}\phi(\nu,\mu)\mu^{\alpha+1}\,d\mu, \tag{33b}$$

$$\nu C_\alpha(\nu) = \pi^{1/2}\int_0^\infty e^{-\mu^2}\phi(-\nu,\mu)\mu^{\alpha+1}e^{-2a/\mu}\,d\mu, \tag{33c}$$

$$\nu D_\alpha(\nu) = \pi^{1/2}\int_0^\infty e^{-\mu^2}\phi(\nu,\mu)\mu^{\alpha+1}e^{-2a/\mu}\,d\mu, \tag{33d}$$

$$K_\alpha = \int_0^\infty e^{-\mu^2}\mu^{\alpha+1}\,d\mu, \tag{34}$$

and

$$T_\alpha(x) = \int_0^\infty e^{-\mu^2}e^{-x/\mu}\mu^\alpha\,d\mu. \tag{35}$$

In order to establish a solution that is accurate for all values of a, we include in Eq. (29) a term multiplied by the step function

$$\theta(a) = 1, \quad 0 \leqslant a < a_*, \tag{36a}$$

$$\theta(a) = 0, \quad a \geqslant a_*, \tag{36b}$$

where a_* is to be selected, as discussed in the next section. It is apparent that

$$K_{2n} = \frac{n!}{2}, \quad n = 0,1,2,\cdots, \tag{37a}$$

and

$$K_{2n+1} = \pi^{1/2}\frac{1\cdot3\cdot5\cdots(2n+1)}{2^{n+2}}. \tag{37b}$$

We note that

$$B_0(\nu) = A_0(\nu) = \int_0^\infty e^{-\mu^2}\mu\frac{d\mu}{\mu+\nu}, \tag{38}$$

and that the remaining $B_\alpha(\nu)$ and $A_\alpha(\nu)$ can be readily generated from

$$B_\alpha(\nu) = \nu B_{\alpha-1}(\nu) - K_{\alpha-1}, \tag{39}$$

and

$$A_\alpha(\nu) = -\nu A_{\alpha-1}(\nu) + K_{\alpha-1}. \tag{40}$$

In addition

$$C_0(\nu) = \int_0^\infty e^{-\mu^2}e^{-2a/\mu}\mu\frac{d\mu}{\mu+\nu}, \tag{41}$$

$$D_0(\nu) = e^{-2a/\nu}B_0(\nu) - \int_0^\infty \mu e^{-\mu^2}\left[\frac{e^{-2a/\mu} - e^{-2a/\nu}}{\mu-\nu}\right]d\mu, \tag{42}$$

$$C_\alpha(\nu) = -\nu C_{\alpha-1}(\nu) + T_\alpha(2a), \tag{43}$$

and

$$D_\alpha(\nu) = \nu D_{\alpha-1}(\nu) - T_\alpha(2a). \tag{44}$$

If we now choose N values of $\nu\in(0,\infty)$, say ν_β, then clearly we can solve the system of algebraic equations

$$\sum_{\alpha=0}^{N} a_\alpha\Big[B_\alpha(\nu_\beta) - \theta(a)(-1)^\alpha D_\alpha(\nu_\beta) + e^{-2a/\nu_\beta} \times [A_\alpha(\nu_\beta) - \theta(a)(-1)^\alpha C_\alpha(\nu_\beta)]\Big] = R(\nu_\beta), \quad \beta = 1,2,3,\cdots,N, \tag{45a}$$

and

$$\sum_{\alpha=0}^{N} a_\alpha\left[K_\alpha - \theta(a)(-1)^\alpha T_{\alpha+1}(2a)\right] = K_2 + aK_1 - \theta(a)[T_3(2a) + aT_2(2a)] \tag{45b}$$

to find the required constants $\{a_\alpha\}$. One of the more attractive features of the F_N method is that the known coefficients in Eqs. (45) are very simply expressed. Note, for example, that the half-width a is not required in $A_\alpha(\nu)$ and $B_\alpha(\nu)$ and that the functions $A_\alpha(\nu)$ and $B_\alpha(\nu)$ are simple combinations of polynomials and the function $B_0(\nu)$. For $a \geqslant a_*$ it is thus evident that very little computer time will be required to compute the coefficients $\{a_\alpha\}$. For $a < a_*$ the coefficients in Eqs. (45) involve also the functions $C_\alpha(\nu)$, $D_\alpha(\nu)$ and $T_\alpha(2a)$; however as can be seen in the next section only a low value of N is required for $a < a_*$ to establish accurate results.

We note that the idea of using the Placzek Lemma[10] and approximating unknown surface distributions by polynomials has been used in the fields of kinetic theory[11,12] and neutron transport theory.[13] The F_N method, with $\theta(a) = 0$, clearly is related to this earlier work though it differs substantially in the way the required constants are determined.

III. NUMERICAL RESULTS

Of course to solve the system of equations given by Eq. (45) we first must select N values of $\nu_\beta\in(0,\infty)$. To have a simple and effective scheme we take the ν_β, $\beta = 1,2,3,\cdots,N$, to be the N positive zeros of the Hermite polynomial $H_{2N}(\xi)$. If we substitute Eq. (29) into Eq. (19) we find that our solution, by the F_N approximation, is

$$Q(a) = \frac{a}{3} + \pi^{-1/2}[1 + 2\theta(a)T_3(2a)] + \frac{2}{a}\pi^{-1/2} \times\Big[\sum_{\alpha=0}^{N} a_\alpha\left[K_{\alpha+1} - \theta(a)(-1)^\alpha T_{\alpha+2}(2a)\right] - \tfrac{1}{4}\pi^{1/2} + 2\theta(a)T_4(2a)\Big] + \frac{2}{a^2}\pi^{-1/2} \times\Big[\sum_{\alpha=0}^{N} a_\alpha\left[K_{\alpha+2} - \theta(a)(-1)^\alpha T_{\alpha+3}(2a)\right] - 1 + \theta(a)T_5(2a)\Big]. \tag{46}$$

TABLE I. The flow rate $Q(a)$.

$2a$	$\theta(a)$	$Q(a)$								
		F_0	F_2	F_4	F_6	F_8	F_{10}	"Present work"	Ref. 5	Ref. 6
0.001	1	4.2736						4.2736	4.2736	...
0.01	1	3.0495	3.0496					3.0496	3.0497	...
0.1	1	2.0314	2.0327					2.0327	2.0327	2.0327
0.5	1	1.5952	1.6018	1.6019				1.6019	1.6019	1.6018
1.0	1	1.5264	1.5385	1.5387				1.5387	1.5387	1.5386
2.0	1	1.5761	1.5944	1.5948				1.5948	1.5948	1.5948
3.0	1	1.6893	1.7099	1.7104	1.7105			1.7105	1.7105	1.7105
5.0	0	1.9504	1.9881	1.9905	1.9906	1.9906	1.9907	1.9907	1.9907	1.9907
7.0	0	2.2708	2.2932	2.2947	2.2948	2.2948	2.2949	2.2949	2.2948	2.2949
8.0	0	2.4304	2.4498	2.4510	2.4511	2.4511	2.4512	2.4512	2.4510	...
9.0	0	2.5906	2.6081	2.6090	2.6091	2.6092		2.6092	2.6090	2.6092
10.0	0	2.7514	2.7677	2.7685	2.7685	2.7686		2.7686	2.7684	2.7686
20.0	0	4.3850	4.3969	4.3973	4.3974	4.3974		4.3974	4.3971	..
30.0	0	6.0381	6.0489	6.0492	6.0492	6.0493		6.0493	6.0479	...
40.0	0	7.6976	7.7077	7.7080	7.7081			7.7081	...	...
100.0	0	17.684	17.693					17.693	...	...

In Table I we show, in addition to the results of Boffi *et al.*,[5] and Loyalka *et al.*,[6] the values obtained by using the solutions of Eq. (45) in Eq. (46). In addition to the results for various orders of the F_N approximation we list as "present work" the stable results we believe to be correct to within $\pm$ 1 in the fifth significant figure.

We have found that the approximation given by Eq. (29) with $\theta(a) = 1$ works well for all values of $2a$ listed in the table. However for $2a \geq 5.0$ we were able to obtain $Q(a)$ accurate to five significant figures with $\theta(a) = 0$ and thus with a greatly reduced requirement for computation time. If the desired accuracy in $Q(a)$ is reduced to four significant figures then $\theta(a) = 0$ can be used for all $2a \geq 1.0$. Finally we note that the F_N solution developed here is especially simple with $\theta(a) = 0$ since only $B_0(\nu)$ and the recursive formulas, Eqs. (39) and (40), are required to define the matrix elements in Eqs. (45).

ACKNOWLEDGMENT

One of the authors (CES) is grateful to the Centre d'Etudes Nucléaires de Saclay for kind hospitality and partial support of this work. This work was also supported by CNEN and IPEN, both of Brasil, and the U.S. National Science Foundation through grant ENG 7709405.

[1] C. Cercignani and A. Daneri, J. Appl. Phys. **34**, 3509 (1963).
[2] C. Cercignani, J. Math. Anal. Appl. **12**, 254 (1965).
[3] P.L. Bhatnagar, E.P. Gross, and M. Krook, Phys. Rev. **94**, 511 (1954).
[4] K.M. Case, Ann. Phys. **9**, 1 (1960).
[5] V. Boffi, L. De Socio, G. Gaffuri, and C. Pescatore, Meccanica **11**, 183 (1976).
[6] S.K. Loyalka, N. Petrellis, and T.S. Storvick, Z.A.M.P. **30**, 514 (1979).
[7] C.E. Siewert and P. Benoist, Nucl. Sci. Eng. **69**, 156 (1979).
[8] C. Cercignani, Ann. Phys. **20**, 219 (1962).
[9] N.I. Muskhelishvili, *Singular Integral Equations* (Noordhoff, Groningen, The Netherlands, 1953).
[10] K.M. Case, F. de Hoffmann, and G. Placzek, *Introduction to the Theory of Neutron Diffusion*, Vol. 1 (U.S. Government Printing Office, Washington, D.C., 1953).
[11] J.K. Buckner and J.H. Ferziger, Phys. Fluids **9**, 2315 (1966).
[12] S.K. Loyalka and J.H. Ferziger, Phys. Fluids **10**, 1833 (1967).
[13] P. Benoist and A. Kavenoky, Nucl. Sci. Eng. **32**, 225 (1968).

Operator methods for time-dependent waves in random media with applications to the case of random particles

K. Furutsu
Radio Research Laboratories, Koganei-shi, Tokyo 184, Japan

(Received 24 April 1980; accepted for publication 20 June 1980)

The random medium is represented by the operator, constructed from the characteristic functional of the medium, and this representation is shown to considerably facilitate the formulation of various equations of waves in random media, as well as obtaining the physical insight into the equations. A specific application is made to waves in the medium of random particles, and the equations obeyed by the characteristic functional of wave are derived with the aid of the effective medium method. Here, the optical condition is exhibited by the condition of an operator in space and time. Independent of this operator method, the general theory is extended, in an unperturbative way, for the equations of the second-order coherence functions, being given in a form of the Bethe–Salpeter equation, and the coherent potential equations are formulated for the basic matrices of two kinds appeared in the equations. The explicit expressions of these matrices are obtained, on utilizing the coherent potential approximation, and are shown to be exactly the same as those obtained by the effective medium method, in both cases of weak-scattering limit and of random particles. Finally, on employing the appropriate Fourier representations in space and time, the theory is presented in a few different forms, one being particularly suited to derive the equations of multifrequency coherence functions.

1. INTRODUCTION AND PRELIMINARIES

The statistical theory of waves in random media has been extensively developed in terms of the mutual coherence function of wave for the typical random media of both turbulent air where the scattering of wave is weak and made mostly in the forward direction, and random particles where the scattering by each particle is usually strong and described in terms of the scattering matrix. In either case, the mutual coherence function of wave has been shown to obey the equation of a form of the Bethe–Salpeter equation,[1–5] and further to obey the ordinary transport equation which can be derived from the former equation to a good approximation.[6]

In case of turbulent air, the higher order coherence functions of wave also have been investigated in connection with the saturation of irradiance scintillation, the probability distribution of irradiance, etc., and the governing equations have been systematically derived with the exact solutions of all orders in the special case when the medium structure function can be given in the parabolic form.[7] In case of random particles, on the other hand, no higher order coherence functions have been investigated nor any systematic way of deriving their governing equations.

Since the complete statistical informations of a wave in random media can be obtained through the characteristic functional of wave, the basic problem is reduced to finding the equations satisfied by this functional, provided the characteristic functional of the medium is on the other side. Here, to find those equations, the medium has been represented by the operator, constructed from the characteristic functional of the medium, which particularly facilitates obtaining the expectation value for any functional of the medium.[7] Here, in the case of the media obeying the Gaussian statistics, this operator representation of the medium leads to the previous formula, which has been extensively used to derive the equations of the coherence functions of wave.[3,8] Also for the medium of random particles, the corresponding representation has been found in the time-independent case[9] and employed to derive the equation for the mutual coherence function of a wave.

With the replacement of the medium and also wave-functions by the corresponding operators, the equations for the characteristic functional of a wave are found to preserve the same forms as the original wave equations. This is a consequence of the more general correspondence principle established between any equations which hold when the medium is deterministic and the corresponding equations when the medium is probablistic; this is the case, e.g., of the equation of continuity of a wave, the equations of energy-momentum conservation, constructed according to the Lagrangian principle, etc.[7]

In this paper, the random medium is first represented by the operator in terms of the characteristic functional of the medium, which is described on the same footing in space and time, and its explicit expressions are derived for both cases of weak-scattering limit and of random particles (Sec. 2). The basic equations obeyed by the characteristic functional of a wave are then derived together with the correspondence principle (Sec. 3). In Sec. 4, the specific application is made to obtain the explicit equations satisfied by the first and second order statistical Green's functions in the medium of random particles, where exclusive use is made of the operator techniques together with the effective medium method. In Sec. 5, entirely independent of the above operator method, the general theory is extended for the same Green's functions, where two basic matrices are introduced and explicitly defined in an unperturbative manner; these matrices are shown to satisfy the optical condition in the

 0022-2488/80/122764-16$1.00

generalized sense. In Sec. 6, the coherent potential equations are formulated to find the two basic matrices according to the definitions in Sec. 5 and their explicit expressions are obtained, on utilizing the coherent potential approximation which has been used successfully for the impurity problems in solid physics, to show their exact equivalence to those by the effective medium method introduced in Sec. 4. Finally in Sec. 7, the equations satisfied by the characteristic functional of a wave are derived for the medium of random particles, based again on the effective medium method, and the optical condition is shown to be exhibited by the condition imposed on a space-time operator.

We employ the following notations: The space coordinate vector is denoted $\mathbf{x} = (x_1, x_2, x_3)$, the time by t, and $x = (\mathbf{x}, t)$ represents the space-time coordinate vector. The space-time element of volume is defined by $dx = d\mathbf{x}\, dt$ with $d\mathbf{x} = dx_1 dx_2 dx_3$. The wave function is designated by $\psi(x)$ and is assumed to satisfy an equation of the form

$$[L(i\partial/\partial x) - q(x)]\psi(x) = j(x),$$
$$[L^*(-i\partial/\partial x) - q(x)]\psi^*(x) = j^*(x). \tag{1.1}$$

Here, the asterisk stands for the complex conjugate, and $q(x) = q^*(x)$ designates the medium, including the random part; $j(x)$ is the external source of a wave and, in the case of the scalar wave,

$$L(i\partial/\partial x) = L^*(-i\partial/\partial x) = \left(\frac{1}{c}\frac{\partial}{\partial t}\right)^2 - \left(\frac{\partial}{\partial \mathbf{x}}\right)^2, \tag{1.2}$$

where c is the wave velocity.

In the case when N particles are enclosed in a finite space of volume V, and $q_\alpha(x)$ is the contribution from one particle characterized by the symbol α, then, $q(x)$ in Eq. (1.1) is given by

$$q(x) = \sum_{j=1}^{N} q_{\alpha_j}(x), \tag{1.3}$$

where α_j is the particular value of α, specifying the jth particle involved. For example, when the particles have time-independent structures and are all moving with a constant velocity $\mathbf{v}$, then, $q_\alpha(x)$ is given in the form

$$q_\alpha(x) = q(\mathbf{x} - \boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = \mathbf{v}t + \mathbf{a}, \tag{1.4}$$

where $\boldsymbol{\alpha}$ is the space coordinates of the center of the particle, and the other parameters, necessary to specify various particle structures and properties, have been suppressed.

2. OPERATOR REPRESENTATION OF RANDOM MEDIUM

It is known that the complete statistical information of any random medium $q(x)$ can be obtained, if the medium is stationary in space and time, from the characteristic functional, defined by

$$Z_q[p] = \left\langle \exp\left[\int dx\, p(x)q(x)\right]\right\rangle, \quad Z_q[0] = 1, \tag{2.1}$$

where $\langle\cdots\rangle$ stands for the statistical averaging of the quantity referred to over all possible values of the medium $q(x)$, and $p(x)$ is an arbitrary function. Whence

$$\left[\frac{\delta}{\delta p(x)}\right]^n Z_q[p]\Big|_{p=0} = \langle q^n(x)\rangle, \quad n = 1, 2, 3, \cdots, \tag{2.2}$$

and, for any functional $f[q]$ of $q(x)$,

$$f[\delta/\delta p] Z_q[p]\,|_{p=0} = \langle f[q]\rangle, \tag{2.3}$$

giving the average value $\langle f[q]\rangle$ in terms of $Z_q[p]$.

Although the expression (2.3) provides us with a convenient means of evaluating the statistical average value of $f[q]$ when it is explicitly given, this is not the case when $f[q]$ is implicitly given, e.g., through its governing equation. To obtain an alternative expression convenient particularly in the latter case, we first note the following relation, valid for an arbitrary function $c(x)$:

$$\begin{aligned} Z_q[\delta/\delta c] f[c] &= \left\langle \exp\left[\int dx\, q(x)\frac{\delta}{\delta c(x)}\right] f[c]\right\rangle \\ &= \left\langle \sum_{n=0}^{\infty}\frac{1}{n!}\left[\int dx\, q(x)\frac{\delta}{\delta c(x)}\right]^n f[c]\right\rangle \\ &= \langle f[c+q]\rangle. \end{aligned} \tag{2.4}$$

Here, in terms of the operator $\mathbf{q}(x)$ defined by

$$\mathbf{q}(x) = Z_q[\delta/\delta c] c(x) Z_q^{-1}[\delta/\delta c], \tag{2.5}$$

when $Z_q = Z_q[\delta/\delta c]$,

$$Z_q c^2(x) Z_q^{-1} = (Z_q c(x) Z_q^{-1})(Z_q c(x) Z_q^{-1}) = \mathbf{q}^2(x), \tag{2.6}$$
$$Z_q c^n(x) Z_q^{-1} = \mathbf{q}^n(x), \quad n = 1, 2, 3, \cdots,$$

and therefore also

$$Z_q[\delta/\delta c] f[c] Z_q^{-1}[\delta/\delta c] = f[Z_q c Z_q^{-1}] = f[\mathbf{q}], \tag{2.7}$$

which enable us to exhibit the result (2.4) by[7,11]

$$\langle f[c+q]\rangle = f[\mathbf{q}] Z_q[\delta/\delta c] = f[\mathbf{q}], \tag{2.8}$$

since, in the last derivation, there is nothing for $Z_q[\delta/\delta c]$ to operate on.

From Eq. (2.8), we learn that the statistical average of any functional $f[q]$ of $q(x)$ can be found simply by replacing $q(x)$ by the operator $\mathbf{q}(x)$, and therefore also that

$$\langle [c(x) + q(x)] f[c+q]\rangle = \mathbf{q}(x) f[\mathbf{q}] = \mathbf{q}(x)\langle f[c+q]\rangle, \tag{2.9}$$

which is particularly convenient in finding $\langle q(x) f[q]\rangle$ for given $\langle f[q]\rangle$.

Here, obtaining the explicit expression of $\mathbf{q}(x)$ is facilitated by introducing the cumulant of $Z_q[p]$, defined by

$$\theta[p] = \ln\{Z_q[p]\}, \tag{2.10}$$

and hence, in terms of the conventional notation for the cummutator $[A, B] = AB - BA$, Eq. (2.5) gives

$$\begin{aligned} \mathbf{q}(x) &= \exp\{\theta[\delta/\delta c]\} c(x) \exp\{-\theta[\delta/\delta c]\} \\ &= c(x) + [\theta, c(x)] + (1/2!)[\theta, [\theta, c(x)]] + \cdots. \end{aligned} \tag{2.11}$$

Here,

$$[\theta[\delta/\delta c], c(x)] = q(x, \delta/\delta c), \tag{2.12}$$

with

$$q(x, p) \equiv (\delta/\delta p(x))\theta[p] = (\delta/\delta p(x)) \ln Z_q[p], \tag{2.13}$$

and therefore it follows that, on the right-hand side of Eq. (2.11), the nonvanishing terms are only the first two terms, yielding

$$\mathbf{q}(x) = c(x) + q(x, \delta/\delta c). \tag{2.14}$$

Here, the operators $\mathbf{q}(x)$ at all points in space and time are commutable with each other, i.e.,

$$[\mathbf{q}(x),\mathbf{q}(x')] = Z_q[\delta/\delta c][c(x)c(x') - c(x')c(x)]$$
$$\times Z_q^{-1}[\delta/\delta c] = 0, \tag{2.15}$$

as follows directly from the definition (2.5), exhibited by the similarity transformation of $c(x)$.

A. Gaussian medium

With the definition (2.10), we obtain, when $\langle q(x)\rangle = 0$,

$$\theta[p] = \tfrac{1}{2}\int dx_1\, dx_2\, \langle q(x_1)q(x_2)\rangle p(x_1)p(x_2)$$
$$+ \tfrac{1}{6}\int dx_1\, dx_2\, dx_3\, \langle q(x_1)q(x_2)q(x_3)\rangle$$
$$\times p(x_1)p(x_2)p(x_3) + \cdots, \tag{2.16}$$

which gives, according to the definition (2.13),

$$q(x, p) = \int dx_1 \langle q(x)q(x_1)\rangle p(x_1)$$
$$+ \tfrac{1}{2}\int dx_1\, dx_2 \langle q(x)q(x_1)q(x_2)\rangle p(x_1)p(x_2) + \cdots. \tag{2.17}$$

Therefore, when the contributions of the second- and higher-order terms on the right-hand side of Eq. (2.17) are negligible, we find, according to Eq. (2.14),

$$\mathbf{q}(x) \sim c(x) + \int dx' D(x - x')\frac{\delta}{\delta c(x')},$$
$$D(x - x') = \langle q(x)q(x')\rangle, \tag{2.18}$$

which describes the medium obeying the Gaussian statistics, as may be seen directly by applying the formula (2.8) to evaluate $Z_q[p]$ according to the definition (2.1). Whence,

$$Z_q[p] = \exp\left[\int dx\, p(x)\mathbf{q}(x)\right]$$
$$= \exp\left\{\int dx\, p(x)\left[c(x) + \int dx'\, D(x - x')\frac{\delta}{\delta c(x')}\right]\right\}, \tag{2.19}$$

which, as $c \to 0$, tends to

$$Z_q[p] = \exp\left[\tfrac{1}{2}\int dx\, dx'\, D(x - x')p(x)p(x')\right], \tag{2.20}$$

with the aid of the formula

$$\exp(A + B) = \exp(A)\exp(B)\exp(\tfrac{1}{2}[B, A]), \tag{2.21}$$

valid for arbitrary operators A and B when the commutator $[B, A]$ is commutable with both A and B, being the present case of $A =$

$$\int dx\, p(x)c(x) \quad \text{and} \quad B = \int dx\, dx'\, p(x)D(x - x')\delta/\delta c(x').$$

Here, the formula (2.9) with the expansion (2.17) for the term $q(x, \delta/\delta c)$ in $\mathbf{q}(x)$, suggests that the assumption that the Gaussian statistics are obeyed will give a good approximation to the real random medium when the magnitude of q is small enough so that the accumulated effect of $q(x)$ over the range of its correlation distance is negligibly small for $\langle f[q]\rangle$. The relation (2.9) with $\mathbf{q}(x)$ given by Eq. (2.18) is equivalent to that previously obtained.[3,8]

B. Multi-component random medium

When the medium is composed of two independent random components, as given by

$$q(x) = q_1(x) + q_2(x), \tag{2.22}$$

then, by Eq. (2.1)

$$Z_q[p] = \left\langle \exp\left\{\int dx\, p(x)[q_1(x) + q_2(x)]\right\}\right\rangle = Z_1[p]Z_2[p], \tag{2.23}$$

$Z_j[p], j = 1, 2$, being the characteristic functional for q_j alone, and hence

$$q(x, p) = \frac{\delta}{\delta p(x)}\ln\{Z_1[p]Z_2[p]\} = q_1(x, p) + q_2(x, p),$$
$$q_j(x, p) = \frac{\delta}{\delta p(x)}\ln Z_j[p], \quad j = 1, 2, \tag{2.24}$$
$$\mathbf{q}(x) = c(x) + q_1(x, \delta/\delta c) + q_2(x, \delta/\delta c), \tag{2.25}$$

which shows that the contributions from the independent components of random medium are simply added up to construct $q(x, \delta/\delta c)$ in $\mathbf{q}(x)$.

C. Medium of random particles

We suppose that N particles are randomly distributed in a space of volume V without any correlation to each other, allowing, strictly speaking, even overlapping of particles, and also that, as in Eq. (1.3), the contribution from each particle to the total $q(x)$ is made through the function $q_\alpha(x)$. Here, the symbol α represents the set of parameters characterizing the structure of one particle and therefore includes the space coordinates of particle's center, say, $\mathbf{a}$ at a particular time, besides other parameters specifying, e.g., its size, shape, orientation, trajectory in space and time, etc. Therefore, the characteristic functional $Z_q[p]$, defined by Eq. (2.1) with $q(x)$ given by Eq. (1.3), is found, in the manner similar to Eq. (2.23), to be

$$Z_q[p] = \prod_{j=1}^{N} V^{-1}\int_V d\mathbf{a}_j \left\langle \exp\left[\int dx\, p(x)q_{\alpha_j}(x)\right]\right\rangle'$$
$$= \left[V^{-1}\int_V d\mathbf{a} \left\langle \exp\left[\int dx\, p(x)q_\alpha(x)\right]\right\rangle'\right]^N, \tag{2.26}$$

where the bracket $\langle\cdots\rangle'$ means the averaging over all possible properties of the involved particles, excluding that over the center coordinates $\mathbf{a}$.

Here, as $V \to \infty$ and $N \to \infty$, keeping a constant density of particles $n = N/V$, Eq. (2.26) tends to

$$Z_q[p] = \exp\left[n\int d\mathbf{a} \left\{\left\langle \exp\left[\int dx\, p(x)q_\alpha(x)\right]\right\rangle' - 1\right\}\right]. \tag{2.27}$$

Thus, according to Eq. (2.13),

$$q(x, p) = n\int d\mathbf{a} \left\langle q_\alpha(x)\exp\left[\int dx'\, p(x')q_\alpha(x')\right]\right\rangle', \tag{2.28}$$

and hence, by Eq. (2.14), $\mathbf{q}(x)$ is found to be given by

$$\mathbf{q}(x) = c(x) + q(x, \delta/\delta c), \tag{2.29}$$

$$q(x, \delta/\delta c) = \left\langle q_\alpha(x) \exp\left[\int dx'\, q_\alpha(x') \frac{\delta}{\delta c(x')}\right]\right\rangle_\alpha,$$

in terms of the notation

$$\langle \cdots \rangle_\alpha \equiv n \int d\mathbf{a}\, \langle \cdots \rangle'. \tag{2.30}$$

The expression (2.29) will be exclusively used in Secs. 4 and 7 to find various statistical equations of a wave in the medium of random particles.[12]

3. EQUATIONS SATISFIED BY THE CHARACTERISTIC FUNCTIONAL OF A WAVE

In exactly the same way as for random media, the complete statistical description of the wave function $\psi(x)$ and of the complex conjugate wave function $\psi^*(x)$ in a random medium can be derived from the characteristic functional of a wave, defined by[10]

$$Z[\bar{j}^*, \bar{j}] = \left\langle \exp\left\{\int dx\, [\bar{j}(x)\psi(x) + \bar{j}^*(x)\psi^*(x)]\right\}\right\rangle. \tag{3.1}$$

To find the equations obeyed by $Z[\bar{j}^*, \bar{j}]$, we assume the wave equations of the form (1.1), i.e.,

$$[L(i\partial/\partial x) - c(x) - q(x)]\psi(x) = j(x) \tag{3.2}$$

and the corresponding complex conjugate wave equation, where $c(x)$ is an infinitesimal function and is to vanish in the final results. Here, it is straightforward, on employing

$$\frac{\delta}{\delta \bar{j}(x)} Z[\bar{j}^*, \bar{j}]$$

$$= \left\langle \psi(x) \exp\left\{\int dx' [\bar{j}(x')\psi(x') + \bar{j}^*(x')\psi^*(x')]\right\}\right\rangle, \tag{3.3}$$

and also the wave equation (3.2), to find

$$\left\langle \{[L(i\partial/\partial x) - c(x) - q(x)]\delta/\delta \bar{j}(x) - j(x)\} \right.$$

$$\left. \times \exp\left\{\int dx' [\bar{j}(x')\psi(x') + \bar{j}^*(x')\psi^*(x')]\right\}\right\rangle = 0, \tag{3.4}$$

where, from the formula (2.9),

$$\langle [c(x) + q(x)] \exp\{\quad\}\rangle = \mathbf{q}(x) Z[\bar{j}^*, \bar{j}]. \tag{3.5}$$

The above result can be expressed in a compact form by introducing the operators, defined by

$$\psi(x) = \frac{\delta}{\delta \bar{j}(x)}, \quad \psi^*(x) = \frac{\delta}{\delta \bar{j}^*(x)}, \tag{3.6}$$

with the commutation relations

$$[\psi(x), \bar{j}(x')] = [\psi^*(x), \bar{j}^*(x')] = \delta(x - x'),$$

$$[\psi(x), \bar{j}^*(x')] = [\psi^*(x), \bar{j}(x')] = [\psi^*(x), \psi(x')] = 0. \tag{3.7}$$

Whence

$$\{[L(i\partial/\partial x) - \mathbf{q}(x)]\psi(x) - j(x)\} Z[\bar{j}^*, \bar{j}] = 0 \tag{3.8}$$

and, in the same way,

$$\{[L^*(-i\partial/\partial x) - \mathbf{q}(x)]\psi^*(x) - j^*(x)\} Z[\bar{j}^*, \bar{j}] = 0, \tag{3.9}$$

which, together with Eq. (3.8), constitutes the basic equations of $Z[\bar{j}^*, \bar{j}]$. The time-independent version of Eqs. (3.8) and (3.9) are exactly the same as previously obtained by an entirely different method.[7]

Comparing Eqs. (3.8) and (3.9) with the original wave equations in Eq. (1.1), we immediately find the obvious one-to-one correspondence; more generally, it is not difficult to show that, if the relation $Q[\psi, \psi^*, q] = 0$ holds among ψ, ψ^*, and q when the medium is deterministic, then

$$Q[\psi, \psi^*, \mathbf{q}] Z[\bar{j}^*, \bar{j}] = 0, \tag{3.10}$$

when the medium is probabilistically given, and Eq. (3.8) is regarded as the particular case of when $Q = (L - q)\psi - j = 0$. For another example, the equation of continuity of the scalar wave, satisfying Eq. (1.1) with (1.2), is given in the form

$$\sum_{j=0}^{3} \frac{\partial}{\partial x_j} F_j + \frac{i}{2}(\psi^* j - \psi j^*) = 0, \tag{3.11}$$

where $x_0 = ct$, c = wave velocity, $\partial/\partial x^j = (\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3, -\partial/\partial x_0)$ and

$$F_j = \frac{i}{2}\left[\psi^* \frac{\partial}{\partial x^j}\psi - \psi \frac{\partial}{\partial x^j}\psi^*\right], \quad j = 0, 1, 2, 3. \tag{3.12}$$

Therefore, $Z[\bar{j}^*, \bar{j}]$ also satisfies the equation corresponding to Eq. (3.11):

$$\left[\sum_{j=0}^{3} \frac{\partial}{\partial x_j}\mathbf{F}_j + \frac{i}{2}(\psi^* j - \psi j^*)\right] Z[\bar{j}^*, \bar{j}] = 0. \tag{3.13}$$

Here, $\mathbf{F}_j$ is the same function of ψ, ψ^*, and $\mathbf{q}$ as F_j is of ψ, ψ^*, and q.

In the same way, we could construct the energy-stress tensor of a wave according to the Lagrangian principle and find the related conservation equations also for $Z[\bar{j}^*, \bar{j}]$. In this case, however, the equations contain the first-order derivatives of $\mathbf{q}(x)$ in space and time, in constrast with the equation of continuity (3.13), showing that neither energy nor momentum is conserved when the medium fluctuates in space and time, as is generally the case of moving media.

In the power series expansion of $Z[\bar{j}^*, \bar{j}]$ with respect to $\bar{j}^*(x)$ and $\bar{j}(y)$, i.e.,

$$Z[\bar{j}^*, \bar{j}] = \sum_{\mu,\nu=0}^{\infty} \frac{1}{\mu!\nu!} \int dx_1\, dx_2 \cdots dx_\mu\, dy_1\, dy_2 \cdots dy_\nu$$
$$\times m_{\mu\nu}(x_1, \cdots, x_\mu; y_1, \cdots, y_\nu) \bar{j}^*(x_1)\bar{j}^*(x_2)$$
$$\cdots \bar{j}^*(x_\mu)\bar{j}(y_1)\bar{j}(y_2)\cdots\bar{j}(y_\nu), \tag{3.14}$$

the expansion coefficients are directly related to the moments of wave functions as:

$$m_{\mu\nu}(x_1, x_2, \cdots, x_\mu;\ y_1, y_2, \cdots, y_\nu)$$
$$= \langle \psi^*(x_1)\psi^*(x_2)\cdots\psi^*(x_\mu)\psi(y_1)\psi(y_2)\cdots\psi(y_\nu)\rangle. \tag{3.15}$$

Here, since Eqs. (3.8) and (3.9) for $Z[\bar{j}^*, \bar{j}]$ contain the undesirable operator $\delta/\delta c(x)$ through $\mathbf{q}(x)$, the next task is to find such an operator $\kappa_{01}(x|x')$ which is free from $\delta/\delta c(x)$ but a functional of $\bar{j}$, $\bar{j}^*$, ψ, and ψ^* instead, defined by [Cf. Eqs. (7.9) and (B11)]

$$\mathbf{q}(x)\psi(x) Z|_{c=0} = \int dx'\, \kappa_{01}(x|x')\psi(x') Z[\bar{j}^*, \bar{j}],$$

by invoking an appropriate approximation. This enables us

to eliminate $\delta/\delta c(x)$ from Eqs. (3.8) and (3.9) and further to derive the equations for the moments $m_{\mu\nu}$ upon substitution of the expansion (3.14).

On the other hand, a formal solution of Eqs. (3.8) and (3.9) is obtained by introducing the symbolic Green's function $\mathbf{G}_q$ and $\mathbf{G}_q^*$, defined according to the equations

$$[L(i\partial/\partial x) - \mathbf{q}(x)]\mathbf{G}_q(x|x') = \delta(x - x'),$$
$$[L^*(-i\partial/\partial x) - \mathbf{q}(x)]\mathbf{G}_q^*(x|x') = \delta(x - x'), \tag{3.16}$$

with the given boundary condition. Here, the solutions are functionals of the operator $\mathbf{q}(x)$ and therefore these Green's functions are also operators involving $\delta/\delta c(x)$; $\mathbf{G}_q$ and $\mathbf{G}_q^*$ can be defined alternatively by the integral equation

$$\mathbf{G}_q(x|x') = G_0(x|x') + \int dx''\, G_0(x|x'')\mathbf{q}(x'')\mathbf{G}_q(x''|x'), \tag{3.17}$$

where $G_0(x|x')$ is the ordinary Green's function, obeying

$$L(i\partial/\partial x)G_0(x|x') = \delta(x - x'), \tag{3.18}$$

with the given boundary condition. Here, it will be noted that, in virture of the mutual commutability of the operators $\mathbf{q}(x)$ at all points in space and time, as exhibited by Eq. (2.15), $\mathbf{G}_q$ and $\mathbf{G}_q^*$ are also mutually commutable and therefore can be treated in exactly the same way as the ordinary Green's functions.

Thus, in terms of the symbolic Green's functions satisfying Eq. (3.16), Eqs. (3.8) and (3.9) can be exhibited by

$$\frac{\delta}{\delta\bar{j}(x)}Z[\bar{j}^*,\bar{j}] = \int dx'\, \mathbf{G}_q(x|x')j(x')Z[\bar{j}^*,\bar{j}] \tag{3.19}$$

and its complex conjugate equation, whose formal solution is obviously

$$Z[\bar{j}^*,\bar{j}] = \exp\left\{\int dx\,dx'[\bar{j}(x)\mathbf{G}_q(x|x')j(x') + \bar{j}^*(x)\mathbf{G}_q^*(x|x')j^*(x')]\right\}Z_0, \tag{3.20}$$

where

$$Z_0 = Z[\bar{j}^*,\bar{j}]\,|_{\bar{j}^*=\bar{j}=0} = 1. \tag{3.21}$$

Here, $\mathbf{q}(x)$ is given by Eq. (2.29) in the medium of random particles and by Eq. (2.18) in media obeying the Gaussian statistics. The expression (3.20) could be derived more directly from Eq. (3.1) with G_q and G_q^* replaced by $\mathbf{G}_q$ and $\mathbf{G}_q^*$, respectively, according to the rule (2.8).

The moments of wave functions of various order are obtained from Eq. (3.20), in the form

$$\langle\psi(x)\rangle = \int dx'\, G(x|x')j(x'),$$
$$\langle\psi^*(x)\rangle = \int dx' G^*(x|x')j^*(x'),$$
$$\langle\psi^*(x_1)\psi(x_2)\rangle = \int dx_1'\,dx_2' G_{11}(x_1;x_2|x_1';x_2')j^*(x_1')j(x_2'),\ \text{etc.}, \tag{3.22}$$

where

$$G(x|x') = \mathbf{G}_q(x|x')Z_0, \quad G^*(x|x') = \mathbf{G}_q^*(x|x')Z_0,$$
$$G_{11}(x_1;x_2|x_1';x_2') = \mathbf{G}_q^*(x_1|x_1')\mathbf{G}_q(x_2|x_2')Z_0,\ \text{etc.}, \tag{3.23}$$

and the auxiliary function $c(x)$ is to vanish in the final results.

4. STATISTICAL GREEN'S FUNCTIONS OF THE FIRST AND SECOND ORDERS IN MEDIA OF RANDOM PARTICLES

The first order statistical Green's function $G(x|x')$, defined by Eq. (3.23), obeys the equation

$$[L(i\partial/\partial x) - \mathbf{q}(x)]G(x|x') = \delta(x - x'), \tag{4.1}$$

as directly follows from Eq. (3.16). Here, on the left-hand side, $\mathbf{q}(x)$ is given by Eq. (2.29) and therefore the contribution from the term of $q(x, \delta/\delta c)G(x|x')$ is expressed, by virtue of the term $c(x)$ contained in $\mathbf{q}(x)$ and in terms of the notation

$$\mathbf{G}_{q+\alpha} \equiv \mathbf{G}_q\,|_{\mathbf{q}(x)\to\mathbf{q}(x)+q_\alpha(x)}, \tag{4.2}$$

as:

$$q(x, \delta/\delta c)G(x|x') = q(x, \delta/\delta c)\mathbf{G}_q(x|x')Z_0 = \langle q_\alpha(x)\mathbf{G}_{q+\alpha}(x|x')\rangle_\alpha Z_0, \tag{4.3}$$

with the aid of the Taylor expansion similar to that in Eq. (2.4).

Here, the right-hand side of Eq. (4.3) can be exhibited in a compact form in terms of the conventional scattering matrix, defined as follows: Let $G_{a+b}(x|x')$ be the Green's function in a medium $a + b$, and also G_{a+b} be the matrix defined by its matrix elements $G_{a+b}(x|x')$ with respect to the coordinates x and x', then the equation of G_{a+b} is expressed, in matrix form, by

$$(L - a - b)G_{a+b} = 1, \tag{4.4}$$

and the solution can be given in the form

$$G_{a+b} = G_b[1 + T_a^b G_b]. \tag{4.5}$$

Here, the scattering matrix T_a^b expresses the effect caused by the scatterer a existing in the medium b, and there are several relations connecting a, b, and T_a^b as:

$$aG_{a+b} = T_a^b G_b, \quad G_{a+b}a = G_b T_a^b, \tag{4.6}$$
$$T_a^b = a(1 - G_b a)^{-1} = (1 - aG_b)^{-1}a = a + aG_b a + aG_b aG_b a + \cdots, \tag{4.7}$$
$$a = (1 + T_a^b G_b)^{-1}T_a^b = T_a^b(1 + G_b T_a^b)^{-1}. \tag{4.8}$$

Thus, applying Lemma 4.6 to the right-hand side of Eq. (4.3), the result can be exhibited, in terms of the notation

$$\mathbf{T}_\alpha^q(x|x'') = T_a^b(x|x'')\,|_{a=q_\alpha,\ b=\mathbf{q}}, \tag{4.9}$$

by

$$q(x, \delta/\delta c)G(x|x') = \int dx''\, \langle\mathbf{T}_\alpha^q(x|x'')\mathbf{G}_q(x''|x')\rangle_\alpha Z_0 = \int dx''\, \langle\mathbf{T}_\alpha^q(x|x'')\rangle_\alpha G(x''|x'). \tag{4.10}$$

Here, we introduce the matrix M, having the matrix

elements $M(x|x')$, defined by

$$\mathbf{q}(x)G(x|x')|_{c=0} = q(x, \delta/\delta c)G(x|x')|_{c=0} = \int dx''\, M(x|x'')G(x''|x'), \tag{4.11}$$

or $\mathbf{q}G = MG$, $c = 0$, in matrix form. Here, on comparing Eq. (4.11) with (4.10), we find that, if the $\mathbf{T}_\alpha^q$ matrix is negligibly correlated with the incident wave, the explicit expression of M may be given by

$$M(x|x'') = \langle \mathbf{T}_\alpha^q(x|x'')\rangle_\alpha \sim \langle T_\alpha^M(x|x'')\rangle_\alpha = n\int d\mathbf{a}\, \langle T_\alpha^M(x|x'')\rangle', \tag{4.12}$$

which has been obtained simply upon substitution of the matrix M for the operator $\mathbf{q}(x)$ in $\mathbf{T}_\alpha^q$. Here the above approximation will be valid when the dimensions of the particles are sufficiently small in comparison with the coherence distance of the wave, allowing $\mathbf{G}_q^n \sim G_M^n$, $n = 1, 2, 3, \ldots$, inside the particle $q_\alpha(x)$, and also when the correlation between its scattering matrix and the incident wave is negligibly small; the latter is possible since the effect of medium fluctuation on the scattering properties of the particles is only through those parts of the medium in the immediate neighborhood of each particle, while the fluctuation of the incident wave is due to the accumulated effect of medium fluctuation along the wave path.

The substitution of Eq. (4.11) in Eq. (4.1) yields

$$L(i\partial/\partial x)G(x|x') - \int dx''\, M(x|x'')G(x''|x') = \delta(x - x'), \tag{4.13}$$

or, in matrix form,

$$(L - M)G = 1, \quad (L^* - M^*)G^* = 1. \tag{4.14}$$

Here, the latter equation is the complex conjugate of the former and, following the notations introduced in Eqs. (4.4)–(4.8), $G = G_M$.

Equation (4.12) provides us a means of finding M for a given $q_\alpha(x)$, being one particle's contribution to the whole medium $q(x)$. Another means of finding M will be shown in Sec. VI in connection with the coherent potential approximation.

We can employ the same method also to find the equation satisfied by the second order Green's function and thus, if G_{11} introduced in Eq. (3.23) is expressed, in matrix form, by

$$G_{11}(1; 2) = \mathbf{G}_q^*(1)\mathbf{G}_q(2)Z_0, \tag{4.15}$$

the multiplication of both sides of Eq. (4.15) to the left with $L^*(1) - \mathbf{q}^*(1)$ and the subsequent use of Eq. (3.16) on the right-hand side, yields

$$[L^*(1) - \mathbf{q}(1)]G_{11}(1; 2) = \delta(1)\mathbf{G}_q(2)Z_0 = \delta(1)G(2), \tag{4.16}$$

where $\delta(1)$ denotes the unit matrix with respect to the coordinates x_1, having the number 1, and similarly $G(2)$ the matrix with respect to x_2.

To evaluate the second term on the left-hand side of Eq. (4.16), we can employ the same procedure as used in Eq. (4.3) and hence

$$q(1, \delta/\delta c)G_{11}(1; 2) = \langle q_\alpha(1)\mathbf{G}_{q+\alpha}^*(1)\mathbf{G}_{q+\alpha}(2)\rangle_\alpha Z_0. \tag{4.17}$$

Here, by formulas (4.5) and (4.6),

$$q_\alpha(1)\mathbf{G}_{q+\alpha}^*(1) = \mathbf{T}_\alpha^{q*}(1)\mathbf{G}_q^*(1), \tag{4.18}$$

$$\mathbf{G}_{q+\alpha}(2) = [1 + \mathbf{G}_q(2)\mathbf{T}_\alpha^q(2)]\mathbf{G}_q(2),$$

which, upon substitution into the right-hand side of Eq. (4.17), yields

$$\langle \mathbf{T}_\alpha^{q*}(1)[1 + \mathbf{G}_q(2)\mathbf{T}_\alpha^q(2)]\rangle_\alpha \mathbf{G}_q^*(1)\mathbf{G}_q(2)Z_0 = [\langle \mathbf{T}_\alpha^{q*}(1)\rangle_\alpha + \mathbf{G}_q(2)\langle \mathbf{T}_\alpha^{q*}(1)\mathbf{T}_\alpha^q(2)\rangle_\alpha]G_{11}(1; 2), \tag{4.19}$$

where use has been made of the commutability of the matrices having different numbers of coordinates.

Here, the consideration similar to that taken in deriving Eq. (4.12) can be applied to the right-hand side of Eq. (4.19), to obtain the approximate expression

$$q(1, \delta/\delta c)G_{11}(1; 2) = [\langle T_\alpha^{M*}(1)\rangle_\alpha + G_M(2)\langle T_\alpha^{M*}(1)T_\alpha^M(2)\rangle_\alpha]G_{11}(1; 2), \tag{4.20}$$

where the operator $\mathbf{q}$ has simply been replaced by the definite matrix M on the right-hand side.

Thus, in terms of the notation

$$M_{11}(1; 2) = \langle T_\alpha^{M*}(1)T_\alpha^M(2)\rangle_\alpha = n\int d\mathbf{a}\, \langle T_\alpha^{M*}(1)T_\alpha^M(2)\rangle', \tag{4.21}$$

Eq. (4.16) becomes, on using Eq. (4.12) after putting $c(x) = 0$,

$$[L(1) - M^*(1) - G(2)M_{11}(1; 2)]G_{11}(1; 2) = \delta(1)G(2), \tag{4.22}$$

which, with the aid of Eq. (4.14), can be given also in a symmetrical form as

$$G_{11}(1; 2) = G^*(1)G(2)[1 + M_{11}(1; 2)G_{11}(1; 2)], \tag{4.23}$$

being given in a form of the Bethe–Salpeter equation.

In the same way, we obtain the equations of the other Green's functions of the second order as

$$G_{02}(1, 2) = G(1)G(2)[1 + M_{02}(1, 2)G_{02}(1, 2)], \tag{4.24}$$

$$G_{20}(1, 2) = G^*(1)G^*(2)[1 + M_{20}(1, 2)G_{20}(1, 2)],$$

where

$$M_{02}(1, 2) = M_{20}^*(1, 2) = \langle T_\alpha^M(1)T_\alpha^M(2)\rangle_\alpha. \tag{4.25}$$

The Green's functions of various orders hold the translational invariance after setting $c(x) = 0$.

5. GENERAL THEORY OF THE MUTUAL COHERENCE FUNCTION OF A WAVE

In Sec. 4, the equations obeyed by the first and second order statistical Green's functions have been found to be given completely in terms of the matrices M, M_{11}, and $M_{02} = M_{20}^*$, where M is defined by Eq. (4.11) and is approxi-

mated, based on the operator method, by Eq. (4.12) while the others are approximated by Eqs. (4.21) and (4.25). In this section, the general theory is extended, independently of the previous operator methods, to see how these matrices are strictly defined in the unperturbative sense, and also the relation inherent between the matrices.

According to the definition (4.11), the matrix M is defined, in view of formula (2.9), by

$$\langle q(x)G_q(x|x')\rangle = \int dx''\, M(x|x'')\langle G_q(x''|x')\rangle \tag{5.1}$$

in terms of the ordinary Green's function $G_q(x|x')$, satisfying Eq. (3.16) with $\mathbf{q}$ replaced by q, and therefore, in matrix form, by

$$\langle qG_q\rangle = M\langle G_q\rangle, \quad \langle qG_q^*\rangle = M^*\langle G_q^*\rangle, \tag{5.2}$$

the latter being the complex conjugate of the former. Hence, in terms of the new quantities

$$\Delta q = q - M, \quad \Delta q^* = q - M^*, \tag{5.3}$$

the equation for G_q is rewritten by

$$(L - M - \Delta q)G_q = 1, \quad \langle \Delta qG_q\rangle = 0, \tag{5.4}$$

whose solution can be given in the form

$$G_q = G + \Delta G_q, \quad G = G_M = (L - M)^{-1}, \tag{5.5}$$

$$\Delta G_q = G\Delta qG_q, \quad \langle \Delta G_q\rangle = 0, \tag{5.6}$$

where G and ΔG_q give the coherent and incoherent parts of the Green's function, respectively.

Here, on employing the expression (5.5) with (5.6) for $G_q(2)$ and the complex conjugate expression for $G_q^*(1)$, we immediately find the equation obeyed by $G_{11}(1;2) = \langle G_q^*(1)G_q(2)\rangle$ strictly in the form (4.23), with $M_{11}(1;2)$ redefined by

$$\langle \Delta q^*(1)\Delta q(2)G_q^*(1)G_q(2)\rangle = M_{11}(1;2)\langle G_q^*(1)G_q(2)\rangle, \tag{5.7}$$

and hence also

$$\langle \psi^*(1)\psi(2)\rangle = \langle \psi^*(1)\rangle\langle \psi(2)\rangle + G^*(1)G(2) \times M_{11}(1;2)\langle \psi^*(1)\psi(2)\rangle. \tag{5.8}$$

Here, the matrices M and $M_{11}(1;2)$ are not entirely independent and, to see this relation, we observe, on using the expression (5.5) for $G_q(2)$ together with Eqs. (5.6) and (5.7), that

$$\langle \Delta q^*(1)G_q^*(1)G_q(2)\rangle = \langle \Delta q^*(1)G_q^*(1)\Delta G_q(2)\rangle = G(2)M_{11}(1;2)\langle G_q^*(1)G_q(2)\rangle, \tag{5.9}$$

or, on using Eq. (5.3) in the left-hand side,

$$\langle q(1)G_q^*(1)G_q(2)\rangle = [M^*(1) + G(2)M_{11}(1;2)]G_{11}(1;2), \tag{5.10}$$

and, in the same way, that

$$\langle q(2)G_q^*(1)G_q(2)\rangle = [M(2) + G^*(1)M_{11}(1;2)]G_{11}(1;2). \tag{5.11}$$

Thus, on letting the coordinates x_1 of (1) and x_2 of (2) coincide in Eqs. (5.10) and (5.11), we find the relation

$$\{M^*(1) - M(2) - [G^*(1) - G(2)]M_{11}(1;2)\}|_{x_1 = x_2} = 0, \tag{5.12}$$

in order that the two expressions become identical.

Here, it is straightforward to show that the above relation guarantees the equation of continuity (3.11), as may be shown first by exhibiting Eq. (5.8) in two ways, one being

$$[L(1) - M^*(1)]\langle \psi^*(1)\psi(2)\rangle = j^*(1)\langle \psi(2)\rangle + G(2)M_{11}(1;2)\langle \psi^*(1)\psi(2)\rangle,$$

and the other the complex conjugate equation with the coordinates (1) and (2) interchanged, and then by deriving their difference with the aid of the relation (5.12). The relation (5.12) gives the optical condition in the sense that the absorbed waves due to the imaginary part of M are perfectly compensated by the same amount of the scattered waves due to the term of M_{11}.

6. COHERENT POTENTIAL EQUATIONS FOR M AND M_{11}

The basic matrices M and M_{11} in the equations of the first and second order Green's functions, are explicitly defined according to Eqs. (5.2) and (5.7) or (5.10), while, independently of these definitions, they have been found to be given approximately by Eqs. (4.12) and (4.21) in the case of random particles. Here, it will be noticed that the latter approximate expressions can be obtained on the more general basis according to the former definitions, by utilizing the coherent potential equations as described in the following.

The Green's function G_q, obtained as the solution of Eq. (5.4), can be exhibited in terms of the scattering matrix $T_{\Delta q}^M$ for Δq, defined by Eq. (4.6) and (4.7), by

$$\Delta qG_q = T_{\Delta q}^M G_M, \quad T_{\Delta q}^M = (1 - \Delta qG_M)^{-1}\Delta q, \tag{6.1}$$

and hence Eq. (5.6) is rewritten as

$$\Delta G_q = GT_{\Delta q}^M G, \quad G = G_M, \tag{6.2}$$

with the condition

$$\langle T_{\Delta q}^M\rangle = 0. \tag{6.3}$$

Here, $G = G_M$ is defined by Eq. (4.14) in terms of the unknown M, and therefore the condition (6.3) provides us with an equation for determining the matrix M.

To find the corresponding equation for determining M_{11} we employ, on both sides of Eq. (5.7), the expression (6.1) with the condition (6.3) and hence

$$\langle T_{\Delta q}^{M*}(1)T_{\Delta q}^M(2)\rangle G^*(1)G(2) = M_{11}(1;2)[1 + G^*(1)G(2)\langle T_{\Delta q}^{M*}(1)T_{\Delta q}^M(2)\rangle] \times G^*(1)G(2), \tag{6.4}$$

which gives the explicit expression of M_{11}, given by

$$M_{11}(1;2) = \langle T_{\Delta q}^{M*}(1)T_{\Delta q}^M(2)\rangle \times [1 + G^*(1)G(2)\langle T_{\Delta q}^{M*}(1)T_{\Delta q}^M(2)\rangle]^{-1}. \tag{6.5}$$

Thus, the matrices M and M_{11} could be found according to Eqs. (6.3) and (6.5), by utilizing the approximation similar to that used in Sec. 4; this sort of approximation has been called the coherent potential approximation in solid physics and been successfully used to treat the impurity problems.[13]

A. Simple example: Weak-scattering limit

From Eq. (6.2), the scattering matrix $T_{\Delta q}^M$ can be given

by

$$T^{M}_{\Delta q} = \Delta q(1 + G_M T^{M}_{\Delta q}), \tag{6.6}$$

where the second term in the parenthesis means the effect of Δq itself to give the effective incident wave on Δq. Hence, averaging both sides of Eq. (6.6) and using the condition (6.3),

$$M = \langle \Delta q G_M T^{M}_{\Delta q} \rangle, \quad \langle q \rangle = 0. \tag{6.7}$$

Therefore, in the weak-scattering limit where q is small enough to retain only the first nonvanishing term on the right-hand side of Eq. (6.7), we obtain

$$M \sim \langle qGq \rangle. \tag{6.8}$$

To obtain the matrix $M_{11}(1; 2)$ according to Eq. (6.5), we again employ the expression (6.6) for both $T^{M*}_{\Delta q}(1)$ and $T^{M}_{\Delta q}(2)$ to obtain

$$\langle T^{M*}_{\Delta q}(1) T^{M}_{\Delta q}(2) \rangle = \langle \Delta q^*(1) \Delta q(2) \times [1 + G^*_M(1) T^{M*}_{\Delta q}(1)][1 + G_M(2) T^{M}_{\Delta q}(2)] \rangle, \tag{6.9}$$

which becomes, on neglecting the correlation between Δq and the effective incident wave,

$$\langle T^{M*}_{\Delta q}(1) T^{M}_{\Delta q}(2) \rangle = \langle \Delta q^*(1) \Delta q(2) \rangle \times [1 + G^*_M(1) G_M(2) \langle T^{M*}_{\Delta q}(1) T^{M}_{\Delta q}(2) \rangle], \tag{6.10}$$

in virtue of the condition (6.3).

Thus, according to the definition (6.5), we find the simple expression

$$M_{11}(1; 2) = \langle \Delta q^*(1) \Delta q(2) \rangle \sim \langle q(1) q(2) \rangle, \tag{6.11}$$

to the lowest order of q, independently of the statistics obeyed by $q(x)$. Here, it will be noted that, on the right-hand side of Eq. (6.10), the last factor [ ] cannot be replaced by unity since the second term in this factor means that part of the incident waves on $q(1)$ and $q(2)$, contributed from the incoherent part of waves scattered within the range of the coherence distance of wave,i.e., the range in which $G^*_M(1)$ and $G_M(2)$ are appreciable. The expression (6.11) has been known as the ladder approximation.

B. Coherent potential approximation for M and M_{11} in case of random particles

It can be shown that the matrices M and M_{11} given by the coherent potential equations (6.3) and (6.5) are equivalent to those given by Eqs. (4.12) and(4.21) according to the effective medium method, as far as the incoherency is assumed between the incident and scattered waves by the random medium, as it is also the case of Sec. 4. The proof is given in Appendix A.

7. EQUATIONS FOR HIGHER-ORDER COHERENCE FUNCTIONS OF A WAVE IN MEDIA OF RANDOM PARTICLES

The equations obeyed by the coherence functions of wave higher than the second order, can also be derived following the procedure similar to that used in Sec. 4 for the first and second order functions. However, it turns out to be more simple to first find the equations obeyed by the characteristic funcitonal of wave, $Z[\bar{j}^*, \bar{j}]$, and then, upon substitution of the moment expansion (3.14), to derive the coherence equations of various orders.

We begin with the basic equation (3.8) and hence, on using the expressions (2.29) for $\mathbf{q}(x)$ and (3.20) for $Z[\bar{j}^*, \bar{j}]$, we obtain

$$[\mathbf{q}(x) - c(x)]\psi(x) Z[\bar{j}^*, \bar{j}] = q(x, \delta/\delta c) \int dx'\, \mathbf{G}_q(x|x') \bar{j}(x') Z[\bar{j}^*, \bar{j}], \tag{7.1}$$

which becomes, in the same manner as in Eq. (4.3),

$$\left\langle \int dx'\, q_\alpha(x) \mathbf{G}_{q+\alpha}(x|x') \bar{j}(x') \exp\left[\int dx_1\, dx_2 \{ \bar{j}(x_1) \mathbf{G}_{q+\alpha}(x_1|x_2) \bar{j}(x_2) + \bar{j}^*(x_1) \mathbf{G}^*_{q+\alpha}(x_1|x_2) \bar{j}^*(x_2) \} \right] \right\rangle_\alpha Z_0. \tag{7.2}$$

Here, in terms of the notation (4.9),

$$q_\alpha(x) \mathbf{G}_{q+\alpha}(x|x') = \int dx'' \mathbf{T}^q_\alpha(x|x'') \mathbf{G}_q(x''|x'), \tag{7.3}$$

$$\mathbf{G}_{q+\alpha}(x_1|x_2) = \mathbf{G}_q(x_1|x_2) + \int dx'\, dx''\, \mathbf{G}_q(x_1|x') \mathbf{T}^q_\alpha(x'|x'') \mathbf{G}_q(x''|x_2), \tag{7.4}$$

by virtue of the formulas (4.5) and (4.6). Hence, Eq. (7.2) further becomes

$$\left\langle \int dx'\, dx'' \mathbf{T}^q_\alpha(x|x') \mathbf{G}_q(x'|x'') \bar{j}(x'') \exp\left[\int dx_1\, dx_2\, dx_3\, dx_4 \left\{ \bar{j}(x_1) \mathbf{G}_q(x_1|x_2) \mathbf{T}^q_\alpha(x_2|x_3) \mathbf{G}_q(x_3|x_4) \bar{j}(x_4) + \bar{j}^*(x_1) \mathbf{G}^*_q(x_1|x_2) \times \mathbf{T}^{q*}_\alpha(x_2|x_3) \mathbf{G}^*_q(x_3|x_4) \bar{j}^*(x_4) \right\} \right] \right\rangle_\alpha Z[\bar{j}^*, \bar{j}]. \tag{7.5}$$

Here, from Eqs. (3.6) and (3.20), it follows that

$$[\psi(x)]^n Z[\bar{j}^*, \bar{j}] = \left[\int dx' \mathbf{G}_q(x|x') \bar{j}(x') \right]^n Z[\bar{j}^*, \bar{j}], \quad n = 1, 2, 3, \cdots, \tag{7.6}$$

and hence Eq. (7.5) can be exhibited by

$$\mathcal{N} \left\langle \exp\left[\int dx_1\, dx_2\, dx_3 \left\{ \bar{j}(x_1) \mathbf{G}_q(x_1|x_2) \mathbf{T}^q_\alpha(x_2|x_3) \psi(x_3) + \bar{j}^*(x_1) \mathbf{G}^*_q(x_1|x_2) \mathbf{T}^{q*}_\alpha(x_2|x_3) \psi^*(x_3) \right\} \right] \int dx'\, \mathbf{T}^q_\alpha(x|x') \right\rangle_\alpha \times \psi(x') Z[\bar{j}^*, \bar{j}]. \tag{7.7}$$

Here, the symbol $\mathcal{N}$ stands for the ordering of the referred function of $\bar{j}, \bar{j}^*$, ψ, and ψ^* in such a manner that, in its power expansion, $\bar{j}$ and $\bar{j}^*$ are always to the left of the operators ψ and ψ^*.

Thus, on replacing the operator $\mathbf{q}$ by the definite matrix M, as has been done in the previous equations (4.12) and (4.20) or (4.21), we finally obtain Eq. (7.1) with (7.7), expressed in the form

$$\mathbf{q}(x)\psi(x)Z[\bar{j}^*,\bar{j}]\,|_{c=0} = \int dx'\,\kappa_{01}(x|x')\psi(x')Z[\bar{j}^*,\bar{j}]\Big|_{c=0}, \tag{7.8}$$

where

$$\kappa_{01}(x|x') = \mathcal{N}\left\langle \exp\left[\int dx_1 dx_2 dx_3 \left\{\bar{j}(x_1)G_M(x_1|x_2)T_\alpha^M(x_2|x_3)\psi(x_3) + \bar{j}^*(x_1)G_M^*(x_1|x_2)T_\alpha^{M*}(x_2|x_3)\psi^*(x_3)\right\}\right] T_\alpha^M(x|x')\right\rangle_\alpha . \tag{7.9}$$

Thus, when $c(x) = 0$, Eq. (3.8) is exhibited by

$$\left[L(i\partial/\partial x)\psi(x) - \int dx'\kappa_{01}(x|x')\psi(x') - j(x)\right] Z[\bar{j}^*,\bar{j}] = 0, \tag{7.10}$$

and, in the same way, Eq. (3.9) by

$$\left[L^*(-i\partial/\partial x)\psi^*(x) - \int dx'\kappa_{10}(x|x')\psi^*(x') - j^*(x)\right] Z[\bar{j}^*,\bar{j}] = 0, \tag{7.11}$$

where $\kappa_{10}(x|x')$ is the same as $\kappa_{01}(x|x')$ with the factor $T_\alpha^M(x|x')$ replaced by $T_\alpha^{M*}(x|x')$.

Equations (7.10) and (7.11) are the basic equations obeyed by the characteristic functional of wave. Here, one of the methods of solving those equations is obviously to substitute the moment expansion (3.14) and derive the equations for $m_{\mu\nu}$ of various orders, but the equations rapidly become complicated with the increase of their orders in the present case, although they can be given in a compact form in the case when the medium can be assumed to follow the Gaussian statistics (Appendix B). Equation (4.13) for the first order Green's function is derived from Eq. (7.10) simply by putting $\bar{j}^* = \bar{j} = 0$.

As far as the irradiance and its moments are concerned, only the symmetrical moments $m_{\nu\nu}$ with the same order for ψ^* and ψ become necessary, while Eqs. (7.10) and (7.11) are not given in a form quite convenient for their derivation and therefore are expected to be unified to an equation symmetrical with respect to ψ^* and ψ. This process will be facilitated by introducing an operator similar to κ_{01}, defined by

$$\kappa = \mathcal{N}\left\langle \exp\left[\int dx_1\,dx_2\,dx_3\,\left\{\bar{j}(x_1)G_M(x_1|x_2)T_\alpha^M(x_2|x_3)\psi(x_3) + \bar{j}^*(x_1)G_M^*(x_1|x_2)T_\alpha^{M*}(x_2|x_3)\psi^*(x_3)\right\}\right]\right\rangle_\alpha , \tag{7.12}$$

and also the associated operators, defined, in matrix form, by

$$\kappa_{mn}(1, 2, \ldots, m; 1, 2, \ldots, n) = \mathcal{N}\langle T_\alpha^{M*}(1)T_\alpha^{M*}(2)\cdots T_\alpha^{M*}(m) \times T_\alpha^M(1)T_\alpha^M(2)\cdots T_\alpha^M(n)\exp[\quad]\rangle_\alpha, \tag{7.13}$$

where exp[ ] is the same as for κ in Eq. (7.12). Here, with the aid of the commutation relation (3.7), it is straightforward to obtain the following commutation relations:

$$[\psi(x), \kappa_{mn}] = \int dx_1\,dx_2\,G(x|x_1)\kappa_{m,n+1}(x_1|x_2)\psi(x_2), \tag{7.14}$$

$$[\psi^*(x), \kappa_{mn}] = \int dx_1\,dx_2\,G^*(x|x_1)\kappa_{m+1,n}(x_1|x_2)\psi^*(x_2),$$

where all the unconcerned coordinates have been suppressed.

Here, in matrix form, Eq. (7.14) is expressed by

$$[\psi(2), \kappa_{mn}] = G(2)\kappa_{m,n+1}(2)\psi(2), \tag{7.15}$$

$$[\psi^*(1), \kappa_{mn}] = G^*(1)\kappa_{m+1,n}(1)\psi^*(1),$$

and Eqs. (7.10) and (7.11) by

$$\{[L(2) - \kappa_{01}(2)]\psi(2) - j(2)\}Z = 0, \tag{7.16}$$

$$\{[L^*(1) - \kappa_{10}(1)]\psi^*(1) - j^*(1)\}Z = 0, \tag{7.17}$$

which still keep the original form of the wave equation (1.1) in terms of the operators.

Here, on multiplying Eq. (7.17) to the left with $\psi(2)$ and subsequently using the commutation relation (7.15), we find

$$\{[L^*(1) - \kappa_{10}(1) - G(2)\kappa_{11}(1; 2)]\psi^*(1)\psi(2) - j^*(1)\psi(2)\}Z = 0, \tag{7.18}$$

and, in the same way, from Eq. (7.16)

$$\{[L(2) - \kappa_{01}(2) - G^*(1)\kappa_{11}(1; 2)]\psi^*(1)\psi(2) - j(2)\psi^*(1)\}Z = 0. \tag{7.19}$$

Here, Eq. (4.22) for $G_{11}(1; 2)$ is directly derived from Eq. (7.18) by putting $\bar{j}^* = \bar{j} = 0$.

Thus, the subtraction of Eq. (7.19) from Eq. (7.18) yields an equation of the form

$$\{[L(1; 2) + \mathbf{V}(1; 2)]\psi^*(1)\psi(2) + \mathbf{J}(1; 2)\}Z = 0, \tag{7.20}$$

where

$$L(1; 2) = (i/2)[L^*(1) - L(2)], \tag{7.21}$$

$$\mathbf{V}(1; 2) = (i/2)\{\kappa_{01}(2) - \kappa_{10}(1) + [G^*(1) - G(2)] \times \kappa_{11}(1; 2)\}, \tag{7.22}$$

$$\mathbf{J}(1; 2) = (i/2)[j(2)\psi^*(1) - j^*(1)\psi(2)]. \tag{7.23}$$

Here it is noticed that, when the coordinates x_1 of (1)

and x_2 of (2) coincide in Eq. (7.20), the equation should be reduced to Eq. (3.13), representing the equation of continuity of wave. This implies that, if $\mathbf{V}(x_1; x_2|x_1'; x_2')$ designates the matrix element of $\mathbf{V}(1; 2)$, the relation

$$\mathbf{V}(x_1; x_2|x_1'; x_2')|_{x_1 = x_2} = 0, \tag{7.24}$$

should hold independently of x_1' and x_2'.

Here, according to Eqs. (7.22) and (7.15), $\mathbf{V}(1; 2)$ can be derived from κ, defined by Eq. (7.12), which is composed of the single-particle scattering matrix T_α^M, the first order Green's function G_M, and their complex conjugates, and, in Appendix C, the proof of the relation (7.24) is given in terms of the relation existing among those quantities, i.e., the optical condition of T_α^M, $M^* \neq M$, in the generalized sense, equivalent to Eq. (5.12). This becomes more explicit by putting $\bar{j}^* = \bar{j} = 0$, in which case, $\kappa_{10}(1) = M^*(1)$, $\kappa_{01}(2) = M(2)$, and $\kappa_{11}(1; 2) = M_{11}(1; 2)$ by Eqs. (4.12) and (4.21), reducing the condition (7.24) with Eq. (7.22) to the condition (5.12) in terms of those in the effective medium approximation introduced in Sec. 4

To derive the higher order moment equations for $m_{\nu\nu'}$, $\nu \geqslant 2$, we first need to evaluate the commutator of the form

$$[\psi^*(3)\psi(4), \mathbf{V}(1; 2)] = \mathbf{V}'(1; 2|3; 4)\psi^*(3)\psi(4). \tag{7.25}$$

Here, from the expression (7.22) for $\mathbf{V}(1; 2)$, it is found to be

$$\begin{aligned}\mathbf{V}'(1; 2|3; 4) &= G^*(3)\mathbf{V}_{10}(1; 2|3) + G(4)\mathbf{V}_{01}(1; 2|4) \\ &\quad + G^*(3)G(4)\mathbf{V}_{11}(1; 2|3; 4) \\ &\neq \mathbf{V}'(3; 4|1; 2),\end{aligned} \tag{7.26}$$

with the new operator $\mathbf{V}_{mn}$, defined, in terms of the notation κ_{mn} in Eq. (7.13), by

$$\begin{aligned}\mathbf{V}_{mn}(1; 2|3, \cdots; 4, \cdots) &= \tfrac{1}{2}i[\kappa_{m,n+1}(2) - \kappa_{m+1,n}(1) \\ &\quad + \{G^*(1) - G(2)\}\kappa_{m+1,n+1}(1; 2)],\end{aligned} \tag{7.27}$$

where the total numbers of the coordinates 3, ..., and 4, ..., are m for the complex conjugate wave functions and n for the original wave functions, and these coordinates have been suppressed on the right-hand side of Eq. (7.27).

Here, by virtue of the condition (7.24) for $\mathbf{V}(1; 2)$, the conditions exhibited by

$$\mathbf{V}_{mn}(1; 2|3, \ldots; 4, \ldots)|_{x_1 = x_2} = 0, \tag{7.28}$$

$$\mathbf{V}'(1; 2|3; 4)|_{x_1 = x_2} = 0, \tag{7.29}$$

also hold in the same sense.

Thus, on multiplying Eq. (7.20) to the left by $\psi^*(3)\psi(4)$ and using the commutation relation (7.25), we find

$$\begin{aligned}&\{[L(1; 2) + \mathbf{V}(1;2) + \mathbf{V}'(1; 2|3; 4)]\psi^*(1)\psi^*(3)\psi(2)\psi(4) \\ &\quad + \mathbf{J}(1; 2)\psi^*(3)\psi(4)\}Z = 0,\end{aligned} \tag{7.30}$$

while, from Eq. (3.14), the moments of the wave functions are given by

$$\begin{aligned}&m_{\mu\nu}(1, 3, \ldots, 2\mu - 1; 2, 4, \ldots, 2\nu) \\ &\quad = \psi^*(1)\psi^*(3)\cdots\psi^*(2\mu - 1)\psi(2)\psi(4)\cdots\psi(2\nu) \\ &\qquad \times Z[\bar{j}^*, \bar{j}]|_{\bar{j} = \bar{j}^* = 0}.\end{aligned} \tag{7.31}$$

To investigate the pronounced features of Eq. (7.20) in the case $\bar{j} = \bar{j}^* = 0$, it is convenient to first introduce the relative coordinates $r_1 = (\mathbf{r}_1, t_1)$ and $\rho_1 = (\boldsymbol{\rho}_1, \tau_1)$, defined by

$$r_1 = x_2 - x_1, \quad \rho_1 = \tfrac{1}{2}(x_2 + x_1), \tag{7.32}$$

yielding

$$L(1; 2) = i\left[\frac{\partial}{\partial \mathbf{r}_1}\cdot\frac{\partial}{\partial \boldsymbol{\rho}_1} - \frac{1}{c^2}\frac{\partial}{\partial t_1}\frac{\partial}{\partial \tau_1}\right], \tag{7.33}$$

and also the matrix elements $V(x_1; x_2|x_1'; x_2')$ of the matrix $V(1; 2) \equiv \mathbf{V}(1; 2)$ for $\bar{j} = \bar{j}^* = 0$, of the form $V(r_1|\rho_1 - \rho_1'|r_1')$, as is required by the translational invariance of $V(1; 2)$; further, the condition (7.24) is exhibited by

$$V(r_1|\rho_1 - \rho_1'|r_1')|_{r_1 = 0} = 0, \quad \bar{j} = \bar{j}^* = 0. \tag{7.34}$$

Here, when the space-time change of the wave functions are mostly due to their phases, with a sufficiently slow change of their amplitudes, then, it follows that the change of $\langle\psi^*(x_1)\psi(x_2)\rangle$ with respect to the coordinates ρ_1 is negligible, as compared with the change with respect to r_1, and therefore also that $V(1; 2)$ [to be substituted for $\mathbf{V}(1; 2)$ in Eq. (7.20)] can be approximated by a new matrix, defined by the matrix elements

$$V(r_1|r_1') = \int_{-\infty}^{\infty} d\rho_1'\ V(r_1|\rho_1 - \rho_1'|r_1'), \tag{7.35}$$

being a matrix with respect to only the coordinates r_1 and r_1'. In view of that, in the present case of random particles, $V(r_1|\rho_1 - \rho_1'|r_1')$ is a very short range function, different from zero only within the range where $|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_1'|$ and $|\tau_1 - \tau_1'|$ are of the order of the particle diameters and the propagation time of wave through the particles, respectively, or smaller. Thus, Eq. (7.20) becomes expressed, when $\bar{j} = \bar{j}^* = 0$, by

$$\begin{aligned}&i\left[\frac{\partial}{\partial \mathbf{r}_1}\cdot\frac{\partial}{\partial \boldsymbol{\rho}_1} - \frac{1}{c^2}\frac{\partial}{\partial t_1}\frac{\partial}{\partial \tau_1}\right]m_{11}(r_1, \rho_1) \\ &\quad + \int dr_1'\ V(r_1|r_1')m_{11}(r_1', \rho_1) + J(r_1, \rho_1) = 0.\end{aligned} \tag{7.36}$$

Here, in the particular case where $V(r_1|r_1')$ depends on t_1 and t_1' only through the difference $t_1 - t_1'$, then, the frequency of the wave function with the periodic time factor $e^{i\omega t}$ is not changed by the scattering and, in terms of the notations

$$\begin{aligned}T_1 &= i\frac{\partial}{\partial \mathbf{r}_1}\cdot\frac{\partial}{\partial \boldsymbol{\rho}_1}, \\ V_1(\mathbf{r}_1|\mathbf{r}_1') &= \int dt_1'\ V(r_1|r_1')\exp[i\omega(t_1' - t_1)],\end{aligned} \tag{7.37}$$

Eq. (7.36) is further simplified, on replacing $-i\partial/\partial t_1 \to \omega$, to the form

$$\left[\frac{\omega}{c^2}\frac{\partial}{\partial \tau_1} + T_1 + V_1\right]m_{11}(\tau_1) + J(\tau_1) = 0, \tag{7.38}$$

where the matrix V_1 is defined by the matrix elements in Eq. (7.37), being a matrix with respect to only the spatial coordinates $\mathbf{r}_1$ and $\mathbf{r}_1'$.

Equation (7.30) also can be simplified by the same procedure, on first introducing the additional coordinates

$$\begin{aligned}&r_2 = x_4 - x_3, \quad \rho_2 = \tfrac{1}{2}(x_4 + x_3), \\ &\rho_{12} = \rho_1 - \rho_2, \quad \rho = \tfrac{1}{2}(\rho_1 + \rho_2),\end{aligned} \tag{7.39}$$

with the time components t_2, τ_2, τ_{12}, and τ, respectively,

which are all (except ρ) translationally invariant and therefore permit the elements of the matrix V', when $\bar{j} = \bar{j}^* = 0$, to be given in the form

$$V'(r_1, r_2, \rho_{12}|\rho - \rho'|r_1', r_2', \rho_{12}');$$

then, this is replaced, on integrating with respect to ρ' as in Eq. (7.35), say, by $V'(r_1, r_2, \rho_{12}|r_1', r_2', \rho_{12}')$ and, in case of the same situation as in Eq. (7.37), further by a matrix with respect to only the spatial coordinates, given by the matrix elements

$$V'_{12}(\mathbf{r}_1, \mathbf{r}_2, \boldsymbol{\rho}_{12}|\mathbf{r}_1', \mathbf{r}_2', \boldsymbol{\rho}_{12}') = \int d\tau_{12}' \int dt_1'\, dt_2' \times \exp[i\omega(t_1' + t_2' - t_1 - t_2)] V'(r_1, r_2, \rho_{12}|r_1', r_2', \rho_{12}'), \tag{7.40}$$

which, in view of the condition (7.29), tends to zero as $\mathbf{r}_1 \to 0$ (although not for $\mathbf{r}_2 \to 0$). Thus, from Eq. (7.30), we finally find an equation, corresponding to Eq. (3.38), in the form

$$\left[\frac{\omega}{c^2}\frac{\partial}{\partial \tau_1} + T_1 + V_1 + V_{12}'\right] m_{22}(\tau_1, \tau_2) + J_{12}'(\tau_1, \tau_2) = 0. \tag{7.41}$$

Here, the matrix elements of V_{12}' are given by Eq. (7.40) with the condition $V_1 = V_{12}'(\neq V_{21}') = 0$ for $\mathbf{r}_1 = 0$, and the coordinates $\mathbf{r}_1, \mathbf{r}_2, \boldsymbol{\rho}_1, \boldsymbol{\rho}_2$ have been suppressed; J_{12}' provides the source term.

Also with respect to the time coordinate τ_2, we obtain the equation similar to Eq. (7.41) and therefore, letting $\tau_1 = \tau_2 = \tau$, the equation with respect to τ is found (since the two equations are linear in $\partial/\partial\tau_1$ and $\partial/\partial\tau_2$, respectively), to be

$$\left[\frac{\omega}{c^2}\frac{\partial}{\partial \tau} + T_1 + T_2 + V_1 + V_2 + V_{12}\right] m_{22}(\tau) + J_{12}(\tau) = 0, \tag{7.42}$$

with

$$V_{12} = V_{12}' + V_{21}' = V_{21}, \quad m_{22}(\tau) = m_{22}(\tau_1, \tau_2)\big|_{\tau_1 = \tau_2 = \tau}.$$

In order to derive the next order moment equation from Eq. (7.30), we need to evaluate the commutator of

$\mathbf{V}'(1; 2|3; 4)$ and $\psi^*(5)\psi(6)$, which gives rise to a higher order correction of V_{12}' due to the interaction with $\psi^*(5)\psi(6)$ and involves the additional factors $T_\alpha^{M*}(5)$ and $T_\alpha^M(6)$ in the $\langle\cdots\rangle_\alpha$ average, besides those of V'. Therefore, when the contribution from this commutator is neglected, all the higher order moment equations are systematically obtained in the form

$$\left[\frac{\omega}{c^2}\frac{\partial}{\partial \tau} + \sum_{j=1}^{\nu}(T_j + V_j) + \sum_{i>j=1}^{\nu} V_{ij}\right] m_{\nu\nu}(\tau) = 0, \tag{7.43}$$

over the region of vanishing wave source. Here, all the spatial coordinates $\mathbf{r}_j, \boldsymbol{\rho}_j, j = 1, 2, \ldots, \nu$, have been suppressed, and T_j, V_j and V_{ij} are the same as those in Eq. (7.42), being functions of a very short range of the order of the particle diameters. It is noted that, in order that the symmetries of Eq. (7.43) with respect to the original coordinates x_j of the even numbers $j = 2, 4, \ldots, 2\nu$ and those of the odd numbers 3, 5, ..., $2\nu - 1$, are respectively secured without violating the condition (7.29), all the terms of $\mathbf{V}'(1; 2|3; 4)$ in Eq. (7.26) are inevitably necessary and consequently given to the fourth order of T_α^M and T_α^{M*}.

8. SUMMARY AND DISCUSSION

The random medium $q(x)$ in the wave equation (1.1) can be represented by the operator $\mathfrak{q}(x)$, as given by Eq. (2.14) with (2.13) in terms of the characteristic functional $Z_q[p]$ of the medium, and this representation particularly facilitates obtaining the expectation value of any functional $f[q]$ of $q(x)$ in space and time; the latter is simply obtained according to Eq. (2.8), and the associated relation (2.9) is especially convenient in finding $\langle q(x)f[q]\rangle$ when $\langle f[q]\rangle$ is given. Here, the operator $\mathfrak{q}(x)$ at different points in space and time are mutually commutable and therefore can be treated in entirely the same way as the ordinary functions. The explicit expression of $\mathfrak{q}(x)$ is given by Eq. (2.18) in the weak-scattering limit and by Eq. (2.29) in the medium of random particles, while, when the medium is composed of several independent components, $\mathfrak{q}(x)$ is obtained according to Eq. (2.25).

On the other hand, when the random medium is prescribed by the characteristic functional $Z_q[p]$, the equations obeyed by the characteristic functional of wave, $Z[\bar{j}^*, \bar{j}]$, are given by Eqs. (3.8) and (3.9) which are exhibited in terms of the medium operator $\mathfrak{q}(x)$ and also the wave operators $\psi(x)$ and $\psi^*(x)$, defined by Eq. (3.6). Here, these equations preserve the forms of the original wave equations (1.1) with the replacement of ψ, ψ^*, and q by ψ, ψ^*, and $\mathfrak{q}$, respectively. This is a consequence of the more general correspondence principle (3.10), and the latter could be applied also, e.g., to the equation of continuity (3.11), the equations of conservation for the energy and momentum of a wave, constructed according to the Lagrangian principle, etc. In this connection, it should be noted that the energy and/or momentum of a wave are generally not conserved in media fluctuating in time and/or space, whereas the equation of continuity (3.11) always holds independently of the medium fluctuation.

The equations for $Z[\bar{j}^*, \bar{j}]$ thus obtained contain the undesirable operator $\delta/\delta c(x)$ through $\mathfrak{q}(x)$ and therefore the next task is to introduce the new operators $\kappa_{01}(x)$ and $\kappa_{10}(x)$, as defined by Eq. (7.8), which are free of $\delta/\delta c(x)$ but are functionals of $\bar{j}, \bar{j}^*$, ψ, and ψ^* instead. To this end, the basic assumption has been made that, as generally accepted in the random media, the correlation between the incident wave and the scattered wave is negligible, and this enables κ_{01} to be given by Eq. (7.9) in case of the random particles while by Eq. (B11) in the case of the weak-scattering limit.

Thus, it follows that the resulting equations for $Z[\bar{j}^*, \bar{j}]$ still preserve the forms of the original wave equations with the replacement of the variables by the corresponding operators, as exhibited by Eqs. (7.10) and (7.11), and this correspondence principle facilitates getting the physical insight into the equations, leading, e.g., to Eq. (7.20), which is given in a form symmetrical with respect to the operators ψ and ψ^*, and which tends to the equation of continuity (3.13) in the special situation of when the two coordinates of (1) and (2) coincide; the latter restriction requires the operator $\mathbf{V}(1; 2)$ to satisfy the condition (7.24), as proved strictly in Appendix C, and turns out to be equivalent to the optical condition (5.12) in the generalized sense. When $\bar{j} = \bar{j}^* = 0$, Eq. (7.20) is reduced to Eq. (5.8) for the mutual coherence function of a wave.

Independently of the operator methods, the general theory is extended specifically for the equation satisfied by the mutual coherence function of a wave in Sec. 5, and the basic matrices M and M_{11} are strictly defined according to Eqs. (5.2) and (5.7) in an unperturbative manner. The matrices thus defined satisfy the optical condition (5.12) rigorously, and the resulting equation for the coherence function necessarily has a form of the Bethe–Salpeter equation. In Sec. 6, the coherent potential equations are constructed to evaluate the matrices M and M_{11} according to the definition in Sec. V, and their explicit expressions are obtained, where use has been made of the usual multiple scattering theory for a many-particle system, together with the coherent potential approximation which has been successfully used in solid physics to treat the impurity problems.[13] It turns out that their expressions are precisely the same as those obtained by the effective medium method introduced in Sec. 4.

So far the various equations have been treated on the same footing in space and time, but they could have been exhibited in terms of those in the wave number space, by means of the Fourier transformation for all the functions involved, according to

$$\tilde{f}(k) = \int dx \exp[ik\cdot x] f(x), \quad k = (\mathbf{k}, \omega),$$
$$\mathbf{k} = (k_1, k_2, k_3), \quad k\cdot x = \mathbf{k}\cdot\mathbf{x} - \omega t. \tag{8.1}$$

The only alteration necessary for this case is the replacement of the function $f(x)$ or matrix $m(x_1|x_2)$ by $\tilde{f}(k)$ or $\tilde{m}(k_1|\kappa_2)$, and dx by $dk = (2\pi)^{-4} d\,\mathbf{k} d\omega$ in all the equations, giving rise to convolution integrals in case of space-time diagonal matrices. For example, in the weak-scattering limit, Eq. (B9) with (B5) would be replaced, on using the specific forms $M(x|x') = M(x - x')$ and $G(x|x') = G(x - x')$, by

$$\left\{[L(k) - \tilde{M}(k)]\tilde{\psi}(k) - \tilde{j}(k) - \int dk'\, \tilde{D}(k')\tilde{\mathbf{Q}}(k')\tilde{\psi}(k - k')\right\} Z[\tilde{j}^*, \tilde{j}] = 0, \tag{8.2}$$

with

$$\tilde{\mathbf{Q}}(k) = \int dk' \,[\tilde{j}(-k')\tilde{G}(k')\tilde{\psi}(k' + k) + \tilde{j}^*(-k') \times \tilde{G}^*(k')\tilde{\psi}^*(k' + k)], \tag{8.3}$$

and the commutation relations

$$[\tilde{\psi}(k), \tilde{j}(-k')] = (2\pi)^4 \delta(k - k'),$$
$$[\tilde{\psi}(k), \tilde{\psi}^*(k')] = 0, \quad \text{etc.} \tag{8.4}$$

It is also possible to utilize the wave number representation with respect to the time only, and this is particularly convenient in the case when the medium is dispersive in time while its temporal fluctuation is slow enough to be negligible within the wave period. In this case, the medium can be well represented by the Fourier transform $q(\mathbf{x}, \omega)$ with respect to the time, and, with the replacement of $f(x) \to f(\mathbf{x}, \omega)$ and $dx \to (2\pi)^{-1} d\mathbf{x} d\omega$, various equations preserve their original forms of the equations, described on the same footing in space and time.

In the special case in which (1) $L(i\partial/\partial x)$ in the wave equation is linear with respect to $i\partial/\partial t$, as in the Schrödinger equation, or, when it is time-independent, with respect to the particular component of $i\partial/\partial \mathbf{x}$ in the direction of wave propagation, as in the forward-scattering approximation, and also (2) the coherence distance of wave in time or space is long engough compared with the corresponding correlation distance of the medium, then, the equation satisfied by $Z[\bar{j}^*, \bar{j}]$ can be given in a form of the Fokker–Planck equation,[14] showing that the wave is effectively described by the Markov process.[5,15]

The equation for $Z[\bar{j}^*, \bar{j}]$ may be solved in terms of the moment equations of wave of all orders, in view of the expansion (3.14), but obtaining their solutions for all the orders is practically impossible even in the weak-scattering limit and with the definite frequency of wave, except the special case when the medium structure function can be given in the parabolic form. In the latter case, the exact solutions have been obtained for all the orders, with the resulting irradiance distribution given by the Rice–Nakagami distribution with respect to the logarithm of irradiance.[7,16,17] But, the assumed model of the medium merely gives rise to the wandering of the wave beam *without* any deformation of the wave beam cross-section, and therefore the model's major interst is mathematical rather than physical.[18,19]

On the other hand, as the medium fluctuation becomes sufficiently large, the moments of irradiance tend to be given by asymptotic expressions, and the latter have been investigated as a function of the order of moment by different methods,[20,21,22] based on the Kolmogorov spectrum of turbulence. However, in order that the obtained expression be valid, it turns out that the larger the order becomes, the larger the medium fluctuation becomes; in fact, in comparison with the experimental values so far obtained, the expression is applicable only up to the third order moment, at most,[21] and, ignoring this fact, it leads to the exponential distribution of the irradiance. Experimentally, however, the irradiance distribution observed in the optical propagation through turbulent air, has been known to be very close to the log-normal distribution, and the theoretical basis for this distribution has been found to be the applicability of the cluster approximation to the solutions of the moment equations, particularly when the essential part of the medium is described by the Kolmogorov spectrum of turbulence.[23] This approximation enables us to exhibit the high order moments of irradiance in terms of the lower order moments in an effective way, and the theory shows a very good agreement with the experimental values so far obtained.[24]

The analytical study for the second order moment of irradiance in turbulent air also has been tried to obtain the expression applicable to the entire range of medium fluctuation, particularly in connection with the saturation phenomenon of irradiance scintillation, but seems to have been unsuccessful. So far the numerical method has been used to obtain the result for two-dimensional space[25] and recently the Monte Carlo method for three-dimensional space.[26]

The equation for the mutual coherence function of a wave is practically most important, and satisfies the equa-

tion of a form of the Bethe–Salpeter equation, independently of the statistics obeyed by the medium (Sec. 5). Consequently, the equation is still difficult to solve in its original form, but, to a good approximation, the ordinary transport equation is known to be derived from this *B-S* equation under the condition that the scattering cross-section of medium undergoes a negligibly small change for the change of wave frequency of the order of the coherence frequency of the wave (or the extinction coefficient times wave velocity) and also of the frequency of space-time change of the wave intensity. Here, since this condition is fulfilled in most cases of interest, obtaining the average intensity of wave can be effected by solving the space-time transport equation. It is furthermore known that, in the particular case when the forward-scattering approximation is possible, the equation of the mutual coherence function of wave and the transport equation are precisely equivalent,[27] and this is also the case of space-time problems, e.g., of pulse wave propagation.[28]

APPENDIX A: DERIVATION OF M AND M_{11} FROM THE COHERENT POTENTIAL EQUATIONS (6.3) AND (6.5)

According to Eqs. (5.3) and (6.3),

$$\Delta q = \sum_\alpha q_\alpha - M, \quad \langle T^M_{\Delta q} \rangle = 0, \tag{A1}$$

where Σ_α means the summation over all the particles involved. Here, on referring to the multiple-scattering theory,[13] we may put the scattering matrix $T^M_{\Delta q}$ for Δq in the form

$$T^M_{\Delta q} = \sum_\alpha Q_\alpha + Q_{-M}. \tag{A2}$$

Here, Q_α represents the effective scattering matrix due to the particle q_α and, similarly, Q_{-M} that due to the part $-M$, obeying respectively the equations

$$Q_\alpha = T^M_\alpha \Big[1 + G_M \Big(\sum_{\beta \neq \alpha} Q_\beta + Q_{-M} \Big) \Big], \tag{A3}$$

$$Q_{-M} = T^M_{-M} \Big[1 + G_M \sum_\alpha Q_\alpha \Big], \tag{A4}$$

where T^M_α is the scattering matrix for q_α alone in the definite medium M, as defined by Eq. (4.7). Equations (A2), (A3), and (A4) are interpreted as follows: The total wave scattered by Δq is a sum of contributions coming from each particle and from the part $-M$. Each particle contribution Q_α is given by the particle T^M_α matrix applied to an effective wave. This effective wave consists of the incident wave and of the contributions from all the other particles and also from $-M$. The contribution from $-M$ is also given formally by the matrix T^M_{-M} applied to an effective wave which consists of the incident wave and of the contributions from all the particles.

Here, averaging both sides of Eq. (A2),

$$\langle T^M_{\Delta q} \rangle = \sum_\alpha \langle Q_\alpha \rangle + \langle Q_{-M} \rangle = 0, \tag{A5}$$

and, to evaluate the right-hand side according to Eqs. (A3) and (A4), we make the basic assumption that the correlations between T^M_α and $Q_\beta, \beta \neq \alpha$, are negligible, as made in Sec. 4 when deriving Eq. (4.12) as well as in the usual coherent potential approximation for disordered alloys.[13] Hence, the averaging of Eq. (A3) and the subsequent use of Eq. (A5) yields

$$\langle Q_\alpha \rangle = \langle T^M_\alpha \rangle [1 - G_M \langle Q_\alpha \rangle], \tag{A6}$$

which gives

$$\langle Q_\alpha \rangle = [1 + \langle T^M_\alpha \rangle G_M]^{-1} \langle T^M_\alpha \rangle \to \langle T^M_\alpha \rangle, \tag{A7}$$

as the volume V of the entire space tends to the infinite, since $\langle T^M_\alpha \rangle$ contains the averaging over the particle's center $\mathbf{a}$ and is given by $V^{-1} \int d\mathbf{a} \, \langle T^M_\alpha \rangle'$, as in Eq. (2.26), tending to zero as $V \to \infty$.

In the same way, from Eq. (A4), we obtain

$$\langle Q_{-M} \rangle = T^M_{-M} [1 - G_M \langle Q_{-M} \rangle], \tag{A8}$$

which gives

$$\langle Q_{-M} \rangle = [1 + T^M_{-M} G_M]^{-1} T^M_{-M} = -M. \tag{A9}$$

Thus, from Eqs. (A5), (A7), and (A9), we find

$$M = \sum_\alpha \langle T^M_\alpha \rangle = N V^{-1} \int d\mathbf{a} \, \langle T^M_\alpha \rangle', \tag{A10}$$

which is exactly the same as given by Eq. (4.12).

To obtain the matrix $M_{11}(1; 2)$, we first employ the expression (A3) with Eq. (A2) to find, with the aid of the condition (A5) and the incoherency between the incident and scattered waves,

$$\begin{aligned} \langle Q^*_\alpha(1) Q_\alpha(2) \rangle &= \langle T^{M*}_\alpha(1) T^M_\alpha(2) \rangle \\ &\times [1 - G^*_M(1) \langle Q^*_\alpha(1) \rangle - G_M(2) \langle Q_\alpha(2) \rangle \\ &+ G^*_M(1) G_M(2) \langle \{ T^{M*}_{\Delta q}(1) - Q^*_\alpha(1) \} \\ &\times \{ T^M_{\Delta q}(2) - Q_\alpha(2) \} \rangle], \end{aligned} \tag{A11}$$

which, as $V \to \infty$, tends to

$$\langle Q^*_\alpha(1) Q_\alpha(2) \rangle = \langle T^{M*}_\alpha(1) T^M_\alpha(2) \rangle F(1; 2), \tag{A12}$$

where, since Q_α is negligible as compared with $T^M_{\Delta q}$ in view of Eq. (A2),

$$F(1; 2) = 1 + G^*_M(1) G_M(2) \langle T^{M*}_{\Delta q}(1) T^M_{\Delta q}(2) \rangle. \tag{A13}$$

In the same way,

$$\langle Q^*_\alpha(1) Q_\beta(2) \rangle = \langle T^{M*}_\alpha(1) \rangle \langle T^M_\beta(2) \rangle F(1; 2), \quad \alpha \neq \beta. \tag{A14}$$

On the other hand, expressing the right-hand side of Eq. (A4) in terms of $T^M_{\Delta q}$ and Q_{-M} by use of Eq. (A2), Q_{-M} can be exhibited, in virtue of the relation (A9), by

$$Q_{-M} = -M [1 + G_M T^M_{\Delta q}]. \tag{A15}$$

Hence

$$\begin{aligned} \langle Q^*_{-M}(1) Q_{-M}(2) \rangle &= M^*(1) M(2) G^*_M(1) \\ &\times G_M(2) F(1; 2). \end{aligned} \tag{A16}$$

With exactly the same procedure, we find from Eqs. (A3) and (A15) that

$$\langle Q^*_\alpha(1) Q_{-M}(2) \rangle = -\langle T^{M*}_\alpha(1) \rangle M(2) F(1; 2). \tag{A17}$$

Thus, from Eq. (A2),

$$\begin{aligned} &\langle T^{M*}_{\Delta q}(1) T^M_{\Delta q}(2) \rangle \\ &= \Big\langle \Big[\sum_\alpha Q^*_\alpha(1) + Q^*_{-M}(1) \Big] \Big[\sum_\beta Q_\beta(2) + Q_{-M}(2) \Big] \Big\rangle, \end{aligned} \tag{A18}$$

where the right-hand side becomes, on employing Eqs. (A12), (A14), (A16), and (A17),

$$\left\langle\left[\sum_\alpha T_\alpha^{M*}(1)-M^*(1)\right]\left[\sum_\beta T_\beta^{M}(2)-M(2)\right]\right\rangle F(1;2), \tag{A19}$$

which, in virtue of Eq. (A10), further becomes

$$\sum_\alpha[\langle T_\alpha^{M*}(1)T_\alpha^{M}(2)\rangle-\langle T_\alpha^{M*}(1)\rangle\langle T_\alpha^{M}(2)\rangle]F(1;2). \tag{A20}$$

Here the terms $\langle T_\alpha^{M*}(1)\rangle\langle T_\alpha^{M}(2)\rangle$ become negligible as $V\to\infty$. Thus, according to Eq. (6.5), Eq. (A18) [with the right-hand side given by Eqs. (A20) and (A13)] provides us with

$$M_{11}(1;2)=\sum_\alpha\langle T_\alpha^{M*}(1)T_\alpha^{M}(2)\rangle, \quad V=\infty, \tag{A21}$$

which becomes the same as Eq. (4.21) with the replacement $\Sigma_\alpha\langle\cdots\rangle\to n\int d\mathbf{a}\langle\cdots\rangle'$.

Thus, the effective medium method introduced in Sec. 4 is found to be equivalent to the coherent potential approximation, but the former is more simple and straightforward than the latter, in the present case at least, in both the method and the physical interpretation.

APPENDIX B: EQUATIONS FOR $Z[\bar{j}^*,\bar{j}]$ IN MEDIA OBEYING GAUSSIAN STATISTICS

So far we have considered only the case of random particles. Also in the other typical case of the media obeying the Gaussian statistics, the various equations can be formulated in entirely the same way, even much more simply than in the former case. In Appendix B are summarized the equations necessary to derive the equation for $Z[\bar{j}^*,\bar{j}]$, together with the equations of coherence functions of wave derived from.

When the medium fluctuation can be assumed to obey the Gaussian statistics, the medium operator $\mathbf{q}(x)$ is given by Eq. (2.18) and therefore, to eliminate the operator $\delta/\delta c(x)$ from Eq. (3.8) for $Z[\bar{j}^*,\bar{j}]$, the only term it is necessary to evaluate becomes, on using Eq. (3.20),

$$\frac{\delta}{\delta c(x'')}Z[\bar{j}^*,\bar{j}]=\int dx\,dx'\left[\bar{j}(x)\left\{\frac{\delta}{\delta c(x'')}\mathbf{G}_q(x|x')\right\}j(x')+\bar{j}^*(x)\left\{\frac{\delta}{\delta c(x'')}\mathbf{G}_q^*(x|x')\right\}j^*(x')\right]Z[\bar{j}^*,\bar{j}]. \tag{B1}$$

Here, from Eq. (3.16),

$$\frac{\delta}{\delta c(x'')}\mathbf{G}_q(x|x')=\mathbf{G}_q(x|x'')\mathbf{G}_q(x''|x'), \tag{B2}$$

and hence, with the aid of the relation (3.19) and the notations (3.6), the right-hand side of Eq. (B1) can be written as

$$\int dx\,[\bar{j}(x)\mathbf{G}_q(\mathbf{x}|\mathbf{x}'')\psi(x'')+\bar{j}^*(x)\mathbf{G}_q^*(x|x'')\psi^*(x'')]Z[\bar{j}^*,\bar{j}]. \tag{B3}$$

Here, to the same approximation as used when deriving Eq. (7.8) from Eq. (7.7), $\mathbf{G}_q$ and $\mathbf{G}_q^*$ in Eq. (B3) may be replaced by the definite Green's functions G_M and G_M^*, respectively, yielding Eqs. (B1) with (B3) in the form

$$\frac{\delta}{\delta c(x'')}Z[\bar{j}^*,\bar{j}]=\mathbf{Q}(x'')Z[\bar{j}^*,\bar{j}], \tag{B4}$$

where $c(x)=0$ and

$$\mathbf{Q}(x)=\int dx'\,[\bar{j}(x')G(x'|x)\psi(x)+\bar{j}^*(x')G^*(x'|x)\psi^*(x)] \tag{B5}$$

obeys the commutation relations

$$[\psi(x),\mathbf{Q}(x')]=G(x|x')\psi(x'),$$
$$[\psi^*(x),\mathbf{Q}(x')]=G^*(x|x')\psi^*(x'). \tag{B6}$$

Thus, when $c(x)=0$, use of Eq. (2.18) with Eqs. (B4)–(B6) leads to the result

$$\mathbf{q}(x)\psi(x)Z=\psi(x)\mathbf{q}(x)Z=\int dx'\,D(x-x')\psi(x)\mathbf{Q}(x')Z$$
$$=\int dx'\,[M(x|x')\psi(x')+D(x-x')\mathbf{Q}(x')\psi(x)]Z, \tag{B7}$$

where

$$M(x|x')=D(x-x')G(x|x'),$$
$$M^*(x|x')=D(x-x')G^*(x|x'). \tag{B8}$$

Hence, Eq. (3.8) is finally exhibited by

$$\left[L(i\partial/\partial x)\psi(x)-\int dx'\,M(x|x')\psi(x')\right]Z[\bar{j}^*,\bar{j}]=\left[j(x)+\int dx'\,D(x-x')\mathbf{Q}(x')\psi(x)\right]Z[\bar{j}^*,\bar{j}], \tag{B9}$$

and Eq. (3.9) by the complex conjugate equation.

Here, the substitution of the moment expansion (3.14) in Eq. (B9) yields the equation for the moments of wave functions, $m_{\mu\nu}$, with respect to one of the coordinates of the original wave functions, say y_1, as

$$L(i\partial/\partial y_1)M_{\mu\nu}(x_1,\ldots,x_\mu;y_1,\ldots,y_\nu)$$
$$-\int dy_1'\,M(y_1|y_1')m_{\mu\nu}(x_1,\ldots,x_\mu;y_1',y_2,\ldots,y_\nu)$$
$$-\sum_{j=1}^{\mu}\int dx_j'\,D(y_1-x_j')G^*(x_j|x_j')\times m_{\mu\nu}(x_1,\ldots,x_j',\ldots,x_\mu;y_1,\ldots,y_\nu)$$
$$-\sum_{j\neq 1}^{\nu}\int dy_j'\,D(y_1-y_j')\times G(y_j|y_j')m_{\mu\nu}(x_1,\ldots,x_\mu;y_1,\ldots,y_j',\ldots,y_\nu)$$
$$=j(y_1)m_{\mu,\nu-1}(x_1,\ldots,x_\mu;y_2,y_3,\ldots,y_\nu). \tag{B10}$$

The corresponding equation with respect to any one of the coordinates of the complex conjugate wave functions, say x_1, is also obtained in the same form from the complex conjugate to Eq. (B9).

Finally, Eq. (B9) and its complex conjugate equation can be combined into the form of Eq. (7.20) with $\kappa_{10}(1)$, $\kappa_{01}(2)$, and $\kappa_{11}(1;2)$ replaced, in matrix elements, by

$$\kappa_{11}(x_1;x_2|x_1';x_2')=D(x_1-x_2)\delta(x_1-x_1')\delta(x_2-x_2'),$$
$$\kappa_{10}(x_1|x_1')=M^*(x_1|x_1')+\int dx\,D(x_1-x)\mathbf{Q}(x)\delta(x_1-x_1'),$$

$$\kappa_{01}(x_2|x_2') = M(x_2|x_2') + \int dx\, D(x_2 - x)\mathbf{Q}(x)\delta(x_2 - x_2'), \tag{B11}$$

which, in view of Eq. (B8), explicitly satisfies the condition (7.24).

APPENDIX C: PROOF OF THE CONDITIONS (5.12) AND (7.24) IN THE EFFECTIVE MEDIUM APPROXIMATION

In view of the definition (7.22) with Eq. (7.13), the condition (7.24) is proved if we can show the relation

$$\{T_\alpha^M(2) - T_\alpha^{M*}(1) + [G_M^*(1) - G_M(2)] \times T_\alpha^{M*}(1)T_\alpha^M(2)\}\big|_{x_1 = x_2} = 0, \tag{C1}$$

which, on averaging by $\langle\cdots\rangle_\alpha$, also becomes the condition (5.12) in view of Eqs. (4.12) and (4.21). To prove Eq. (C1), we first introduce the unitary matrix U_λ, defined by its matrix element

$$U_\lambda(x|x') = \exp[i\lambda\cdot x/2]\delta(x - x'), \tag{C2}$$

and then define the matrix A_λ for any matrix A, by the transformation

$$A_\lambda = U_\lambda A U_\lambda^{-1}, \tag{C3}$$

whose matrix elements are therefore given by

$$A_\lambda(x|x') = A(x|x')\exp[i\lambda\cdot(x - x')/2]. \tag{C4}$$

Here, since $q_\alpha(x)$ is a diagonal matrix, it follows that $q_{\alpha,\lambda} = q_\alpha$. Generally, the relation

$$(AB\cdots C)_\lambda = A_\lambda B_\lambda \cdots C_\lambda, \quad f_\lambda(A, B, \ldots) = f(A_\lambda, B_\lambda, \cdots) \tag{C5}$$

holds.

Here, we also introduce the Hermitian conjugate matrix $A^\dagger$ of A, defined by the matrix elements $A^\dagger(x|x') = A^*(x'|x)$, and hence, by Eqs. (C2) and (C3),

$$U_\lambda^\dagger = U_\lambda^{-1}, \quad A_\lambda^\dagger = U_\lambda A^\dagger U_\lambda^{-1} = (A^\dagger)_\lambda. \tag{C6}$$

Particularly, the matrix elements of G_λ and $G^\dagger{}_{-\lambda}$ are given by

$$G_\lambda(x|x') = G(x - x')\exp[i\lambda\cdot(x - x')/2],$$
$$G^\dagger{}_{-\lambda}(x|x') = G^*(x' - x)\exp[i\lambda\cdot(x' - x)/2]. \tag{C7}$$

The T_α^M matrix for q_α in the medium M is connected to q_α, according to the formula (4.8), by

$$q_\alpha = T_\alpha^M(1 + G\,T_\alpha^M)^{-1} = (1 + T_\alpha^{M\dagger}G^\dagger)^{-1}T_\alpha^{M\dagger}, \tag{C8}$$

where the last expression is the Hermitian conjugate of the former. Hence, performing the unitary transformation of Eq. (C8) by U_λ and $U_{-\lambda}$, separately, the use of the formula (C5) leads to

$$q_\alpha = q_{\alpha,\lambda} = T_{\alpha,\lambda}^M(1 + G_\lambda T_{\alpha,\lambda}^M)^{-1} = g_{\alpha,-\lambda} = (1 + T_{\alpha,-\lambda}^{M\dagger}G^\dagger{}_{-\lambda})^{-1}T_{\alpha,-\lambda}^{M\dagger}, \tag{C9}$$

which gives the relation between $T_{\alpha,\lambda}^M$ and $T_{\alpha,-\lambda}^{M\dagger}$ as

$$\tilde{v}_\lambda \equiv T_{\alpha,\lambda}^M - T_{\alpha,-\lambda}^{M\dagger} + T_{\alpha,-\lambda}^{M\dagger}(G^\dagger{}_{-\lambda} - G_\lambda)T_{\alpha,\lambda}^M = 0, \tag{C10}$$

whose matrix elements are given, on reference to Eq. (C7), by

$$\tilde{v}_\lambda(x_1|x_2) = T_\alpha^M(x_1|x_2)e^{i\lambda\cdot(x_1 - x_2)/2} - T_\alpha^{M*}(x_2|x_1)e^{i\lambda\cdot(x_2 - x_1)/2} + \int dx\,dx'\,[G^*(x' - x)e^{i\lambda\cdot(x' - x)/2} - G(x - x')e^{i\lambda\cdot(x - x')/2}] \times T_\alpha^{M*}(x|x_1)T_\alpha^M(x'|x_2)e^{i\lambda\cdot(x + x' - x_1 - x_2)/2} = 0. \tag{C11}$$

Thus, on multiplying both sides of Eq. (C11) by $\exp[i\lambda\cdot\{(x_1 + x_2)/2 - \rho\}]$ and then performing the λ integration over the entire range of $\infty \geqslant \lambda \geqslant -\infty$, we obtain, say, $v_\rho(x_1|x_2)$, given by

$$v_\rho(x_1|x_2) = T_\alpha^M(x_1|x_2)\delta(x_1 - \rho) - T_\alpha^{M*}(x_2|x_1)\delta(x_2 - \rho) + \int dx\,dx'[G^*(x' - x)\delta(x' - \rho) - G(x - x')\delta(x - \rho)] \times T_\alpha^{M*}(x|x_1)T_\alpha^M(x'|x_2) = 0, \tag{C12}$$

which is equivalent to Eq. (C1) since, for arbitrary function $f(x_1; x_2)$, the relation holds, in matrix form, as

$$\int dx_1\,dx_2 v_\rho(x_1|x_2)f(x_1; x_2) = \{T_\alpha^M(2) - T_\alpha^{M*}(1) + [G^*(1) - G(2)]T_\alpha^{M*}(1)T_\alpha^M(2)\}f(1;\ 2)\big|_{x_1 = x_2 = \rho} = 0. \tag{C13}$$

The relation (C12) corresponds to the usual optical condition, indicating that the total scattering cross-section is proportional to the imaginary part of the scattering amplitude, but is more general in the two points that $v_\rho(x_1|x_2) = 0$ not only for arbitrary x_1 and x_2 but also for arbitrary ρ and further for any absorbing medium, described by a complex M; for the latter point, the Fourier transform of $G^* - G$, say $\tilde{G}^*(s) - \tilde{G}(s) = [L(s) - \tilde{M}^*(s)]^{-1} - [L(s) - \tilde{M}(s)]^{-1}$, tends to $2\pi i\delta[L(s)]$ as $M \to 0$, and therefore the integral in Eq. (C12) tends, when integrated with respect to ρ over the entire range, to be contributed to only by those components of the Fourier transform of the integrand obeying $L(s) = 0$, or from the "shell" components. Here, on the other hand, when $M \neq 0$ and/or the change with respect to the coordinates ρ is large enough, the off-shell components also become important enough to make appreciable contributions to the integral.

[1] L. L. Foldy, Phys. Rev. **67**, 107 (1945).
[2] R. C. Bourret, Nuovo Cimento **26**, 1 (1962); Can. J. Phys. **40**, 782 (1962).
[3] K. Furutsu, J. Research. of the National Bureau of Standards (U. S. Dept. of Commerce) NBS. D **67**, 303 (1963).
[4] U. Frisch, in *Probabilistic methods in applied mathematics*, edited by A. T. Barucha-Reid, (Academic, New York, 1968), Vol. 1, p. 76.
[5] V. I. Tatarski, "The effects of the turbulence atmosphere on wave propagation," Reproduced by National Technical Information Service (1971). (English translation available from the U. S. Dept. of Commerce, Springfield, VA 22151).
[6] See, for example, Yu. N. Barabanenkov, Yu. A. Kravtsov, S. M. Rytov, and V. I. Tatarskii, Sov. Phys. Usp. **13**, 551 (1971).
[7] K. Furutsu, J. Opt. Soc. Am. **62**, 240 (1972).
[8] E. A. Novikov, Zh. Eksp. Teor. Fiz. **47**, 1919 (1964).
[9] K. Furutsu, Radio Sci. **10**, 29 (1975).
[10] J. B. Keller, in *Proceedings of the Symposium on Applied Mathematics*, Vol. XVI (Am. Math. Soc., Providence, R. I., 1964); M. J. Beran, *Statistical Continuum Theories* (Wiley, New York, 1968).
[11] The result (2.8) has a close connection with that in: J. Schwinger, J. Math. Phys. **2**, 407 (1961).
[12] The time-independent version of Eq. (2.29) was obtained in Ref. 9.
[13] See, for example, B. Velický, Phys. Rev. **184**, 614 (1969); B. Velický, S. Kirkpatrick, and H. Ehrenreich, Phys. Rev. **175**, 747 (1968).

[14]K. Furutsu, Phys. Rev. **168**, 167 (1968).
[15]L. C. Lee, J. Math. Phys. **15**, 1431 (1974).
[16]K. Furutsu and Y. Furuhama, Opt. Acta **20**, 707 (1973).
[17]I. M. Besieris, J. Math. Phys. **19**, 2533 (1978).
[18]M. Tateiba, IEEE Trans. Antennas Propagation **AP-23**, 493 (1975).
[19]K. Furutsu, Radio Sci. **14**, 287 (1979).
[20]V. U. Zavovotnyi, V. I. Klyatskin, and V. I. Tatarskii, Zh. Eksp. Teor. Fiz. **73**, 481 (1977) [Sov. Phys. JETP **46**, 252 (1977)].
[21]I. G. Yakushkin, Izv. VUZ Radiofiz. **21**, 1194 (1978) [Radiophys. Quantum Electron., **21**, 835 (1978)].
[22]R. Dashen, J. Math. Phys. **20**, 894 (1979).
[23]K. Furutsu, J. Math. Phys. **17**, 1252 (1976).
[24]M. E. Gracheva, A. S. Gurvich, S. S. Kashkarov, and V. V. Pokasov, Preprint, Dept. of Oceanology, Physics, Atmoshperes, and Geography, USSR Academy of Sciences [Trans. LRG-73-T-28, available from Aerospace Corp., Library Series, P. O. Box 92957, Los Angeles, CA 90009].
[25]W. P. Brown, Jr., J. Opt. Soc. Am. **62**, 966 (1972).
[26]V. I. Tatarskii, A. S. Gurvich, B. S. Elepov, V. V. Pokasov, and K. K. Sabelfeld, Opt. Acta **26**, 531 (1979).
[27]L. S. Dolin, Izv. VUZ Radiofiz. **7**, 380 (1964).
[28]K. Furutsu, J. Math. Phys. **20**, 617 (1979).

A note on the Schrödinger equation for the $x^2 + \lambda x^2/(1 + gx^2)$ potential

N. Bessis and G. Bessis
Laboratoire de Spectroscopie Théorique, Université Claude Bernard, Lyon I, 69622 Villeurbanne, France.

(Received 15 April 1980; accepted for publication 13 June 1980)

The energy levels and wave functions of the Schrödinger equation involving the potential $x^2 + \lambda x^2/(1 + gx^2)$ are calculated by the variational method, for any range of λ and g, without having to resort to numerical quadrature. Using properly scaled (in λ and g) harmonic oscillator functions as a basis set, an easy to compute analytical expression of the current Hamiltonian matrix element is derived. Perturbative results are also given.

I. INTRODUCTION

Recently, special interest has been drawn to the resolution of the following eigenequation:

$$\left(\frac{d^2}{dx^2} - V(x) + E\right)\psi(x) = 0, \tag{1}$$

where

$$V(x) = x^2 + \lambda x^2/(1 + gx^2). \tag{2}$$

Interest in this type of interaction arises in several areas and these have been summarized by Mitra[1] and Kaushal[2]. In particular, this type of potential occurs when considering models in laser theory.[3,4] The ground state and the two first energy levels were first computed by Mitra,[1] for a large range of λ and g ($\lambda, g = 0$ to 100) within the variational Rayleigh–Ritz framework. Properly λ-scaled harmonic oscillator eigenfunctions have been chosen as a basis set for the representation of the Hamiltonian operator. By repeated use of the recurrence formula for the Hermite polynomials, it has been shown by Mitra that the current variational matrix element can be obtained by a recursive procedure from the knowledge of one unique matrix element H_{11}. Nevertheless, it was possibly overlooked by Mitra that this H_{11} can be directly expressed in terms of the Error function. Therefore Mitra had to resort to numerical quadrature for obtaining H_{11}, and encountered some difficulties, especially for large values of g. Furthermore, the above recursive procedure could lead to numerical instabilities and therefore further discussion, based on some other algorithm may be necessary, even though the actual results for the eigenvalues may not differ significantly.

On the other hand, Kaushal[2] has used a relatively complex perturbation algorithm in order to obtain an asymptotic expansion of the eigenspectrum but had to restrict the calculation to rather small range of g ($g = 0$ to 1) and large range of λ ($\lambda = 0$ to 100). At the same time, the work of Kaushal may be questionable in that the author expands $1/(1 + gx^2)$ in a power series for $gx^2 < 1$, but does not present any estimate on the error made in restricting the domain of x in this way.

It is shown, in the present paper, that as long as the harmonic oscillator eigenfunctions are used as basis set, the current matrix element of the Hamiltonian can be calculated exactly for any value of λ and of g without having to resort numerical integration. Taking advantage of a two-parameter λ and g-scale transformation, the determination of the spectrum of the eigenequation (1) is reinvestigated both within a variational and a perturbational scheme.

II. METHOD

It seems quite reasonable to consider the eigenequation (1) as a perturbed (and properly scaled via b) harmonic oscillator wave equation

$$\left(\frac{d^2}{dx^2} - b^2x^2 + \epsilon_v\right)\phi_v = 0 \tag{3}$$

and to use a basis set for a variational procedure or a Rayleigh–Schrödinger perturbation scheme, the well-known orthonormal eigenfunctions

$$\phi_v = \left(\frac{b}{\pi}\right)^{1/4}\left(\frac{1}{2^v v!}\right)^{1/2} e^{-bx^2/2} H_v(b^{1/2}x), \tag{4}$$

where $v = 0, 1, 2, \ldots$ and H_v is a Hermite polynomial of degree v.

The associated eigenvalues are

$$\epsilon_v = 2b\,(v + 1/2). \tag{5}$$

Since ϕ_v involves a polynomial and since the eigenequation (1) is of even parity, it follows that, either in a variational or in a perturbation treatment, the critical part of the calculation is the evaluation of the following basic integral:

$$I_k = 2\left(\frac{b}{\pi}\right)^{1/2}\int_0^\infty e^{-bx^2}(b^{1/2}x)^{2k}\frac{dx}{1 + gx^2}. \tag{6}$$

A. Analytical expression of the basic integral

It is easily found that I_k can be determined from the very simple recursion relationship

$$I_{k+1} = -\frac{b}{g}I_k + 2\left(\frac{b}{\pi}\right)^{1/2}\frac{1}{g}b^{k+1} \times\int_0^\infty e^{-bx^2}x^{2k}\,dx \tag{7}$$

or

$$I_{k+1} = -\frac{b}{g}\left(I_k - \frac{(2k-1)!!}{2^k}\right). \tag{8}$$

This relation yields

$$I_k = -\left(\frac{b}{g}\right)^k I_0 - \sum_{s=0}^{k-1}\frac{(2s-1)!!}{2^s}\left(-\frac{b}{g}\right)^{k-s}. \tag{9}$$

Finally, the only integral to be calculated is

 0022-2488/80/122780-06$1.00

$$I_0 = 2\left(\frac{b}{\pi}\right)^{1/2}\int_0^\infty e^{-bx^2}\frac{dx}{1+gx^2}. \tag{10}$$

It should be noted that Mitra,[1] also found that the central element to be computed was $H_{11} = I_0$. The integral (10) simply defines the complementary error function, Erfc (see, for instance, Ref. 5, p. 302), namely

$$I_0 = \sqrt{\pi} z e^{z^2}\mathrm{Erfc}(z) \quad \text{with } z = (b/g)^{1/2}. \tag{11}$$

The Erf or Erfc = 1 − Erf functions have been, for many years, extensively and accurately calculated and tabulated (see, for instance, Refs. 5–8). Furthermore, series and asymptotic expansions are available.[5,9,10] For large z, the following asymptotic expansion of I_0 can be used[5]:

$$I_0 \simeq 1 + \sum_{m=1}^{\infty}(-)^m(2m-1)!!\left(\frac{1}{2z^2}\right)^m. \tag{12}$$

On the other hand, for small z, one can use[5]

$$I_0 = \sqrt{\pi} z e^{z^2}\left(1 - \frac{2}{\sqrt{\pi}} z \sum_{m=0}^{\infty}(-)^m \frac{z^{2m}}{m!(2m+1)}\right). \tag{13}$$

Let us mention that we have verified formula (13) for $b = 1$, $g = 100$ (i.e., $z = 0.1$). Limiting ourselves to m up to 3, we found $I_0 = 0.15889286$. From tables and expression (11), we obtained exactly the same result up to the last figure.

One can also use the following Hasting's formula

$$I_0 = \frac{2}{\sqrt{\pi}} z \sum_{i=1}^{5} a_i t^i + \epsilon(z); \quad |\epsilon(z)| \leqslant 1.5\times 10^{-7}, \tag{14}$$

where

$t = 1/(1+pz)$; $p = 0.3275911$; $a_1 = 0.254829592$;

$a_2 = -0.284496736$;

$a_3 = 1.421413741$; $a_4 = -1.453152027$; $a_5 = 1.061405429$.

At the expected limit of accuracy, one obtains $I_0 = 0.15889290$.

From a comparative study of the accuracy of all these formulas for several ranges of z, we found that the most convenient expression (except for really very small z) is the one given by Henrici[10]

$$I_0 = z^2 \sum_{n=0}^{\infty} \frac{(1/2)_n}{(n+1)!} \frac{1}{L_n^{-1/2}(-z^2) L_{n+1}^{-1/2}(-z^2)}, \tag{15}$$

where $(a)_n = a(a+1)(a+2)\cdots(a+n-1)$; $(a)_0 = 1$. The associated Laguerre polynomial $L_n^{-1/2}$ as well as the factorials are very easily generated by recursion.

B. Choice of the scale transform

It is clear that for $g \ll \lambda$ and also for very large g, (1) mainly behaves as an oscillator wave equation (3) with $b^2 = 1+\lambda$ or $b^2 = 1$ respectively. Hence, a physically good scaling has to depend simultaneously upon λ and g. In order to extract a harmonic potential from the second term of $V(x)$ [Eq. (2)], one can replace gx^2 by its average and choose, for each state "v" to be computed,

$$b^2 = 1+\alpha = 1+\lambda/(1+g(v+1/2)). \tag{16}$$

Nevertheless, the drawback of this choice is the creation of a nonorthogonal basic set. In order to avoid this disadvantage, we have chosen a unique average scaling $\alpha = \lambda/(1+g/2)$ which is well adapted to most low lying states. Of course, if one is interested solely in one specific state, for example "v", it is more convenient to choose $\alpha = \lambda/(1+g(v+1/2))$.

It is interesting to compare our scaling formula (16) with that of Mitra[1] and Kaushal,[2] $b^2 = 1+\lambda$. In Table I, the zeroth order ground state energies, which are obtained when alternately using Mitra's scaling and ours (16), are compared with the exact value. It appears that, if for small g both scalings are fairly good, for large g the former scaling fails to predict even the order of magnitude of the exact ground state energy. Our scaling gives the correct order of magnitude of the energy levels for all λ and g values.

C. Representation matrix of the Hamiltonian

Since the eigenequation (1) is of even parity, one can treat separately even and odd states and distinguish between the even and odd normalized basis set

$$\phi_{2n} = \left(\frac{b}{\pi}\right)^{1/4} e^{-bx^2/2} \sum_{s=0}^{n} d_{ns}^{(P)} (b^{1/2}x)^{2s},$$

$$\phi_{2n+1} = \left(\frac{b}{\pi}\right)^{1/4} e^{-bx^2/2} \sum_{s=0}^{n} d_{ns}^{(I)} (b^{1/2}x)^{2s+1}, \tag{17}$$

where the $d_{ns}^{(P)}$ and the $d_{ns}^{(I)}$ are the coefficients of the normalized Hermite polynomials of degree $2n$ and $2n+1$, respectively,

$$d_{ns}^{(P)} = (-)^{n+s} 2^{2s-n} \sqrt{(2n)!}/((2s)!(n-s)!),$$

$$d_{ns}^{(I)} = (n+1/2)^{1/2} d_{ns}^{(P)}/(s+1/2). \tag{18}$$

Setting $b^2 = 1+\alpha$, the potential $V(x)$ [Eq. (2)] can be rewritten as

$$V(x) = b^2x^2 + [((\lambda-\alpha)x^2 - \alpha g x^4)/(1+gx^2)]. \tag{19}$$

TABLE I. Zeroth order ground state energies calculated from different scaling procedures

(a) g-dependent scaling $E = \sqrt{1+\lambda/(1+g/2)}$;
(b) exact value;
(c) $E = \sqrt{1+\lambda}$

g \ λ		0.1	1	10	100	200
0.1	(a)	1.0465	1.3973	3.2440	9.8101	13.8375
	(b)	1.0432	1.3805	3.2503	9.9762	14.1032
1	(a)	1.0328	1.2910	2.7688	8.2260	10.0499
	(b)	1.02426	1.2324	2.7823	9.3594	13.4687
10	(a)	1.0083	1.0801	1.6330	4.2031	5.8595
	(b)	1.0059	1.0592	1.5800	5.7939	9.2811
100	(a)	1.00098	1.00976	1.0936	1.7207	2.2184
	(b)	1.00084	1.00841	1.0840	1.8363	2.6631
500	(a)	1.000495	1.00494	1.04833	1.4107	1.7263
	(b)	1.00044	1.00442	1.04419	1.4413	1.8812
(c)		1.0488	1.4142	3.3166	10.050	14.177

TABLE II. The first four energy levels for different values of λ and g in increasing order of excitation (Jacobi diagonalization of an 18×18 matrix).

g \ λ	0.1	0.5	1	2	5	10	20	50	100	200	500
0.1	1.04317371	1.20303955	1.38053180	1.68561740	2.38954155	3.25026122	4.51242099	7.0686947	9.97618009	14.1032168	22.3084299
	3.12008186	3.57080929	4.07988301	4.96859933	7.05096392	9.61906641	13.3973600	21.0607383	29.7811911	42.1612060	66.7760954
	5.18109479	5.87158370	6.66791910	8.08680404	11.4848086	15.7293363	22.0055699	34.7638297	49.2926905	69.9231255	110.945882
	7.23100998	8.12187144	9.16656747	11. 0627486	15.7066621	21.5910055	30.3432738	48.1814982	68.5130522	97.3906193	154.818813
0.5	1.03121454	1.15156359	1.29295052	1.55104915	2.19211847	3.01685429	4.25506611	6.79278953	9.69215782	13.8139887	22.0149518
	3.07390256	3.36380139	3.71390237	4.37658192	6.12105873	8.48227060	12.1236133	19.6850376	28.3625979	40.7156819	65.3089021
	5.09306915	5.46321387	5.92063165	6.81529745	9.32076345	12.9480334	18.7961421	31.2380423	45.6365729	66.1860474	93.6226964
	7.10585043	7.52788119	8.05237875	9.09000674	12.0931670	16.6793649	24.4519211	41.549253	61.5778732	90.2681487	147.546529
1	1.02418675	1.11858946	1.23237205	1.44732998	2.01300219	2.782330	3.977692	6.47811496	9.35941803	13.4687482	21.6587477
	3.05165067	3.25584210	3.50742053	3.99841495	5.37944	7.417506	10.7906303	18.1287122	26.705965	38.992519	63.528936
	5.05928655	5.29506292	5.58986086	6.17851432	7.92192614	10.7010259	15.698561	27.3753456	41.4410998	61.7775337	102.558118
	7.06549833	7.32454029	7.64831681	8.29493343	10.224358	13.3883239	19.409653	34.6454207	53.839093	82.0052851	138.855208
2	1.01789466	1.0870649	1.170485	1.331863	1.782435	2.442570	3.534937	5.931990	8.758278	12.827070	20.979385
	3.031773	3.186776	3.329042	3.649514	4.593627	6.09516618	8.838714	15.497575	23.743326	35.803455	60.139256
	5.035846	5.175886	5.34849066	5.6940304	6.739675	8.490523	11.94156	21.395858	34.257779	53.80779	93.926914
	7.03474084	7.226541	7.381135	7.750547	8.868673	10.732244	14.448137	25.294215	41.494948	67.626030	122.80434
5	1.009787	1.048807	1.0972941	1.193317	1.47402433	1.918909	2.7446638	4.7584713	7.342857	11.215761	19.168545
	3.01608085	3.0803718	3.1606623	3.320997	3.80001389	4.5915684	6.152962	10.586344	17.1828134	27.99277	51.18189
	5.01560022	5.0780463	5.1562048	5.312852	5.78531644	6.5803330	8.1988138	13.107241	21.205211	36.237616	72.140082
	7.0169852	7.0849202	7.1698253	7.339591	7.8485497	8.695784	10.39439	15.451712	23.895738	40.581092	85.176092
10	1.0059428	1.0296851	1.05929700	1.1183019	1.293580	1.5800249	2.132445	3.6443906	5.793947	9.2811627	16.73919
	3.0088109	3.04405055	3.0880908	3.1761407	3.4400419	3.8790372	4.7537844	7.350187	11.572198	19.551651	39.580823
	5.00828042	5.0414117	5.0828477	5.1657921	5.415200	5.8327692	6.6746838	9.2463907	13.62879	22.490906	48.071034
	7.0090376	7.04518677	7.0903704	7.1807285	7.4517292	7.9031549	8.8051293	11.504728	15.988706	24.95478	51.883453
20	1.0034334	1.0171614	1.0343083	1.068558	1.1709608	1.3404716	1.6751703	2.6454669	4.157188	6.850189	13.278094
	3.0046566	3.023282	3.0465640	3.093123	3.2327765	3.4654425	3.9304376	5.3226407	7.633095	12.212003	25.5030372
	5.0043275	5.0216391	5.0432824	5.0865814	5.2165782	5.4335727	5.8688196	7.1846197	9.409245	13.950132	27.870513
	7.0047083	7.0235415	7.0470827	7.0941635	7.235395	7.4707439	7.9413024	9.3518730	11.699220	16.382007	30.39511
50	1.001569	1.0078473	1.0156933	1.0313808	1.0784008	1.1566708	1.3127555	1.7774654	2.5401081	4.0209960	8.1288805
	3.0019372	3.0096860	3.0193720	3.0387439	3.0968584	3.1937121	3.3874053	3.9683697	4.9362526	6.8704990	12.659978
	5.001808	5.0090444	5.0180891	5.0361795	5.0904585	5.18094088	5.3620252	5.9060391	6.8154661	8.6452110	14.220266
	7.001943	7.0097191	7.0194382	7.0388763	7.0971901	7.1943775	7.3887445	7.9717804	8.9432892	10.885480	16.706681
100	1.0008411	1.0042054	1.0084106	1.0168203	1.0420438	1.0840643	1.1680354	1.4193826	1.8363850	2.66311244	5.0840857
	3.0009831	3.0049158	3.0098317	3.0196635	3.0491587	3.0983170	3.1966324	3.4915694	3.9830992	4.9660377	7.9138556
	5.0009257	5.0046377	5.0092755	5.0185512	5.0463792	5.09276246	5.1855409	5.4639723	5.9283525	6.8584230	9.6605099
	7.0009845	7.0049224	7.0098449	7.0196899	7.0492246	7.0984491	7.1968972	7.4922353	7.9844448	8.9687849	11.921169
200	1.0004420	1.0022101	1.0044203	1.0088404	1.0221001	1.0441967	1.0883796	1.2208439	1.4413330	1.8812271	3.1920300
	3.0004955	3.0024779	3.0049558	3.0099115	3.0247789	3.0495579	3.0991156	3.2477883	3.4955736	3.9911350	5.4777444
	5.0004729	5.0023648	5.0047296	5.0094592	5.0236482	5.0472970	5.0945959	5.2365039	5.4730537	5.9462928	7.3672267
	7.0004958	7.0024791	7.0049583	7.0099166	7.0247916	7.0495833	7.0991665	7.2479157	7.4958292	7.9916496	9.4790563
500	1.00011849	1.0009245	1.0018491	1.0036983	1.0092456	1.0184910	1.0369811	1.0924449	1.1848632	1.3696191	1.9232260
	3.0001992	3.0009963	3.0019926	3.0039852	3.0099630	3.0199260	3.0398520	3.0996301	3.1992601	3.3985199	3.9962969
	5.0001928	5.0009640	5.0019279	5.0038559	5.0096399	5.0192799	5.0385600	5.0964009	5.1928043	5.3856183	5.9641161
	7.0001992	7.0009964	7.0019928	7.0039857	7.0099644	7.0199288	7.0398576	7.0996440	7.1992879	7.3985755	7.9964367

Using the definition (6) of I_k and the recursion relation (8), one obtains the following expression of the current Hamiltonian matrix element between basis functions, even and odd respectively (see Appendix),

$$\mathscr{H}^{(P)}_{nm} = b\,\delta_{nm}(4n+1) + \sum_{s=0}^{n}\sum_{t=0}^{m} d^{(P)}_{ns} d^{(P)}_{mt} \mathscr{V}_{s+t+1},$$

$$\mathscr{H}^{(I)}_{nm} = b\,\delta_{nm}(4n+3) + \sum_{s=0}^{n}\sum_{t=0}^{m} d^{(I)}_{ns} d^{(I)}_{mt} \mathscr{V}_{s+t+2}, \tag{20}$$

where

$$\mathscr{V}_k = \frac{\lambda}{b} I_k - \frac{\alpha}{b}\frac{(2k-1)!!}{2^k}. \tag{21}$$

Finally, using (8) the $\mathscr{H}_{nm}$ are very easily computed recursively either by hand or by a very simple routine in terms of I_0 [Eq. (15)].

III. RESULTS AND DISCUSSION

Eigenvalues and eigenfunctions of wave equation (1) have been obtained on the basis of scaled orthonormal harmonic oscillator functions by the variational method for a large range of g and λ ($g, \lambda = 0.1$ to 500). Moreover, since our zeroth order energies (see Table I) are in good agreement with the order of magnitude of the exact values, we have verified the accuracy of the results given by a traditional Rayleigh–Schrödinger perturbation calculation (first and second order). (We have considered that the values obtained by Mitra and corroborated afterwards by our variational calculations converge towards the exact values.)

A. Variational calculations

Since the basis set is orthonormal, it is well known that the variational procedure reduces to the diagonalization of the even (or odd) $\mathscr{H}_{nm}$ matrix representation. We have used the Jacobi diagonalization procedure for different sizes of $N \times N$ matrices ($N = 4$ to $N = 18$). It is worthwhile to note that for $N = 4$, we obtain an overall accuracy of three significant figures. For $N = 16$, our results are identical with those calculated by Mitra (except for a discrepancy for the first excited state when $g = 0.5, \lambda = 100$). In order to obtain eight significant figures, the calculations have been performed for $N = 18$. The first six energy levels for different values of λ and g ($\lambda = 0.1$ to 500, $g = 0.1$ to 500) are given in Table II.

B. Perturbational results

Using perturbation theory, we have numerically diagonalized the Hamiltonian matrix. We found that either for small g/b ratios (this is the case, in particular, when g is small and λ is large), or for large g/b ratios, the eigenvalues thus obtained compare favorably with the exact values, even when the perturbation process is limited to the first order. In this last case, analytical expressions of the energies are very easy to obtain in terms of $I_0(b, g)$ and g/b (or b/g). For the ground state and the two first excited states, for instance, one obtains

$$E(v=0) = b - \frac{\alpha}{2b} - \frac{\lambda}{g}(I_0 - 1),$$

$$E(v=1) = 3\left(b - \frac{\alpha}{2b}\right) + \frac{\lambda}{g}\left(2\frac{b}{g}(I_0 - 1) + 1\right), \tag{22}$$

TABLE III. Perturbative results for the gound state energies (a) first order [Eq. (22)]; (b) second order; (c) exact value; (d) Kaushal (Ref. 2)

g \ λ		0.1	0.5	1	10	50	100
0.1	a	1.0432	1.2034	1.3814	3.25114	7.0689	9.9768
	b	1.04318	1.20312	1.3807	3.25066	7.0689	9.9763
	c	1.04317	1.203039	1.380532	3.250261	7.068696	9.97618
	d	(1.04305)	(1.20290)	(1.38045)	(3.250244)	(7.068692)	(9.9761778)
0.5	a	1.0315	1.1547	1.3001	3.0306	6.7977	9.7023
	b	1.03140	1.1523	1.29459	3.0231	6.7973	9.6960
	c	1.03121	1.15156	1.29295	3.01685	6.79278	9.69215
	d	(1.032)	(1.103)	(1.263)	(3.0139)	(6.7922)	(9.69185)
1	a	1.0248	1.1239	1.2446	2.8190	6.4929	9.3860
	b	1.0246	1.1207	1.2361	2.7989	6.4928	9.3727
	c	1.0241	1.1185	1.2323	2.7823	6.4781	9.3594
	d	(1.015)	(1.100)	(1.227)	(2.754)	(6.472)	(9.3567)
10	a	1.00596	1.0302	1.0613	1.6439	3.8029	5.9362
	b	1.00594	1.0297	1.0593	1.5907	3.7266	5.9098
	c	1.00594	1.02968	1.059297	1.580025	3.64439	5.7939
50	a	1.00157	1.00788	1.01584	1.1671	1.8781	2.731
	b	1.00157	1.007848	1.10569	1.1570	1.7916	2.588
	c	1.001569	1.007847	1.01569	1.15667	1.77746	2.540108
100	a	1.00084	1.00421	1.00845	1.0874	1.4665	1.949
	b	"	1.00420	1.00841	1.0841	1.4222	1.8504
	c	"	"	"	1.08406	1.41938	1.8364

TABLE IV. Perturbative results for the first three excited states energies (a) first order [Eqs. (22) and (23)]; (b) Kaushal (Ref. 2) to be compared to the exact values in Table II.

g \ λ	v	0.1	0.5	1	10	50	100
0.1	1	3.1201 (3.1189)	3.5725 (3.5695)	4.0834 (4.0789)	9.621 (9.61843)	21.073 (21.06057)	29.810 (29.78110)
	2	5.1823 (5.175)	5.8852 (5.864)	6.6974 (6.661)	15.7789 (15.7228)	34.7789 (34.7621)	49.298 (49.29177)
	3	7.2288 (7.208)	8.1233 (8.093)	9.189 (9.132)	21.618 (21.554)	48.160 (48.1716)	68.503 (68.5079)
0.5	1	3.073 (3.061)	3.366 (3.362)	3.724 (3.459)	8.509 (8.40)	19.785 (19.663)	28.657 (28.3514)
	2	5.098 (5.022)	5.5147 (5.311)	6.053 (5.852)	13.586 (12.560)	31.562 (31.016)	45.783 (45.521)
	3	7.101 (6.916)	7.531 (7.011)	8.098 (7.556)	17.144 (15.675)	41.324 (41.170)	61.194 (61.321)
1	1	3.050 (2.976)	3.255 (3.049)	3.513 (3.305)	7.478 (6.754)	18.228 (17.952)	27.138 (26.615)
	2	5.130 (4.801)	5.346 (4.499)	5.7318 (4.634)	11.937 (9.355)	28.413 (26.779)	42.022 (41.030)
	3	7.061 (6.489)	7.322 (5.448)	7.674 (5.212)	14.168 (10.091)	34.667 (33.092)	52.991 (52.802)

$$E(v=2)=5\left[b-\frac{\alpha}{2b}\right]-\frac{\lambda}{g}\times\left\{\left[\frac{1}{2}+2\frac{b}{g}+2\left(\frac{b}{g}\right)^2\right](I_0-1)+\left(\frac{b}{g}-\frac{1}{2}\right)\right\}, \tag{23}$$

$$E(v=3)=7\left[b-\frac{\alpha}{2b}\right]+\frac{\lambda}{g}\left\{\left[3\frac{b}{g}+4\left(\frac{b}{g}\right)^2+\frac{4}{3}\left(\frac{b}{g}\right)^3\right]\times(I_0-1)+\left[\frac{2}{3}\left(\frac{b}{g}\right)^2+\frac{b}{g}+1\right]\right\},$$

where $b^2=1+\alpha$; $\alpha=\lambda/(1+g/2)$. For small g/b ratios, one can use these formulas in conjunction with the asymptotic expression (12) of I_0, and then determine the energies as a series of $g/b \ll 1$. Only a few terms (3 to 4) are needed in order to obtain a good degree of accuracy. For large g/b ratios, the same formulas (22) and (23) can be used in conjunction with (13) or (14) and yield an expansion of the energies in powers of $g/b \ll 1$.

First order and second order energies for the ground state are furnished in Table III in comparison with the exact values. From Table IV, it is shown that the first excited states energy levels which are obtained from our simple formulas (22) and (23) are, on the whole, closer to the exact values than those obtained by Kaushal[2] from a more complex formula. This is certainly due to the introduction of the g-dependent scaling [Eq. (16)].

It should be noted that, from the expression (20) of $\mathscr{H}_{nm}$, second order energies can also be obtained by a series expansion in b/g (or g/b). Nevertheless, in our opinion, a perturbational treatment of the wave equation (1) could be more conveniently tackled in the framework of the "perturbed ladder operator method".[11] Analytical results will be given elsewhere.

Finally, one can say that, after properly scaling the harmonic oscillator basis set, eigenvalues and eigenfunctions of the wave equation (1) can be obtained for a wide range λ and of g without difficulties. (The coefficients of the eigenfunctions are not given here but are available on request to the authors.) Indeed the computer program for the calculations is straightforward: double precision is sufficient and only a few seconds of computer time (IBM 168) were needed in order to obtain the results of Table II.

APPENDIX

Using the definition (6) of I_k, one gets by termwise integration of the product $\phi_{2n}(x)\phi_{2m}(x)$,

$$\left\langle\phi_{2n}\left|\frac{(\lambda-\alpha)x^2-\alpha g x^4}{1+gx^2}\right|\phi_{2m}\right\rangle=\sum_{s=0}^{n}\sum_{t=0}^{m}d_{ns}^{(P)}d_{mt}^{(P)}\times\left(\frac{(\lambda-\alpha)}{b}I_{s+t+1}-\frac{\alpha g}{b^2}I_{s+t+2}\right).$$

Using the recursion relation (8), one can write

$$I_{s+t+2}=-\frac{b}{g}\left(I_{s+t+1}-\frac{[2(s+t+1)-1]!!}{2^{s+t+1}}\right).$$

From (5)

$$\left\langle \phi_{2n} \left| -\frac{d^2}{dx^2} + b^2 x^2 \right| \phi_{2m} \right\rangle = b\,\delta_{nm}(4n+1).$$

Finally, one obtains the expression (20) of $\mathscr{H}^{(P)}_{nm}$. Obviously, $\mathscr{H}^{(I)}_{nm}$ is derived in the same way.

[1]A. K. Mitra, J. Math. Phys. **19**, 2018 (1978).
[2]S. K. Kaushal, J. Phys. A: Math. Gen. **12**, L253 (1979).
[3]H. Haken, "Laser Theory," in *Encyclopedia of Physics*, Vol. XXV/2c (Van Nostrand, Princeton, N. J., 1970).
[4]H. Risken and H. D. Vollmer, Z. Phys. **201**, 323 (1967).
[5]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[6]A. Lowan, *Tables of Probability Functions*, Vol. I, N. B. S. Series, New York, 1941.
[7]*Tables of the Error Function and its Derivative.* Applied Math. Series, N. B. S., Washington, D. C., 1954.
[8]J. Kaye, J. Math. Phys. **34**, 119 (1955).
[9]C. Hastings Jr., *Approximation For Digital Computers* (Princeton U. P., New Jersey, 1955).
[10]P. Henrici, *Applied and Computational Complex Analysis*, Vol. II (Wiley, New York, 1977).
[11]N. Bessis, G. Bessis, and G. Hadinger, J. Phys. A: Math. Gen. **13**, 1651 (1980).

Exactly solvable eigenvalue problem with hypergeometric eigenfunctions [a),b)]

Michael J. King and Fritz Rohrlich
Department of Physics, Syracuse University, Syracuse, New York 13210

(Received 29 April 1980; accepted for publication 7 August 1980)

A Schrödinger equation with a momentum dependent interaction leads to exact solutions $\psi \in L^2(\mathbf{R}^3, d^3x)$ with radial parts of the wave function which are hypergeometric functions and their appropriate analytic continuations. The normalization integrals are obtained in closed form.

The Schrödinger equation to be discussed below arises in the relativistic two-body problem of a Hamiltonian formulation of dynamics in which the interaction is momentum dependent.[1] The corresponding classical interaction has the form $-\beta^2(\mathbf{x}\cdot\mathbf{p})$, where $\mathbf{x}$ and $\mathbf{p}$ are relative positions and momenta. Its solution is presented here for at least three reasons: The method of solution is somewhat unconventional, the wave functions are physically acceptable despite a singularity on the interval $[0,\infty)$, and the normalization integrals, although apparently not known in the literature, can be obtained in closed form. The whole problem is an instructive application of the theory of hypergeometric functions.

1. THE DIFFERENTIAL EQUATION

We consider the equation

$$[\nabla^2 - \beta^2(\mathbf{x}\cdot\nabla)^2_{\text{ord}}]\psi(\mathbf{x}) = -\eta\psi(\mathbf{x}), \tag{1}$$

where the ordered differential operator is defined by

$$\begin{aligned}(\mathbf{x}\cdot\nabla)^2_{\text{ord}} &= [\tfrac{1}{2}(\mathbf{x}\cdot\nabla + \nabla\cdot\mathbf{x})]^2 \\ &= (\mathbf{x}\mathbf{x}\cdot\nabla)\cdot\nabla + 4\mathbf{x}\cdot\nabla + \tfrac{9}{4}.\end{aligned}$$

The first term can be written as

$$x^2\nabla^2 + L^2 - 2\mathbf{x}\cdot\nabla,$$

where

$$\mathbf{L} = -i\mathbf{x}\times\nabla.$$

The differential equation (1) thus becomes

$$[\nabla^2 - \beta^2(\mathbf{x}^2\nabla^2 + 2\mathbf{x}\cdot\nabla + L^2 + \tfrac{9}{4})]\psi_\eta = -\eta\psi_\eta. \tag{2}$$

This equation separates in spherical coordinates,

$$\psi_\eta(\mathbf{x}) = \frac{1}{r}\sum_{l=0}^{\infty}\sum_{m=-l}^{l} c_l^m Y_l^m(\theta,\varphi) R_{e\eta}(r).$$

The radial functions $R_{l\eta}(r)$ satisfy

$$(1-\beta^2r^2)R''_{l\eta} - 2\beta^2 r R'_{l\eta} + \left(\eta - \tfrac{1}{4}\beta^2 - \frac{l(l+1)}{r^2}\right)R_{l\eta} = 0,$$

which is conveniently written in terms of the dimensionless variables

$$\rho \equiv \beta r, \quad \nu(\nu+1) = -\tfrac{1}{4} + \eta/\beta^2, \tag{3}$$

$$P_{l\nu}(\rho) = R_{l\eta}(r)/\sqrt{\beta},$$

in the form

$$\frac{d}{d\rho}\left((1-\rho^2)\frac{d}{d\rho}P_{l\nu}\right) + \left(\nu(\nu+1) - \frac{l(l+1)}{\rho^2}\right)P_{l\nu} = 0.$$

For small ρ the centrifugal term will dominate for all $l > 0$. It is therefore convenient to separate the behavior at the origin

$$P_{l\nu}(\rho) = \rho^{l+1} g_{l\nu}(\rho), \tag{4}$$

leading to

$$(1-\rho^2)g''_{l\nu} + 2\left(\frac{l+1}{\rho} - (l+2)\rho\right)g'_{l\nu} + [\nu(\nu+1) - (l+1)(l+2)]g_{l\nu} = 0.$$

The equation reduces to the hypergeometric equation by the substitution

$$t \equiv \rho^2, \quad u_{l\nu}(t) \equiv g_{l\nu}(\rho). \tag{5}$$

One finds

$$t(1-t)u''_{l\nu} + [c - (a+b+1)t]u'_{l\nu} - abu_{l\nu} = 0, \tag{6a}$$

where

$$a \equiv \tfrac{1}{2}(l+\nu)+1, \quad b \equiv \tfrac{1}{2}(l-\nu)+\tfrac{1}{2}, \quad c = a+b. \tag{6b}$$

The square integrability requirement on $R_{l\nu}(r)$,

$$R_{l\nu}(r) \in L^2([0,\infty), dr), \tag{7a}$$

translates via (3), (4), and (5) into

$$u_{l\nu}(t) \in L^2([0,\infty), t^{l+1/2}\,dt). \tag{7b}$$

The problem is to find those values of ν for which the solutions $u_{l\nu}$ of (6a) satisfy (7b).

2. THE EIGENVALUES

The hypergeometric differential equation (6a) has a solution which is analytic in the complex t plane cut along the real axis from $t = 1$ to $t = +\infty$. At $t = 1$ it has a logarithmic singularity. This solution consists of the hypergeometric function

$$u_{l\nu}(t)\Big|_{t<1} = F(a,b;a+b;t), \tag{8}$$

which is analytic in $|t| < 1$, and its analytic continuation $G_{l\nu}$ into the rest of the cut plane. Since the singularity at $t = 1$ is only logarithmic

[a)]Parts of this work are contained in a Ph.D. thesis by the first author (Syracuse University, 1980).
[b)]Supported in part by a grant from the National Science Foundation.

 0022-2488/80/122786-03$1.00

$$u_{l\nu}(t)\Big|_{t<1} \in L^2([0,1], t^{l+1/2}\,dt). \tag{9}$$

The problem of satisfying (7b) consists in finding suitable λ and ν such that $u_{l\nu}(t)|_{t>1}$ defined by

$$u_{l\nu}(t)\Big|_{t>1} = \lim_{\epsilon\to 0}[\lambda G_{l\nu}(t+i\epsilon) + (1-\lambda)G_{l\nu}(t-i\epsilon)], \tag{10}$$

satisfies the condition

$$u_{l\nu}(t)\Big|_{t>1} \in L^2([1,\infty), t^{l+1/2}\,dt). \tag{11}$$

Now the analytic continuation of (8), $G_{l\nu}(t)$, for $|t|>1$ and valid in $|\arg(-t)|<\pi$ is[3]

$$G_{l\nu}(t) = \frac{\Gamma(b+a)\Gamma(b-a)}{\Gamma(b)^2}\frac{1}{(-t)^a}F\left(a,1-b;1-b+a;\frac{1}{t}\right) + \frac{\Gamma(a+b)\Gamma(a-b)}{\Gamma(a)^2}\frac{1}{(-t)^b} \times F\left(b,1-a,1-a+b;\frac{1}{t}\right).$$

From (6b) we see that these two terms require $\nu > -\frac{1}{2}$ and $\nu < -\frac{1}{2}$, respectively, in order to satisfy (11); one of them must be eliminated. Since the solution is symmetric in a and b, it is arbitrary which one survives: the result will be the same. We choose to eliminate the second term. Equation (10) then requires

$$\lim_{\epsilon\to 0}\left(\frac{\lambda}{(-e^{i\epsilon})^b} + \frac{1-\lambda}{[-e^{i(2\pi-\epsilon)}]^b}\right) = 0,$$

which can be satisfied only with $\lambda = \frac{1}{2}$ and

$$\cos\pi b = 0, \quad 2b = -2k+1 \quad (k \text{ integer}).$$

It follows, therefore, from (6b) that (11) will be satisfied when

$$\lambda = \tfrac{1}{2}, \quad \nu = l + 2k \equiv n > 0 \quad (k \text{ integer}). \tag{12}$$

The last inequality results from the above-mentioned requirement $\nu > -\frac{1}{2}$. The eigenvalues ν are thus determined and give with (3),

$$\eta = \beta^2[\nu(\nu+1) + \tfrac{1}{4}] = \beta^2(n+\tfrac{1}{2})^2. \tag{13}$$

3. THE EIGENFUNCTIONS

Substitution of the result (12) into (8) and (10) leads to the eigenfunctions $u_{ln}(t)$ where

$$u_{ln}(t)\Big|_{t<1} = F[\tfrac{1}{2}(l+n)+1; \tfrac{1}{2}(l-n)+\tfrac{1}{2}; l+\tfrac{3}{2}; t], \tag{14a}$$

$$u_{ln}(t)\Big|_{t>1} = \frac{\Gamma(l+\frac{3}{2})\Gamma(-n-\frac{1}{2})}{\Gamma(\frac{1}{2}-k)^2}\frac{1}{(-t)^{l+k+1}} \times F(\tfrac{1}{2}(l+n)+1, \tfrac{1}{2}+k; n+\tfrac{3}{2}; 1/t). \tag{14b}$$

These must now be normalized.

The normalization integral

$$\int_0^\infty R_{ln}^2(r)\,dr = 1 \quad \text{becomes} \quad \tfrac{1}{2}\int_0^\infty U_{ln}^2(t)\,t^{l+1/2}\,dt = 1. \tag{15}$$

These integrals cannot be found in the customary tables. However, they can be expressed in terms of known functions as follows.

The differential equation (6a) can be written in "Sturm–Liouville" form as

$$\frac{d}{dt}\left[t^c(1-t)^{a+b-c+1}\frac{dF}{dt}\right] = abt^{c-1}(1-t)^{a+b-c}F, \tag{16}$$

Here we wrote

$$F \equiv F(a,b;c;t)$$

instead of $u_{l\nu}$ since the following will be valid also when $c = a+b$ is not satisfied.

Let $\bar F \equiv F(\bar a,\bar b;\bar c;t)$ be a solution of (16) when a,b,c are replaced by $\bar a,\bar b,\bar c$. Then one easily derives by integration by parts the identity

$$\int_\alpha^\beta dt\,\bar F F\,[abt^{c-1}(1-t)^{a+b-c} - \bar a\bar b t^{\bar c-1}(1-t)^{\bar a+\bar b-\bar c}]$$
$$= [t^c(1-t)^{a+b-c+1}\bar F F' - t^{\bar c}(1-t)^{\bar a+\bar b-\bar c+1}F\bar F']\Big|_\alpha^\beta$$
$$- \int_\alpha^\beta dt\,\bar F' F'[t^c(1-t)^{a+b-c+1} - t^{\bar c}(1-t)^{\bar a+\bar b-\bar c+1}].$$

If one choses

$$\bar a + \bar b = a + b, \quad \bar c = c$$

the last integral vanishes identically and one finds[2] with $\bar a = a+e$, $\bar b = b-e$

$$e(a-b+e)\int_\alpha^\beta dt\,F(a,b;c;t)F(a+e,b-e;c;t)t^{c-1} \times (1-t)^{a+b-c} = t^c(1-t)^{a+b-c+1}W(a,b;c,e;t)\Big|_\alpha^\beta, \tag{17}$$

where $W(a,b;c,e;t)$ is the Wronskian

$$W(a,b;c,e;t) \equiv \begin{vmatrix} F(a+e,b-e;c,t) & F(a,b;c;t) \\ F'(a+e,b-e;c;t) & F'(a,b;c;t) \end{vmatrix}.$$

A special case of this result is obtained in the limit $e\to 0$. One then finds

$$\lim_{e\to 0}\frac{1}{e}W(a,b;c;e;t) = -F^2(a,b;c;t)\frac{\partial}{\partial t}\left(\frac{\partial}{\partial e}\ln F(a+e,b-e;c;t)\right)_{e=0},$$

so that one has the integral

$$(a-b)\int_\alpha^\beta dt\,F^2(a,b;c;t)t^{c-1}(1-t)^{a+b-c} = -t^c(1-t)^{a+b-c+1}F^2(a,b,c;t) \times \left(\frac{\partial^2}{\partial e\partial t}\ln F(a+e,b-e;c;t)\right)_{e=0}\Big|_\alpha^\beta. \tag{18}$$

The integrals needed for (15) are the special case of this result when $a+b=c$.

For this special case one needs the right-hand side of (18) near $t=1$ as approached from below and from above. From standard expansions[3] one deduces with $2(a-b) = 2n+1$ from (6) and (12)

$$\int_0^1 dt\, t^{l+1/2} u_{ln}^2(t) = \frac{2}{2n+1} B^{-2}(a,b)[\psi'(b) - \psi'(a)], \tag{19a}$$

$$\int_1^\infty dt\, t^{l+1/2} u_{ln}^2(t) = \frac{2}{2n+1} B^{-2}(a,b)[\psi'(a) + \psi'(1-b)], \tag{19b}$$

where $B(a,b)$ is the Gaussian beta function; ψ is the logarithmic derivative of the gamma function. Since

$$\psi'(b) + \psi'(1-b) = \pi^2,$$

for $b = \frac{1}{2} - k$, the sum of the two integrals (19) is

$$\int_0^\infty u_{ln}^2(t) t^{l+1/2}\, dt = \frac{2\pi^2}{2n+1} B^{-2}(a,b). \tag{20}$$

The correctly normalized wave function satisfying (14) is therefore

$$U_{ln}(t) = (1/\pi)\sqrt{2n+1}\, B[\tfrac{1}{2}(l+n)+1, \tfrac{1}{2}(l-n)+\tfrac{1}{2}] \times u_{ln}(t), \tag{21}$$

with $u_{ln}(t)$ given by (14). This can also be written as

$$U_{ln}(t)\Big|_{t<1} = (1/\pi)\sqrt{2n+1}\, B[\tfrac{1}{2}(l+n)+1, \tfrac{1}{2}(l-n)+\tfrac{1}{2}] \times F[\tfrac{1}{2}(l+n)+1, \tfrac{1}{2}(l-n)+\tfrac{1}{2}; l+\tfrac{3}{2}; t], \tag{22a}$$

$$U_{ln}(t)\Big|_{t>1} = (1/\pi)\sqrt{2n+1}\, B[\tfrac{1}{2}(l+n)+1, -n-\tfrac{1}{2}] \times \frac{1}{(-t)^{(1/2)(l+n)+1}} \times F\left[\tfrac{1}{2}(l+n)+1, \tfrac{1}{2}(n-l)+\tfrac{1}{2}; n+\tfrac{3}{2}; \frac{1}{t}\right]. \tag{22b}$$

The normalized radial functions $R_{ln}(r)$ are expressed in terms of these functions by

$$R_{ln}(r) = \sqrt{\beta}\,(\beta r)^{l+1} U_{ln}(\beta^2 r^2). \tag{23}$$

4. DISCUSSION

Our results can be summarized by saying that the Schrödinger equation (2) has the eigenvalues η given by (13) and the eigenfunctions

$$\psi_{nl}^m(\mathbf{x}) = \frac{1}{r} R_{ln}(r) Y_l^m(\theta,\varphi).$$

The R_{ln} are given in terms of hypergeometric functions by Eqs. (22) and (23). The energy eigenvalues depend only on the quantum number n which is a non-negative integer. The angular momentum quantum number can take on all values $l = n - 2k \geqslant 0$ where k is an integer.

However, there are various additional points of interest. For distances r such that $\beta r > 1$ the eigenfunctions must be defined as averages of the two edges of a branch cut from $\beta r = 1$ to $+\infty$. Furthermore, the eigenfunctions have a logarithmic singularity at $\beta r = 1$. This does not cause difficulties in physical interpretation because the probability is well defined over any finite interval Δr including the point $\beta r = 1$.

Furthermore, the normalization problem led to the derivation of a new class of integrals over two hypergeometric functions. These are given by (17) for finite e and by (18) for the limit $e \to 0$.

Of some interest is also the special case $l = 0$. For that case the equation for the radial part of the wave function reduces to

$$\frac{d}{d\rho}\left((1-\rho^2)\frac{dP_{0n}}{d\rho}\right) + n(n+1)P_{0n} = 0,$$

which is Legendre's equation, and n must be an *even* integer according to (12). The solutions are now given by the Legendre functions of the second kind $Q_n(\rho)$. More precisely one obtains from (22) and (23) the result

$$R_{0n}(r) = (2/\pi)\sqrt{\beta}\,\sqrt{2n+1}\, Q_n(\beta r), \tag{24}$$

using the known relations[3] between the Legendre functions and the hypergeometric functions. The normalization of (24) can be checked from known integrals.[4]

After completion of this work our attention was drawn to a paper which deals with an interaction that is the same as ours when a certain parameter is suitably chosen.[5] However, their solution (for that special case) differs from ours because they restrict the domain of t to [0,1] while we have no such restriction. Their quantization condition is therefore also different from ours.

[1] M. King and F. Rohrlich, Phys. Rev. Lett. **44**, 621 (1980); a more detailed discussion of the relativistic Hamiltonian dynamics of which this is a very special example is contained in M. King and F. Rohrlich, "Relativistic Hamiltonian Dynamics II: Momentum Dependent Interactions, Confinement, and Quantization", Ann. Physics (to be published).

[2] This result for the special case of negative integer a and $a + e$ was apparently first obtained by I. I. Hirschman, Jr., Proc. Am. Math. Soc. **8**, 286 (1957).

[3] A. Erdelyi *et al.*, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. I; *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (U. S. National Bureau of Standards, Washing, D.C., 1964).

[4] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1965). We note that formula 7.113.2 in this reference is incorrect and does not agree with 7.112.4 in the limit $\nu = \sigma$.

[5] M. Lakshmanan and K. Eswaran, J. Phys. A **8**, 1658 (1975).

K-surfaces in the Schwarzschild space-time and the construction of lattice cosmologies [a)]

Dieter R. Brill, John M. Cavallo, James A. Isenberg [b)]
Department of Physics, University of Maryland, College Park, Maryland 20742

(Received 7 August 1979; accepted for publication 10 December 1979)

We investigate spacelike spherically symmetric hypersurfaces of constant mean curvature K (which we call K-surfaces) in spherically symmetric static spacetimes. We obtain the differential equation satisfied by these surfaces from a variational principle. The spacetime Killing vector leads to a first integral in the form of a conservation of energy for a particle moving in an effective potential. An embedding of the K-surfaces' intrinsic geometry in flat space likewise follows from an effective potential motion. We apply the formalism to the Schwarzschild solution, and display results of numerical integrations for a variety of K-surfaces and their flat space embeddings. We use these to construct "lattice" cosmological models, and obtain a foliation of K-surfaces of such models with large scale behavior of both the open and closed Friedmann type.

I. INTRODUCTION

Hypersurfaces of constant mean curvature K (called K-surfaces in this paper) have long been considered interesting objects in studying the dynamics of space-time.[1] However, there is a dearth of explicit examples of families of nontrivial K-surfaces in inhomogeneous space-times.

Given a single K-surface, one can obtain a local K-surface foliation by solving an elliptic equation on the lapse function. This approach has been exploited by Estabrook *et al.*[2] for maximal surfaces ($K = 0$), and more recently by Eardley and Smarr[3] for surfaces of $K \neq 0$.

Alternatively, it is well known that each K-surface separately satisfies a variational principle. In this paper we use this principle as a computational tool to find directly the family of spherically symmetric K-surfaces in any spherically symmetric static space-time. We discuss in detail the behavior of these surfaces in Schwarzschild–Kruskal space-time.

The behavior of K-surfaces ($K \neq 0$) in Schwarzschild–Kruskal spacetime differs from that of maximal surfaces both in the asymptotic and in the inner regions. Asymptotically, the K-surfaces become null and go to null infinity $\mathscr{I}$, whereas maximal surfaces—like the familiar $t = \text{const}$ slices of Schwarzschild space-time—go to spacelike infinity i_0. In the interior region, regular spherically symmetric maximal slices do not exist[2] in the region $r < 1.5m$, whereas regular K-surfaces can approach the singularity at $r = 0$ arbitrarily closely if $|K|$ is large enough. In fact, we show numerically that the entire spacetime can be foliated by K-surfaces. A K-surface foliation therefore suggests itself as particularly adapted to radiation problems, and for studying the regions of large curvature in a more general asymptotically flat space-time.

The importance of K-slice foliations in cosmological space-times is well motivated by the Friedmann example and by the fact that, even locally, closed spacetimes cannot be foliated by maximal slices. Thus, the K-slices of Schwarzschild are useful if one wants to patch the Schwarzschild space-time onto a closed cosmological model. We discuss this problem in the context of building "lattice cosmological models."

II. VARIATIONAL PRINCIPLE FOR $K = \text{CONSTANT}$ SURFACES

To obtain the variational principle for K-surfaces in space-time we generalize another well-known extremum property: In Euclidean space the spheres have constant mean curvature, and they have the least surface area for a fixed enclosed volume. Similarly, in space-time we extremize the three-dimensional area $A(S)$ of the hypersurface S, holding constant the 4-volume $V(S, S_1)$ enclosed by S together with any fixed surface S_1. We use a Lagrange multiplier to include this constraint in the variational principle and obtain[4]

$$\delta I = 0, \tag{1a}$$

with

$$I = A(S) + \lambda(S,S_1) = \int_S n^\mu \, d^3S_\mu + \lambda \int_V d^4V. \tag{1b}$$

Here S is any (finite) achronal surface subject to variation with fixed boundary. S_1 is a fixed hypersurface, homotopic to S, with identical boundary, and n^μ is a field of unit vectors normal to S and S_1. $A(S)$ is the three-dimensional area of S, and $V(S, S_1)$ is the four-dimensional volume bounded by S and S_1. The arbitrariness of S_1 corresponds to an arbitrary additive constant in I. Alternatively, one can make the unique choice $S_1 = H^-(S)$, the past horizon of S. The variational principle (1) then agrees with that of Goddard.[5]

To show that Eq. (1) leads to K-surfaces, note that the boundary of V is $S - S_1$, and use the divergence theorem to rewrite I as an integral over V

$$\begin{aligned} I &= A(S) - A(S_1) + \lambda V(S, S_1) + A(S_1) \\ &= \int_V [n^\mu{}_{,\mu} + \lambda] \, d^4V + A(S_1). \end{aligned} \tag{2}$$

[a)]Supported in part by the National Science Foundation under grants PHY-7906940 and PHY-7909281

[b)]Present address: Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1.

 0022-2488/80/122789-08$1.00

The variation of this expression can vanish for arbitrary variations of S only if the integrand vanishes everywhere on S. However, on S, the divergence of the unit normal is related to the mean curvature K. Hence, we have

$$n^{\mu}{}_{;\mu} + \lambda = -K + \lambda = 0. \tag{3}$$

This shows that K is constant and its value is just the Lagrange multiplier λ.

III. APPLICATION TO SPHERICALLY SYMMETRIC STATIC SPACE-TIME

The variational principle (1) provides a convenient way to derive the equation for K-surfaces in spherically symmetric static space-times, and particularly to find first integrals of this equation. In suitable coordinates the metric takes the form

$$ds^2 = -B(r)dt^2 + C(r)dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{4}$$

Let the spacelike surface S be described by $t = t(r, \theta, \phi)$, and choose for S_1 the surface $t = 0$. The variational principle then becomes

$$\delta \int [\Sigma + \lambda (B(r)C(r))^{1/2} t(r)] r^2 \sin\theta\, dr\, d\theta\, d\phi = 0, \tag{5}$$

where

$$\Sigma^2 := -Bt_r^2 - (BC/r^2)[t_\theta^2 + (t_\phi/\sin\theta)^2] + C \tag{6}$$

is positive for a spacelike surface, and subscripts denote partial derivatives. The variational equation obtained from Eq. (5) by varying t is

$$-\lambda r^2 (BC)^{1/2} = -(Br^2 t_r/\Sigma)_r + (BC\sin\theta t_\theta/\Sigma)_\theta/\sin\theta + (BCt_\phi/\Sigma)_\phi/\sin^2\theta. \tag{7}$$

To simplify this equation we restrict attention to spherically symmetric hypersurfaces $t = t(r)$. The resulting ordinary differential equation can be integrated once with respect to r, and then solved for the "rate of change of proper time with proper distance" $dt^*/dr^* = (B/C)^{1/2} t_r$:

$$(dt^*/dr^*)^2 = (H - J)^2/[(H - J)^2 + Br^4], \tag{8}$$

where

$$J := \lambda \int^r [B(u)C(u)]^{1/2} u^2\, du \quad \text{(indefinite integral)},$$

and H is a constant of integration.

If B and C are negative ("inside the horizon"), the spherically symmetric spacelike surface S is more appropriately described by $r = r(t)$, and the variational principle (1) takes the Lagrangian form (with $\dot r =: dr/dt$)

$$0 = \delta \int L\, dt = \delta \int [r^2(-B + C\dot r^2)^{1/2} - J]\, dt. \tag{9}$$

Here we can also obtain a first integral: because L is time independent, the Hamiltonian $H := \dot r(\partial L/\partial \dot r) - L$ is conserved:

$$H = Br^2(-B + C\dot r^2)^{-1/2} + J = \text{const}. \tag{10}$$

The solution of Eq. (10) for dr^*/dt^*, i.e.,

$$(dr^*/dt^*)^2 - Br^4(H - J)^{-2} = 1, \tag{11}$$

is equivalent to Eq. (8). Thus, either Eq. (8) or Eq. (11) can be used to solve for K-surfaces, both inside and outside the horizon [except in special cases like $t(r) = \text{const}$, where the inverse $r(t)$ does not exist]. Equation (11) is particularly useful for a qualitative discussion, because it is analogous to the energy conservation law for a particle of unit total energy moving in a potential given by the second term on the left.

For given H, K, the surfaces differ only by isometry. Therefore the intrinsic metric γ and extrinsic curvature $K = -\frac{1}{2}\mathcal{L}_n \gamma$ are uniquely determined:

$$\gamma_{ij} dx^i dx^j = \frac{BCr^4}{(H - J)^2 + Br^4} dr^2 + r^2 d\Omega^2, \tag{12}$$

$$K_r{}^r = K + 2(H - J)r^{-3}(BC)^{-1/2},$$

$$K_\theta{}^\theta = K_\phi{}^\phi = -(H - J)r^{-3}(BC)^{-1/2}. \tag{13}$$

IV. ISOMETRIC EMBEDDINGS

The intrinsic geometry of two-dimensional Riemannian spaces can often be visualized by an isometric embedding in three-dimensional flat space. The embedding condition is that the geometry inherited by the surface as a subspace of flat space be the same as the surface's given intrinsic geometry. In general, the embedding is local only and cannot be extended to the whole two-dimensional space.

To visualize a spherically symmetric three-dimensional surface by means of an embedding it is customary to suppress one of the angle variables and embed the resulting "reduced" two-dimensional spacelike surface. For the $K = \text{const}$ surfaces of interest here we find that the positively curved portions of the reduced surface can be embedded as a rotationally symmetric surface in three-dimensional Euclidean space. The negatively curved portions can be similarly embedded in three-dimensional Minkowski space.

The intrinsic geometry of the spacelike K-surfaces can be read off from Eq. (4), and by setting $d\phi = 0$ we find the metric of the corresponding reduced surface

$$d\sigma^2 = [C - B(dt/dr)^2]dr^2 + r^2 d\theta^2. \tag{14}$$

Let the metric of flat Euclidean or Minkowski space be written in cylindrical coordinates z, r, ϕ:

$$dl^2 = \pm dz^2 + dr^2 + r^2 d\theta^2. \tag{15}$$

Let $z(r)$ describe an axially symmetric surface, and on this surface equate dl^2 and $d\sigma^2$ to find the embedding condition

$$\pm (dz/dr)^2 + 1 = -B(dt/dr)^2 + C. \tag{16}$$

To solve explicitly for $z(r)$ we substitute $(dt/dr)^2$ from Eq. (8); the problem then reduces to a quadrature.

Like Eq. (11), the isometric embedding equation can also be written as an "energy conservation law" for potential motion

$$\pm\left(\frac{dr}{dz}\right)^2 + \frac{r^4 BC}{r^4 B(C - 1) - (H - J)^2} + 1 = 0, \tag{17}$$

a form which is useful for a qualitative analysis of the reduced surface.

V. EXAMPLES OF SPHERICALLY SYMMETRIC K-SURFACES

In all asymptotically flat space-times, the spherically symmetric K-surfaces which are not confined by horizons

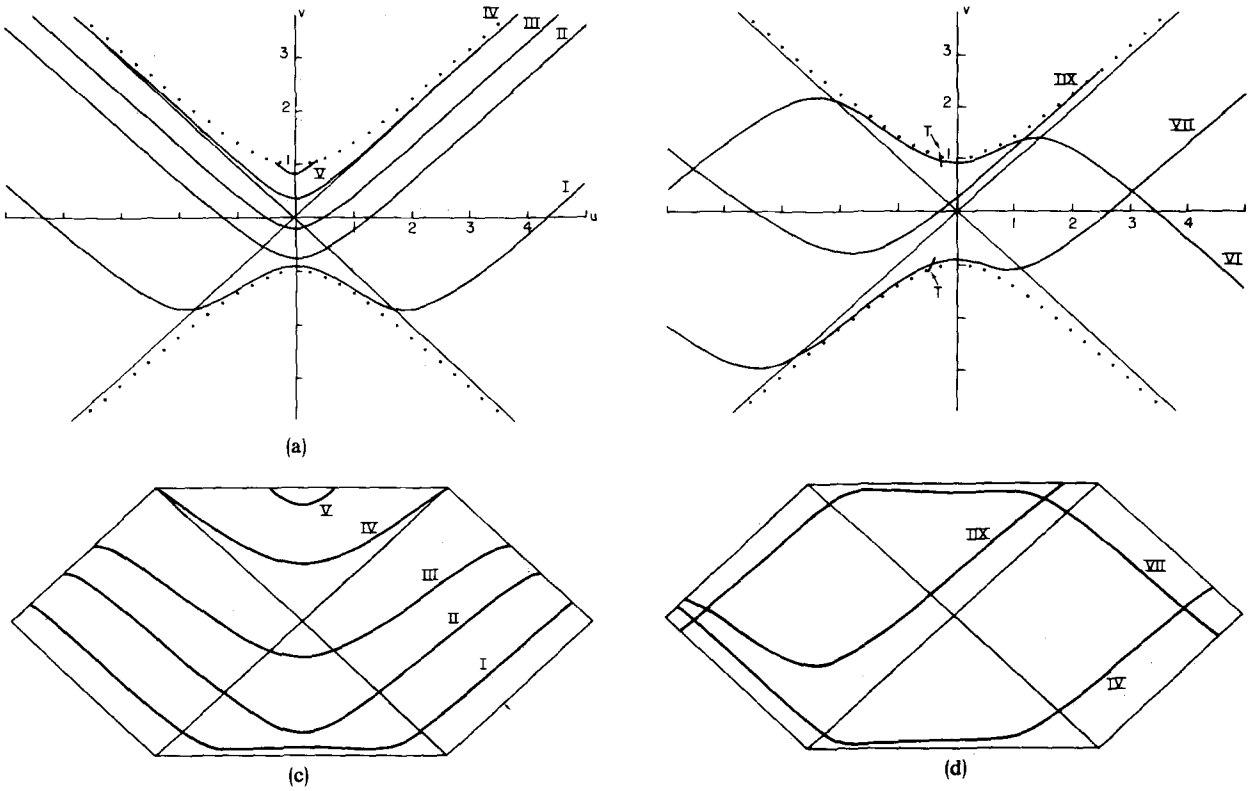


FIG. 1. (a) A family of K-surfaces with $K = 2$ and various values of H in a Kruskal diagram. The throat of all surfaces is on the v axis, and the value of H for each surface is given in the table below. The surfaces that do not reach the sigularity at $r = 0$ have H values in the range $H_- = -1/6 \leqslant H \leqslant H_+ = 0.77871$. (b) Various K-surfaces which are not reflection symmetric about the v axis. Surfaces VI and VII are reflection symmetric about the axes shown. Each axis intersects its surfaces at its throat, as defined by its intrinsic geometry. Surface IIX has no reflection symmetry and terminates at the singularity. (c), (d) Penrose diagrams of these surfaces. The transformation to the coordinates u', v' used here, as well as in Fig. 4 and 6, is given by $\begin{matrix} u' \\ v' \end{matrix} = \frac{1}{2}[\tan^{-1}(u+v) \pm \tan^{-1}(u-v)]$. The slope of these surfaces near $\mathscr{I}^+$ is determined by the value of K. They dip in the interior to avoid the singularity.

Surface #	I	II	III	IV	V	VI	VII	IIX
$K =$	2	2	2	2	2	-2	2	1.2
$H =$	-0.166667	$-1/12$	1/2	H_+	1/2	0.166667	-0.166667	3/4

have a common asymptotic form—that of K-surfaces in Minkowski space-time. In Minkowski space-time the only everywhere smooth, spherically symmetric K-surfaces correspond to the integration constant's value $H = 0$, and they are the familiar constant-interval hyperboloids (analogs of spheres in Euclidean geometry), which become lightlike at $r \to \infty$. If $H \neq 0$, the surfaces become lightlike also at the (space) origin $r = 0$, and hence they are singular there. They can be obtained by numerical integration of Eq. (8) with $B = C = 1$ (even if $K = 0$ the solution is an elliptic integral).

Our main application concerns the Schwarzschild solution. Here we find (Fig. 1) a larger variety of K-surfaces, depending on the value of H. The regular surfaces correspond to a limited range of H values $H_- \leqslant H \leqslant H_+$; for values outside this range the surfaces hit one of the singularities at $r = 0$. (Contrary to the situation in Minkowski space-time, there is a curvature singularity at $r = 0$ in the Schwarzschild space-time. Hence, surfaces which are "irregular at $r = 0$" are of some interest in the latter space-time). Whereas in Minkowski space-time all the regular surfaces are homogeneous (invariant under Lorentz transformations), only some of the surfaces in Schwarzschild space-time—namely those with $H = H_+$ and $H = H_-$— are homogeneous (invariant under t-translation and space rotations). However, all the regular surfaces have an inversion symmetry about their "throat," characterized in the simplest cases by symmetry across the v axis in the Kruskal diagram.

Although numerical integration is necessary to obtain these surfaces explicitly, some of their properties can be qualitatively understood: The Schwarzschild version of Eq. (11), obtained by setting $B = 1/C = 1 - 1/r$,

$$(dr^*/dt^*)^2 = (1 - 1/r)^{-2}(dr/dt)^2 = 1 + r^3(r-1)(H - \tfrac{1}{3}Kr^3)^{-2} \tag{18}$$

or

$$(dr^*/dt^*)^2 + V(r) = 1$$

can be considered as the energy equation for a particle with total energy unity, in an effective potential $V(r) = -r^3(r-1)(H - \frac{1}{3}Kr^3)^{-2}$ (Fig. 2). Here we have set $2m = 1$ without loss of generality, since this amounts to using new dimensionless variables $r/2m$, $t/2m$, $H/4m^2$, and $2mK$. For example, the $r = \text{const}$ K-surfaces mentioned above correspond to "unstable equilibria" of the potential: $V(H, K, r) = 1$ and $dV/dr(H, K, r) = 0$. Now for $r \geqslant 1$, V is negative definite, so there are no corresponding $r = \text{const}$ K-surfaces. However, for any $r < 1$,

$$K^2 = (4/r^3)(r - \tfrac{3}{4})^2/(1-r), \quad H^2 = (r^3/9)(r - \tfrac{3}{2})^2/(1-r), \tag{19}$$

gives us such a surface.[6] K-surfaces which do not satisfy Eq.

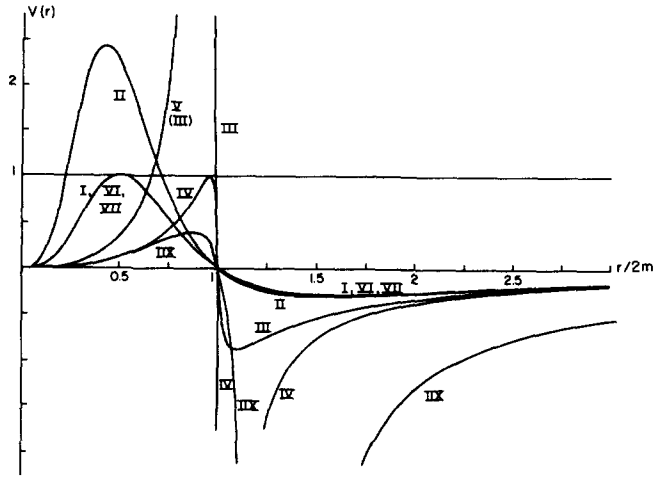


FIG. 2. Plot of effective potential $V(r) = -r^3(r-1)(H-\frac{1}{3}Kr^3)^{-2}$ for the curves shown in Fig. 1.

(19) may or may not contain a "turning point" at which $V = 1$ [and therefore $(dr^*/dt^*)^2 = 0$]. Surfaces with no such point necessarily reach the singularity.

Because of the asymptotic flatness of Schwarzschild, those K-surfaces which escape to r-infinity asymptotically resemble the hyperboloids in Minkowski space. We note that since $V(r) < 0$ for all $r > 1$, if a given surface reaches $r > 1$, it necessarily reaches infinity in r. For large r we have $V(r) \to -9/K^2r^2$, and hence the solution of Eq. (18) becomes the hyperboloid

$$r^2 = t^2 - (3/K^2). \tag{20}$$

The surfaces plotted in Fig. 1 were obtained by numerical integration of Eq. (18), or of the equivalent differential equation in terms of the Kruskal coordinates

$$u = (r-1)^{1/2}\, e^{r/2}\cosh(t/2), \quad v = (r-1)^{1/2}\, e^{r/2}\sinh(t/2),$$

which takes the form

$$\frac{dv}{du} = \frac{Av + Eu}{Au + Ev} \quad \text{with } E := H - \tfrac{1}{3}Kr^3,$$
$$A^2 := E^2 + r^3(r-1). \tag{21}$$

To fix the sign of A we demand that K be the divergence of the future pointing normal (or convergence of past pointing normal). Surfaces of positive K are then concave up in the asymptotic regions $t = +(r^2 + (3/K)^2)^{1/2}$. This rule about the sign of A leads to a smooth surface through the turning point (where $A^2 = 0$), and implies that A switches sign at this point.

A particular K-surface is specified, and hence can be integrated by computer, if we give H, K, one "point" (r, t) on the surface, and the sign of A. [This latter choice specifies on which side of the throat of this K-surface the point (r, t) will be.] We can regard r as a parameter within the surface; therefore, the spherically symmetric K-surfaces in Schwarzschild space-time form a three-parameter family. One parameter is the surface's constant mean curvature K itself, another the t-translation from some fiducial surface (e.g., that surface, of the same K and H, whose throat occurs at $u = 0$). The third parameter H measures how much the intrinsic and extrinsic curvature varies on the K-surface. This is shown by Eq. (13) which becomes in the Schwarzschild case (or for any spherically symmetric metric with $BC = 1$)

$$K_r{}^r = K/3 + 2H/r^3, \quad K_\theta{}^\theta = K_\phi{}^\phi = K/3 - H/r^3. \tag{22}$$

When $H = 0$, the extrinsic curvature tensor has the covariant constant, isotropic, "pure trace" form $K_{ab} = \frac{1}{3}Kg_{ab}$. Similarly, when $H = 0$ the Ricci tensor has the simple form

$$R_r{}^r = -2m/r^3 - 2K^2/9,$$
$$R_\theta{}^\theta = R_\phi{}^\phi = m/r^3 - 2K^2/9, \tag{23}$$

i.e., the Schwarzschild maximal slice value supplemented by the isotropic, covariant constant tensor $-(2K^2/9)g_{ab}$. The intrinsic metric is given by Eq. (12):

$$\gamma_{ij}dx^i dx^j = (1 - 2m/r + K^2r^2/9)^{-1}dr^2 + r^2 d\Omega^2. \tag{24}$$

(This is the unique family of spherically symmetric three-dimensional metrics with constant scalar curvature $R = 2K^2/3$, which also occurs as the geometry of the maximal, $t =$ const slice of the "Schwarzschild solution" with nonvanishing cosmological constant $\Lambda = K^2/3$). In the general case, when $H \neq 0$, we can consider the value H to be a measure of the deviation from this "homogeneous and isotropic" behavior of the curvatures.

For each fixed value of K there exist values H_+ and H_- such that all surfaces with $H < H_-$ or $H > H_+$ contain one and only one singularity, while those with $H_- < H < H_+$ contain either two singularities or none at all. The $H = H_+$ or $H = H_-$ surfaces are the homogeneous, $r =$ const sufaces of Eq. (19) (see also Fig. 1). From the explicit form of A^2 of Eq. (21), we see that the constant values $r_\pm$ corresponding to $H_\pm$ must satisfy $0 < r_- < .75 < r_+ < 1$. Further, we find that for $K > 0$ all nonsingular surfaces ($H_- < H < H_+$) which intersect the region $u > 0$ have minimum r greater that r_+, while the nonsingular, $K > 0$, surfaces intersecting $u < 0$ have minimum r greater than r_-. The maximum values for r on the double-singular surfaces obey similar inequalities.[6]

For any given K-surface, there is a corresponding embedding which may be obtained by integrating Eq. (17) evaluated for the Schwarzschild case

$$\pm(dr/dz)^2 = 1 - r^4/[r^3 - (H - \tfrac{1}{3}Kr^3)^2]. \tag{25}$$

The appropriate sign for $(dr/dz)^2$ is determined by the sign of the right-hand side of Eq. (25). This also determines whether the embedding is Euclidean or Minkowskian. Note that it often happens that part of a given surface requires Euclidean embedding while the rest requires Minkowskian embedding. This happens with some of the surfaces of Fig. 1. In Fig. 3, we give the embeddings of the surfaces in Fig. 1. All are obtained numerically. However, Eq. (25) is simple enough to permit a qualitative analysis by the "particle in a potential" analogy similar to the above analysis of Eqs. (18) and (21). We omit the details.

Can one foliate the Schwarzschild space-time using K-surfaces? Previous experience with attempts to foliate with maximal slices indicates that while the region outside the horizon is easily filled (indeed, one may use a set of surfaces that all have the same H and K), filling the inner region is more difficult.

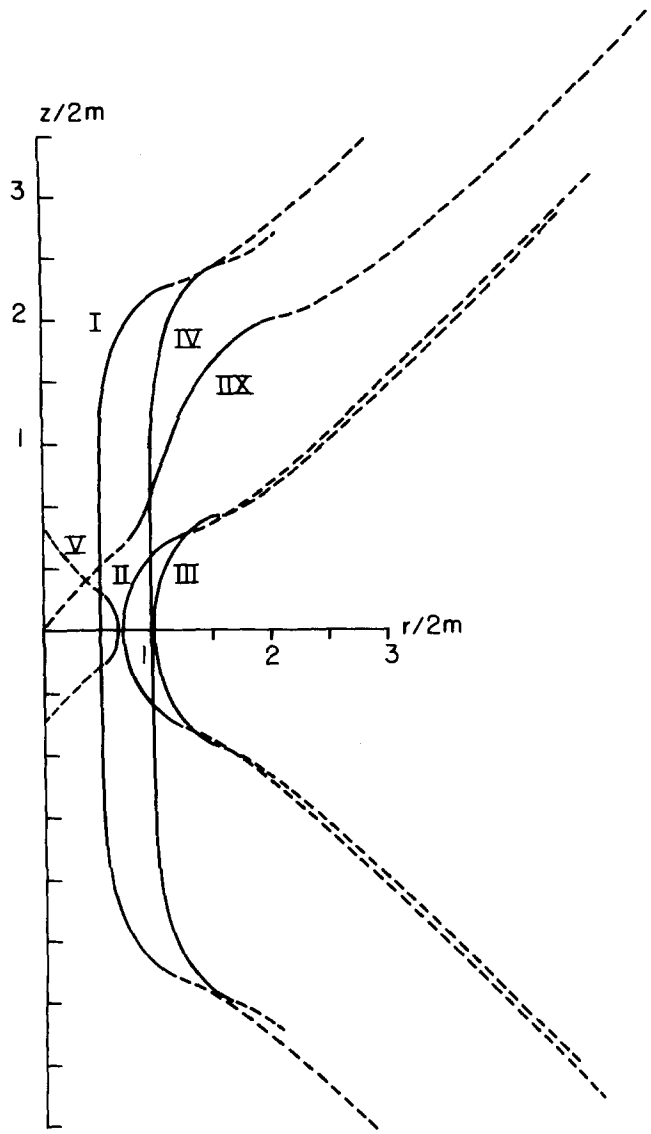


FIG. 3. Embeddings of the K-surfaces in Fig. 1. Solid curves show Euclidean embeddings, dotted curves show Minkowskian embeddings. Curves need to be rotated about the z axis to generate the embedded surfaces. Note that surfaces VI and VII of Fig. 1(b) have the same embedding as surface I of Fig. 1(a).

It has been conjectured, however, that if the full set of values of K is used, then a foliation can be achieved. Numerical support for this conjecture is provided by the family of slices sketched in Fig. 4. For simplicity, all slices were chosen with the throat at $u = 0$. Note that the foliation must consist entirely of $r = \text{const}$ surfaces for $K \leqslant 0$ (see figure caption). We thank the referee for pointing out that the existence of such foliations can, in fact, be proved on the basis of the results of Eardley and Smarr.[3] They show that *any* K-surface foliation of the exterior (not only spherically symmetric) with constant K approaches the corresponding $r = r_+$ surface in the limit. It can therefore be extended to a complete foliation, as above, by $r = \text{const}$ surfaces down to $r = 0$.

VI. CONSTRUCTION OF LATTICE UNIVERSES USING K-SLICES

Whereas K-surfaces in asymptotically flat space-times are relatively unfamiliar, such surfaces have always been used in relativistic cosmology, typically as the spaces seen by comoving observers. An interesting connection between these two realms is provided by "lattice universes." The K-surface foliations of Schwarzschild space-times prove to be quite useful for the construction of such models. These slices provide the original Lindquist–Wheeler[7] lattice universes with a smooth foliation of constant "extrinsic time," which corresponds closely to the Friedmann comoving time. In addition, a larger class of models can be built from K-slices of Schwarzschild than from maximal ($K = 0$) slices.

Lattice universes are space-times consisting of a number of Schwarzschild regions (appropriately truncated at some finite, time-dependent radius) which are patched together as closely as possible—so that the violation of the Israel matching condition[8] is minimized. Explicitly, Lindquist and Wheeler built their version by (a) choosing a (finite spherical) maximal hypersurface S from Schwarzschild, (b) fitting together N copies of S (with a "comparison hypersphere" serving as a template for the fitting) to form an initial

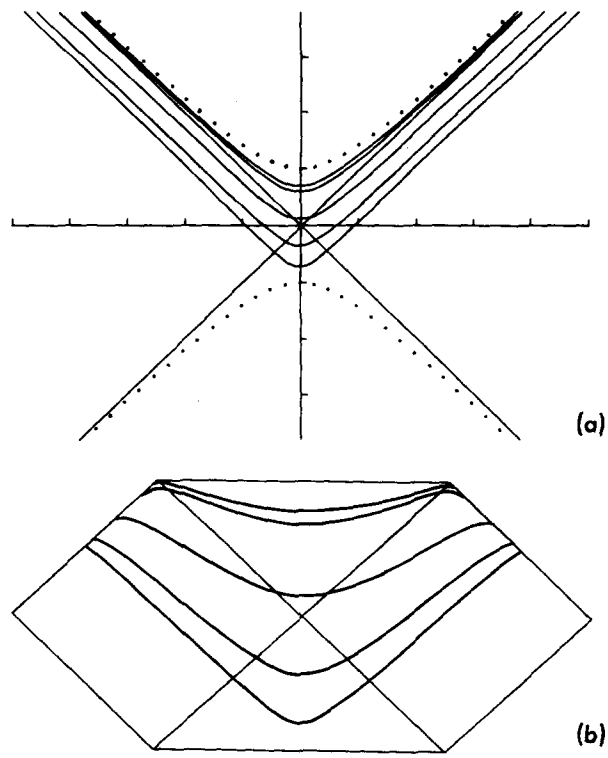


FIG. 4. Foliation of Schwarzschild–Kruskal spacetime by K-surfaces in Kruskal (a) and Penrose (b) diagrams. Only a few typical surfaces are shown. Whereas there is some arbitrariness (e.g., in the location of the throat) for the surfaces in the past of $r = 1.5m$, all surfaces in the future of $r = 1.5m$ must be of the $r = \text{const}$ type. Namely, since the future $r = 0$ singularity corresponds to collapse (converging normals), surfaces with $K < 0$ must lie to the future of surface with $K > 0$. However from the concavity of K-surfaces in the asymptotic region as discussed in the text, one knows that K-surfaces which emerge from the horizon reach $\mathscr{I}^+$ if $K > 0$, and $\mathscr{I}^-$ if $K < 0$. Hence, in a foliation, K-surfaces of both signs cannot reach null infinity.

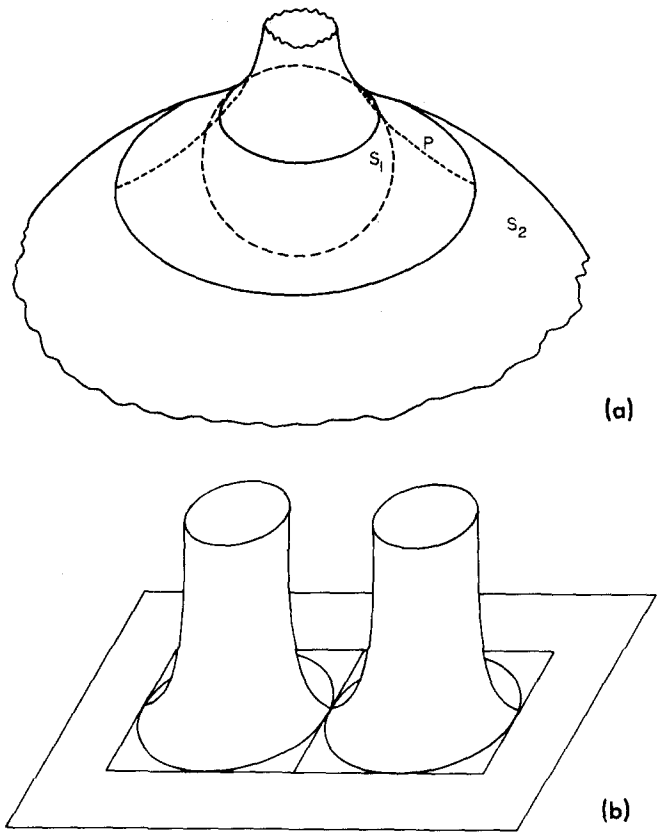


FIG. 5. Embedding diagrams of lattice universes. Figure 5(a) shows the typical paraboloid of revolution, generated by the parabola P, which has the intrinsic geometry of the equatorial plane of a maximal (t = const) surface in the Schwarzschild solution. This surface can be matched only to "comparison hyperspheres", of which two examples S_1, S_2 are shown, but not to a "comparison hyperplane". Figure 5(b) shows how some K-surface embeddings do match to a comparison hyperplane H. Two copies of the embedded surface are shown to illustrate adjacent cells of a lattice universe. The curve generating the surface is half of curve I of Fig. 3. All surfaces are shown only up to their throat, and should be continued symmetrically on the top.

surface S^N, and (c) evolving S^N into a space-time by letting the N regions effectively free fall into each other. Such a scheme can never produce an exact solution to Einstein's equations because of the irremovable gaps between the packed spheres. Nevertheless, Lindquist and Wheeler show that a remarkably good approximation can be achieved, one measure of which is the accuracy with which the dynamics follows that of the "comparison Friedmann model."

The Lindquist–Wheeler scheme, based on maximal slices, can only produce lattice universes which approximate the positive curvature (3-sphere) Friedmann models. This follows from the fact that the initial surface S^N of the lattice spacetime constructed from maximal slices is necessarily a surface of maximum expansion (recall that the only Friedmann universes which contain such a maximal slice are the positive curvature models). Figure 5, which shows how the embedding diagrams of the N copies of S patch together, further illustrates this point. We note also that in the Lindquist–Wheeler scheme, only the original "$t = 0$" surface can be smoothly constructed from maximal slices; later and earlier $t_{\text{Schwarzschild}}$ = const surfaces have discontinuous slopes at the cell boundaries.

If, however, we use general K-slices rather than maximal slices only, then we can build lattice space-times which approximate Friedmann cosmologies of every sort—negative curvature and flat versions as well as the positive curvature types. The construction scheme for these generalized lattice space-times is basically the same as that of Lindquist and Wheeler. Note, however, that the nonmaximal K-slices permit us to follow the evolution of the lattice space-time into the future and past, using a preferred, smooth time choice. We now describe how this is done, using as an example a lattice model which approximates a (intrinsically) flat Friedmann cosmology.

Any K-slice S_K in Schwarzschild for which there exists a radius ρ_0 at which the imbedded surface becomes horizontal (i.e., $dz/dr|_{r=\rho_0} = 0$) may be used as the basic cell for the initial surface. Such a K-surface (truncated at $r = \rho_0$) can be matched smoothly onto the "comparison hyperplane" (which is simply flat Euclidean 3-space E^3); we can therefore construct the initial slice for our lattice model by packing copies of S_K (2-spheres packed in E^3). Of course, with equal approximation we could identify cell boundaries and obtain a closed lattice universe with 3-torus topology, containing one or several Schwarzschild masses.

As in Lindquist and Wheeler, the dynamical evolution of the cell boundaries must be radial free fall in order to fulfill the Israel matching conditions[8] at the point of contact of cell boundaries. We therefore find $\rho(t)$ by solving the Schwarzschild radial geodesic equation. Its solution is

$$\frac{t}{m} = \frac{2}{3}\left(\frac{\rho}{2m}\right)^{3/2} + 2\left(\frac{\rho}{2m}\right)^{1/2} - \ln\left[\frac{(\rho/2m)^{1/2}+1}{(\rho/2m)+1}\right] + \tau, \tag{26}$$

where τ is chosen so that $t = t_0$ where $\rho = \rho_0$. [Note that Eq. (26) describes radial free fall for a particle at escape velocity, i.e., which reaches infinity with vanishing speed. This is appropriate *only* for the flat Friedmann example now being discussed.] The relation $r = \rho(t)$ from Eq. (26) of course does not determine a slice, since we still need H and K. One relation between H and K is obtained by demanding that the slices be smooth across the cell boundaries for all times. Since all cells are equivalent and symmetric, this means that the slice is orthogonal to the path $\rho(t)$ of the boundary. This relation can be visualized even more directly as the demand that the embedding of the evolving slice stay "horizontal" at $r = \rho(t)$:

$$2m\rho^3 = (H - \tfrac{1}{3}K\rho^3)^2. \tag{27}$$

We might try to complete the job of determining the evolving slices by demanding that K match that of the appropriate comparison hyperplane. Such a condition is consistent with, but not required by, the Israel junction conditions. (For the present case, these conditions demand that γ_{rr}, $\gamma_{\theta\theta}$, $\gamma_{\phi\phi}$, $K^{\theta}{}_{\theta}$, and $K^{\phi}{}_{\phi}$ all be continuous. However, they permit $K^{r}{}_{r}$, and therefore also K, to be arbitrarily discontinuous.) If we do this, we find $K(t) = 3(2m/\rho(t)^3)^{1/2}$ and $H = 0$ (all time). Thus, the extrinsic curvature is isotropic not only on the boundary (where it matches that of the comparison hyper-

plane), but everywhere. However, these surfaces have the undesirable feature of intersecting each other, and they are not all symmetric across the same throat $u = 0$. A successful alternative is the condition that the surface be, in fact, symmetric across this throat. (This condition is necessary if one wants to join pairs of throats and thereby form "wormholes.") This prescription gives us a K-slice foliation of a part of our lattice model space-time. A computer generated picture of this foliation is given in Fig. 6.

The lattice space-time cannot, however, be completely foliated by this prescription. The reason is similar to that given in the caption of Fig. 4: Near the "bang" K must approach $+\infty$, and near the individual black holes' collapse it must approach $-\infty$; hence, on some surface it must be zero. However, this $K = 0$ maximal surface cannot satisfy the cosmological boundary condition (27), as one can see from the behavior of the effective potential in Eq. (25). (Alternately, one can use the 3-torus identification and note that a spacelike surface of local volume maximum cannot exist in this ever expanding universe.[4]) In fact, $K = 0$ is approached only asymptotically by the surfaces constructed according to the prescription. The surface $r = 1.5m$ is the limit surface of

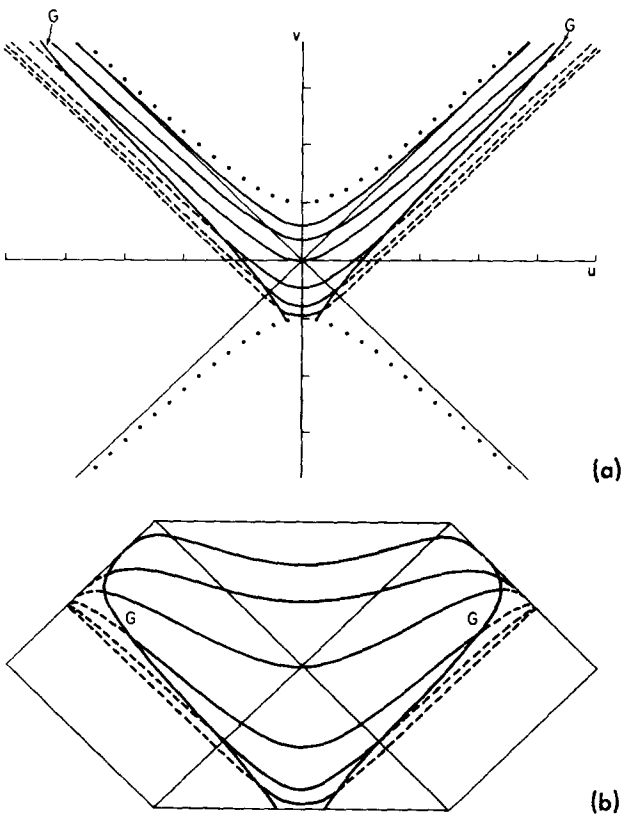


FIG. 6. Foliation of a lattice universe by K-surfaces shown in Kruskal (a) and Penrose (b) diagrams. Since all cells are identical, only one is shown. The edge of the cell is a geodesic G, and it indicates how the universe as a whole is collapsing. Only the interior part of the K-surfaces (shown in solid lines) corresponds to the lattice universe. The foliation was constructed according to the presciption in the text. All surfaces have $K > 0$, with K decreasing in time. The foliation needs to be supplemented by $K = \text{const} < 0$ surfaces in the collapse region near the future singularity. The values of K and H are as follows, starting from the bottom surface:

$K =$	2.3	1.3	0.90	0.60	0.42	0.24
$H =$	−0.1	−0.06	0.18	0.40	0.45	0.40.

infinite expansion. To foliate the remainder of the space-time we can fall back on the $r = \text{const} < 1.5m$ surfaces. These surfaces have one disconnected component of infinite volume for each of the model's black hole masses, and do not connect across the entire universe as the cosmological slices do.

The construction of the other two types of lattice universes, with their preferred foliations, is achieved in a way very similar to that just described. The only changes are that, for the positive curvature Friedmann approximation, we use K-slices for which dz/dr is positive at the boundary and replace Eq. (26) with the radial geodesic infall relation of a particle dropped from rest at some finite $r = R$ (the maximum radius); while for the negative curvature Friedmann approximation we use K-slices for which the embedding is Minkowskian at the boundary, and replace Eq. (26) with the free fall relation for a particle reaching infinity with some nonvanishing speed. In the former case the prescription gives the complete foliation of the space-time,[9] while in the latter case the constructed surfaces again have to be supplemented by disconnected $r = \text{const}$ surfaces. Of course, the packing of the regions is somewhat different for the three types as well.

VII. CONCLUSIONS

We have shown how to construct explicitly K-surfaces in the Schwarzschild solution, and given some applications where these surfaces provide a smoothly matched slicing of space-times which are patched together out of Schwarzschild regions. One can use these slices for numerous other matching problems; for example, they provide a smooth slicing of collapsing spherically symmetric interior solutions. In particular, it is well known[10] that an exterior Schwarzschild solution, truncated in the center at some radial geodesic, can be matched to a section of a dust-filled Friedmann universe ("collapsing dust ball"). The standard description uses the Friedmann homogeneous time coordinate in the interior, and Schwarzschild time in the exterior, so that the $t = \text{const}$ slices, even if continuous, are not smooth. However, the same interior slices will smootly fit onto K-surfaces of Schwarzschild—in the particular case of collapse from infinity they would be the $H = 0$ slices we mentioned above, but their exterior rather than their interior parts. The slicing of matched interior and exterior solutions by K-surfaces for marginally bound collapse has been treated by Eardley and Smarr.[3] They discuss the details of this type of slicing, note some of its disadvantages, and obtain nonhomogeneous interior K-surfaces as well.

We have shown by explicit numerical example that there appears to be no obstacle to foliating by K-surfaces the Schwarzschild space-time, and the various space-times constructed from it by patching or approximate patching. This is of some interest because the theorems[1,3] available today about existence of K-surface foliations always involve some avoidance assumption which is not *a priori* known to be satisfied in specific examples.[11]

The K-surface slicing of the Lindquist–Wheeler lattice universe provides a very simple explicit example of the "simultaneity" of the cosmological bang and "crunch" in a sit-

uation involving processes of very different proper time duration—the collapse of the individual Schwarzschild regions, and the collapse of the universe as a whole. This solution also gives a simple example where one can see the mixmaster oscillations of Belinsky, Khalatnikov, and Lifshitz[12] played out. Our K-slices of a Schwarzschild region, in the limit $K \to \infty$, show the behavior that one would expect in general for a "black hole" region in a cosmological solution, and by which one might recognize, when using such slicing, that one is approaching that region of the bang or crunch to be associated with a black hole: Rather than undergoing the general mixmaster oscillations, the metric on these slices corresponds to the "zero frequency" behavior where transverse distances shrink to zero, and radial distances expand to infinity. No simpler illustration of this behavior can be given than the particular K-surfaces on which $r = \text{const}$: For small r, Eq. (19) becomes $K = \frac{3}{2}r^{-3/2}$. Hence, the metric on the surface, as a function of K-time, is

$$d\sigma^2 = (2K/3)^{2/3}dt^2 + (2K/3)^{-4/3}d\Omega^2.$$

The other regions which one would expect to appear in general, and which would show the more general mixmaster oscillations corresponding to cosmological collapse, as opposed to black hole formation, are of course shrunk to zero by the matching assumptions in the special example of the lattice universe.

The lattice universes which approximate "flat" and "open" Friedmann universes also illustrate the behavior of K-slices when there is local black hole collapse but infinite cosmological expansion. In this case, of course, the collapse regions cannot be viewed simultaneously as one single crunch, because the universe as a whole in fact does not collapse. Here we found that the K-surface foliation initially provides connected Cauchy surfaces for the whole universe. These surfaces foliate the entire region outside the black holes' horizons, and their volume tends to infinity as K approaches zero. However, to complete the foliation, disconnected surfaces are needed, with one component collapsing down on each black hole singularity.

ACKNOWLEDGMENTS

We are pleased to acknowledge the generous support of the University of Maryland Computer Science Center. We are also grateful to D. Eardley, L. Smarr, and A. Qadir for helpful discussions.

[1]A. Lichnerowicz, J. Math. Pure Appl. **23**, 37 (1944) (maximal slices only); J. York, Phys. Rev. Lett. **28**, 1082 (1972) (nonmaximal slices); J. Marsden, and F. Tipler, "Maximal Hypersurfaces and Foliations of Constant Mean Curvature in General Relativity," Preprint, University of California at Berkeley (1978).
[2]F. Estabrook, H. Wahlquist, S. Christensen, B. DeWitt, L. Smarr, and E. Tsiang, Phys. Rev. D **7**, 2814 (1973).
[3]D. Eardley and L. Smarr, Phys. Rev. D **19**, 2239 (1978).
[4]D. Brill and F. Flaherty, Commun. Math. Phys. **50**, 157 (1976); Ann. Inst. Henri Poincaré **28**, 335 (1978).
[5]A. Goddard, "Spacelike Surfaces of Constant Mean Curvature," Ph.D. Dissertation, Oxford University (1975); Comm. Math. Phys. **54**, 279 (1977); Math. Proc. Cambridge Philos. Soc. **82**, 489 (1977); Gen. Rel. Grav. **8**, 525 (1977).
[6]We thank the referee for pointing out that some of these results appear in Eardley and Smarr, Ref. 3.
[7]R. Lindquist and J. Wheeler, Rev. Mod. Phys. **29**, 432 (1957).
[8]S. O'Brian and J. Synge, Commun. Dublin Inst. Adv. Study A **9**, 1 (1952); W. Israel, Nuovo Cimento B **44**, 1 (1966); B **48**, 463 (1967); E. Robson, Ann. Inst. H. Poincaré **16**, 41 (1972).
[9]K-surface foliations of lattice universes which approximate the positive curvature Friedmann model have been studied by A. Qadir and J.A. Wheeler Gehenian Festschr. (to appear).
[10]C. Misner, K. Thorne, and J. Wheeler, *Gravitation* (Freeman, New York, 1973), pp. 852–3.
[11]For space-times in which a crushing function can be found, the necessary avoidance property is guaranteed by the results of Eardley and Smarr, Ref. 3.
[12]V. Belinsky, I. Khalatnikov, and E. Lifshitz, Adv. Phys. **19**, 525 (1970).

Gödel-like cosmological solutions

Dipankar Ray
Department of Physics, Queen Mary College, London, England

(Received 24 August 1979; accepted for publication 14 November 1979)

Gödels cosmological solutions have been generalized by Novello and Reboucas [Astrophys. J. **225**, 719–24 (1978)]. An attempt is made further to generalize their work. A class of solutions is obtained which are Gödel-like in the sense of Novello and Reboucas, but which have singularities at both ends of time.

1. INTRODUCTION

Solutions of Einstein equations for the energy-momentum tensor

$$T_{\mu\nu} = (\rho + p)V_\mu V_\nu - pg_{\mu\nu} + q_\mu V_\nu + q_\nu V_\mu, \tag{1}$$

where p is pressure, ρ is density, V_μ is velocity, and q_μ is heat flux, and the metric

$$ds^2 = dt^2 + 2A(x,t)\,dy\,dt - (m-1)A^2(x)dy^2 - H^2(t)\,dz^2 - F^2(t)\,dx^2, \tag{2}$$

where m is a constant and the coordinate system is co-moving, i.e.,

$$V^\mu = 0, \quad \text{for } \mu \neq 0, \qquad = 1, \quad \text{for } \mu = 0, \tag{3}$$

and

$$(t,x,y,z) = (x^0,x^1,x^2,x^3),$$

have been sought by Novello and Reboucas.[1] They have shown that such solutions have the interesting property that $A(x,t)$ can be expressed as

$$A(x,t) = e^{Cx}A_2(t), \quad C \text{ a constant.} \tag{4}$$

Such solutions have been called Gödel-like by the above authors. For the metric (2) with a co-moving frame and energy-momentum tensor (1) the field equations without any specified equation of state reduce to

$$HF = m^*C, \quad m^* \text{ a constant,} \tag{5a}$$

$$2(m-1)\left[\frac{\ddot{F}}{F} + \frac{\dot{F}\dot{A}_2}{FA_2} - \frac{\dot{F}^2}{F^2}\right] - \frac{(2m-1)C^2}{2F^2} = 0 \tag{5b}$$

$$(m-1)\left[\frac{\ddot{A}_2}{A_2} + \frac{\dot{F}^2}{F^2} - \frac{\dot{A}\dot{F}}{AF} - \frac{\ddot{F}}{F}\right] + (2m-1)\frac{C^2}{F^2} - \frac{2\dot{F}^2}{F^2} = 0. \tag{5c}$$

Equations (5) have been solved by Novello and Reboucas for $F = 1$. This gives $m = \frac{1}{2}, p - \Lambda = \frac{1}{2}C^2, \rho + \Lambda = \frac{1}{2}C^2$, where Λ is the cosmological constant, and

$$A_2(t) = \Theta_0(t) + 1, \quad \text{where } \Theta_0 \text{ is a constant.}$$

In the present note we shall try to generalize these solutions by solving (4) for an equation of state

$$p - \Lambda = \rho + \Lambda. \tag{6}$$

2. SOLUTIONS

It has been shown by Novello and Reboucas[1] that from (1), (2), and (3) one gets the tetrad components as

$$T_1{}^1 = T_2{}^2 = T_3{}^3. \tag{7}$$

From (6), (7), and Einstein equation,

$$R_3{}^3 = 0. \tag{8}$$

From (2), (3), (4), (5), and (8)

$$A_2\dot{F}/F = k, \quad \text{where } k \text{ is a constant.} \tag{9}$$

(9) and (5b) then give

$$m = 1/2. \tag{10}$$

From (9), (10), and (5c) we get

$$\dot{A}_2^2 + 8k^2 \ln A_2 = l^2. \tag{11}$$

Now, if $k = 0$, then we get the solutions by Novello and Reboucas, which need not be discussed here.

If $k \neq 0$, we get from (11)

$$A_2 = e^{(l^2 - u^2)/8k^2}, \tag{12a}$$

where

$$(e^{l^2/8k^2}/4k^2)\int e^{-u^2/8k^2}\,du = t + B, \tag{12b}$$

and from (9) and (5a)

$$F = De^{u/4k},$$

and

$$H = \frac{m^*C}{D}e^{-u/4k}, \tag{13}$$

where B, D, and l are constants.

It can be easily checked that (10), (12), and (13) together satisfy equations (5). They also satisfy (7) and (8) and hence the equation of state (6). Moreover in (12) we note that, although it has not been possible to express u and hence A_2 explicitly in terms of t, the integral in the left-hand side of (12a) is the familiar distribution function of normal distribution whose tables are available.

3. CONCLUSIONS

From (2) we note that $A_2 = 0$ gives $\det g_{\mu\nu} = 0$ and hence a singularity. Therefore from (11) we see that A_2 can range from 0 to $e^{l^2/8k^2}$, i.e., increase from 0 to $e^{l^2/8k^2}$ and then decrease from $e^{l^2/8k^2}$ to 0. From (12a) we note that at $A_2 = 0$, $u = \infty$, and at $A_2 = e^{l^2/8k^2}$, $u = 0$. Therefore, the time taken in going from $A_2 = 0$ to $A_2 = e^{l^2/8k^2}$ is $(e^{l^2/8k^2})/4k^2 \times \int_0^\infty e^{-u^2/8k^2}\,du$, which is finite for $k \neq 0$.

Likewise the time taken in going from $A_2 = e^{l^2/8k^2}$ to

 0022-2488/80/122797-02$1.00

$A_2 = 0$ is also finite, i.e., the system goes from one singularity to another in a finite time.

Thus summarily the solutions of Einstein's equation with a metric (2), energy-momentum tensor (1), and equation of state (6) are either the solutions by Novello and Reboucas or are given by (10), (12), and (13). In the second case the system has singularities at both ends of time.

It can also be noted that, if the equation of state is unspecified, then another class of solutions of (5) can be easily obtained by taking $\dot{F} = 1$. However, that gives the unphysical result, namely $p + \rho < 0$.

ACKNOWLEDGMENT

The author thank Professor W.B. Bonnar of Queen Elizabeth College, London, for useful discussions.

[1]M. Novello, and M.J. Reboucas, Astrophys. J. **225**, 719–24 (1978).

Static gravitational and Maxwell fields in the general scalar tensor theory

A. Banerjee [a)]
Instituto de Física, Ilha do Fundão, U.F.R.J., Rio de Janeiro, Brazil

S. B. Duttachoudhury
Department of Physics, St. Anthony's College, Shillong 793003, India

(Received 30 August 1979; accepted for publication 9 November 1979)

The expression for g_{00} as a function of the scalar field Ψ is obtained in the general scalar tensor theory of gravitation proposed by Nordtvedt and later discussed by Barker, assuming that there exists a functional relationship between them. Exact solutions for a plane symmetric static gravitational field are also obtained in this theory. Further the calculations are extended for the static electrovac with the assumption that here both g_{00} and the scalar field Ψ are functions of the electrostatic potential ϕ, and the results are different from those previously obtained in the corresponding situation of Brans–Dicke theory.

I. INTRODUCTION

Within the framework of the general scalar tensor theory of gravitation (Nordtvedt[1]) one can allow the parameter ω to be an arbitrary function of the scalar field Ψ. Recently Barker[2] proposed a special case of the Nordtvedt's general class of scalar tensor theories where the Newtonian gravitational constant G does not vary with time in the homogeneous cosmological situation and arguments in favor of the this theory were put forward.

It is worthwhile to discuss the static space–time in this theory and one arrives at some new results in this special case of general scalar tensor theory where the exact form of ω as a function of the scalar field Ψ is obtained from the condition that $G = \text{const}$. One can further generalize some of the results of Raychaudhuri and Bandyopadhyaya[3] for a static electrovac in Brans–Dicke theory[4] of gravitation where $\omega = \text{const}$.

In Sec. II we consider a general static space–time and find the exact form of g_{00} as a function of the scalar field Ψ assuming, however, that there exists a functional relationship between them. Such a relation was previously obtained by Banerjee and Bhattacharya[5] in Brans–Dicke (B–D) theory. Further we give here an exact plane symmetric static solution in Nordtvedt's general scalar tensor theory with ω given in Barker's form: $\omega = (4 - 3\Psi)/(2\Psi - 2)$. These solutions are new and reduce to those of Taub[6] in Einstein's theory when the scalar field is absent.

In Sec. III we consider a static electrovac representing an electrostatic field alone in the general scalar tensor theory of gravitation. Assuming that both g_{00} and the scalar field Ψ are functions of the electrostatic potential ϕ we get two relations, one connecting g_{00}, Ψ, and ϕ, and the other is a differential equation relating Ψ, $\omega(\Psi)$, and ϕ. These relations, however, reduce to those previously obtained in B–D theory for $\omega = \text{const}$. Further explicit expressions for both g_{00} and Ψ are obtained as functions of the electrostatic potential ϕ in Barker's special case. We have not yet succeeded to get exact solutions in particular cases of this static electrovac.

II. STATIC GRAVITATIONAL FIELD

The field equations in the metric formulation of Nordtvedt can be expressed in the form

$$G_{\mu\nu} = -\frac{8\pi}{\Psi}T_{\mu\nu} - \frac{\omega}{\Psi^2}(\Psi_\mu\Psi_\nu - \tfrac{1}{2}g_{\mu\nu}\Psi_\alpha\Psi^\alpha) - \frac{1}{\Psi}(\Psi_{\mu;\nu} - g_{\mu\nu}\Box\Psi) \tag{1}$$

and

$$\Box\Psi = -\frac{\Psi_\alpha\Psi^\alpha}{(2\omega+3)}\left(\frac{d\omega}{d\Psi}\right). \tag{2}$$

The line element for a static space–time can be written as

$$ds^2 = g_{00}dt^2 + g_{ij}dx^i dx^j, \tag{3}$$

with g_{00} and g_{ij} being functions of space coordinates only, and i, j being 1, 2, and 3. One of the field equations, which is of interest, can be written as

$$\Psi R^0{}_0 = -(\omega/\Psi)\Psi^0\Psi_0 - \Psi^0{}_{;0} - \tfrac{1}{2}\Box\Psi. \tag{4}$$

Here any subscript μ indicates derivative with respect to x^μ coordinate. Now since for a static metric (1)

$$R^0{}_0 = \frac{1}{2(-g)^{1/2}}(g^{00}g^{ij}(-g)^{1/2}g_{00,j})_{,i} \tag{5}$$

and $\Psi_{,0} = 0$, Eq. (4) leads to

$$\Psi[g^{ij}g^{00}(-g)^{1/2}g_{00j}]_{,i} + [g^{ij}\sqrt{-g}\Psi_i]_{,j} + g^{00}g^{ij}(-g)^{1/2}\Psi_i g_{00j} = 0. \tag{6}$$

Assuming now that a functional relationship exists between g_{00} and the scalar field Ψ, and using the wave equation (2), one can, in turn, write Eq. (6) in the form

$$\left(\Psi\frac{g'_{00}}{g_{00}} + 1\right)_{,i}(\Psi^i(-g)^{1/2}) = \left(\frac{\Psi g'_{00}}{g_{00}} + 1\right)\frac{(-g)^{1/2}(d\omega/d\Psi)}{(2\omega+3)}\Psi_j\Psi^j. \tag{7}$$

Here the prime indicates differentiation with respect to the

[a)] On leave from the Department of Physics, Jadavpur University, Calcutta 700032, India

 0022-2488/80/122799-03$1.00

scalar Ψ. Now in view of the fact that $\omega = \omega(\Psi)$, Eq. (7) can be written in a slightly modified form,

$$\frac{[\Psi(g'_{00}/g_{00})+1]_{,i}}{[\Psi(g'_{00}/g_{00})+1]}\Psi^i = \frac{(2\omega+3)_{,i}\Psi^i}{2(2\omega+3)}. \tag{8}$$

We write $\ln[\Psi(g'_{00}/g_{00})+1] = X$ and $\ln(2\omega+3)^{1/2} = Y$, and thus get from (8) the relation

$$X_{,i}\Psi^i = Y_{,i}\Psi^i. \tag{9}$$

Again since $X = X(\Psi)$ and $Y = Y(\Psi)$, Eq. (9) leads to

$$X' = Y',$$

which in turn, on integration yields

$$X = Y + \text{const.} \tag{10}$$

In B–D theory $\omega = \text{const}$ and consequently $Y' = X' = 0$ and we get $X = \text{const}$. This is the known result in B–D theory previously obtained by Banerjee and Bhattacharya. Equation (10) gives a relation between g_{00}, ω, and Ψ in the general scalar tensor theory of Nordtvedt as

$$[\Psi(g'_{00}/g_{00})+1] = A(2\omega+3)^{1/2}, \tag{11}$$

A being the integration constant. If ω is a known function of Ψ, g_{00} can be expressed explicitly as a function of Ψ on integration of (11). In the special case proposed by Barker, $\omega = (4-3\Psi)/2(\Psi-1)$ which has a consequence that the Newtonian G turns out to be a constant. It is now easy to integrate Eq. (11) if one substitutes for ω given by Barker and the integration yields the exact form of g_{00} as a function of Ψ, in the form

$$g_{00}\Psi = \text{const} \times e^{2A\tan^{-1}(\Psi-1)^{1/2}}. \tag{12}$$

Next we proceed to give here an exact solution for a plane symmetric static gravitational field in Nordtvedt's general scalar tensor theory with ω given by Barker's form. The line element in this case is given by

$$ds^2 = e^{2\alpha}(dt^2 - dx^2) - e^{2\beta}(dy^2 + dz^2),$$

where α and β are functions of the x coordinate. The field equations (1) and the equation (2) are now explicitly written for this metric as

$$\begin{aligned}
&\Psi(\beta_1^2 + 2\alpha_1\beta_1) = \omega\Psi_1^2/2\Psi - \alpha_1\Psi_1 - 2\beta_1\Psi_1,\\
&\Psi(\alpha_{11} + \beta_{11} + \beta_1^2) = -\omega\Psi_1^2/2\Psi - \beta_1\Psi_1 - \Psi_{11},\\
&\Psi(2\beta_{11} + 3\beta_1^2 - 2\alpha_1\beta_1)\\
&\quad = -\omega\Psi_1^2/2\Psi + \alpha_1\Psi_1 - \Psi_{11} - 2\beta_1\Psi_1,\\
&\Psi_{11} + 2\beta_1\Psi_1 = -\omega_1\Psi_1/(2\omega+3).
\end{aligned} \tag{13}$$

Omitting details of the steps for integration procedure the solutions of the set of equations (13) can be finally written in the form

$$\begin{aligned}
e^{2\alpha} &= k(ax+b)^{(c-1)/2}\cos^2\ln[d(ax+b)^{\sqrt{c}/2}],\\
e^{2\beta} &= (ax+b)\cos^2\ln[d(ax+b)^{\sqrt{c}/2}],
\end{aligned} \tag{14}$$

and

$$\Psi = \sec^2\ln[d(ax+b)^{\sqrt{c}/2}],$$

where a, b, c, d, and k are all arbitrary constants appearing in the processes of integration. It can be easily verified that the solution (14) obtained in the static plane symmetric case are consistent with the relation previously derived in (2) in a more general situation. When we put $c = 0$, $\Psi = \text{const}$ and the solutions (14) reduce to those given by Taub[6] in general relativity.

III. STATIC ELECTROVAC

In this section we extend our discussions to the case of a static electrovac, where the electrostatic field existing can be represented by the nonvanishing field tensor F_{0i} with $i = 1,2,3$, and further $F_{0i} = \phi_{,i}$, ϕ being the electrostatic potential. Then Maxwell's equation can be easily expressed as

$$[g^{00}g^{ij}(-g)^{1/2}\phi_j]_{,i} = 0. \tag{15}$$

For an electromagnetic field the energy momentum tensor is written in the form

$$4\pi T_{\mu\nu} = -F_{\mu\alpha}F_{\nu}{}^{\alpha} + \tfrac{1}{4}g_{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}, \tag{16}$$

and in view of (16) one gets from the field equations (1) the relation

$$\Psi R^0{}_0 = g^{00}g^{ij}\phi_i\phi_j - (\omega/\Psi)\Psi^0\Psi_0 - \Psi^0{}_{;0} - \tfrac{1}{2}\Box\Psi. \tag{17}$$

Since the field is static $\Psi_{,0} = 0$ and Eq. (17) reduces to

$$\Psi R^0{}_0 = g^{00}g^{ij}\phi_i\phi_j - \Psi^0{}_{;0} - \tfrac{1}{2}\Box\Psi, \tag{18}$$

and in view of (2), (5), and (18) one can immediately write

$$\begin{aligned}
&\Psi(g^{00}g^{ij}(-g)^{1/2}g_{00,j})_{,i}\\
&\quad = 2g^{00}g^{ij}(-g)^{1/2}\phi_i\phi_j - g^{00}g^{ij}(-g)^{1/2}g_{00,j}\Psi_i\\
&\qquad - (g^{ij}(-g)^{1/2}\Psi_j)_{,i}.
\end{aligned} \tag{19}$$

Now if one assumes that both g_{00} and Ψ are functions of the electric potential ϕ, which is, however, trivial for many special cases of symmetry, one can obtain from Eq. (19) the relation

$$\begin{aligned}
&[g^{00}g^{ij}(-g)^{1/2}g'_{00}\phi_j\Psi]_{,i}\\
&\quad = 2g^{00}g^{ij}(-g)^{1/2}\phi_i\phi_j - (g^{ij}(-g)^{1/2}\Psi_j)_{,i},
\end{aligned} \tag{20}$$

with prime denoting differentiation with respect to ϕ. Again since $g_{00}g^{00} = 1$, it is easy to show in view of Maxwell's equation (15) that

$$(g^{ij}(-g)^{1/2}\Psi_j)_{,i} = g^{00}g^{ij}(-g)^{1/2}\phi_j(g_{00}\Psi')_{,i}. \tag{21}$$

In Brans–Dicke theory $\Box\Psi = 0$, and in consequence one gets the condition $g_{00}\Psi' = \text{const}$. In Nordtvedt's extended theory, however, in general $g_{00}\Psi'$ is a variable. Using (21) and Maxwell's equation (15) in (20), one can immediately obtain

$$g^{00}g^{ij}(-g)^{1/2}\phi_j[(g_{00}\Psi)' - 2\phi]_{,i} = 0. \tag{22}$$

Writing now $\xi(\phi)$ for $[(g_{00}\Psi)' - 2\phi]$, which is a function of ϕ, Eq. (22) can be represented in the form

$$(g^{ij}\phi_i\phi_j)\xi' = 0,$$

which leads one to the conclusion that $\xi' = 0$ for a nonvanishing electric field, or in other words,

$$(g_{00}\Psi)' - 2\phi = a, \tag{23}$$

a being the integration constant. Integrating (23) one can write finally the relation connecting g_{00}, Ψ, and ϕ in the form

$$g_{00}\Psi = (\phi^2 + a\phi + b), \tag{24}$$

b being another integration constant. It may be noted that the relation (24) is identical with that obtained by Raychaudhuri and Bandyopadhyaya in Brans–Dicke theory even if in our case ω is not a constant.

In the next step we prove that for a static electrovac in Nordtvedt's general scalar tensor theory $g_{00}\Psi' \propto (2\omega+3)^{-1/2}$, with ω as function of Ψ. The proof is as follows.

The wave equation (2) for the scalar field can be written as

$$(g^{ij}(-g)^{1/2}\Psi'\phi_j)_{,i} = -\frac{(-g)^{1/2}}{(2\omega+3)} g^{ij}(\omega'\Psi')\phi_i\phi_j. \tag{25}$$

Since in view of $g_{00}g^{00} = 1$ and Maxwell's equations the left-hand side of (25) can be replaced by

$$[g^{00}g^{ij}(-g)^{1/2}\phi_j]\, g_{00,i}\Psi' + g^{ij}(-g)^{1/2}\phi_i\phi_j\Psi'',$$

one can easily reduce Eq. (25) to the form

$$\frac{g'_{00}}{g_{00}} + \frac{\Psi''}{\Psi'} = -\frac{\omega'}{(2\omega+3)}, \tag{26}$$

which, in turn, on integration finally yields the relation

$$g_{00}\Psi' = \frac{C}{(2\omega+3)^{1/2}}, \tag{27}$$

where C is the integration constant.

One can now write in a straightforward way from (24) and (27)

$$\frac{\Psi'}{\Psi} = \frac{C}{(2\omega+3)^{1/2}(\phi^2 + a\phi + b)}. \tag{28}$$

If $\omega = \text{const}$, the relations (27) and (28) reduce to those in Brans–Dicke theory. We can proceed further to find the explicit functional relationship of g_{00} and Ψ with the electric potential ϕ, provided we know the exact form of ω as a function of Ψ. For this we choose Barker's form $\omega = (4-3\Psi)/(2\Psi-2)$ mentioned previously in Sec. II. With this for ω Eq. (28) can be written in a modified form

$$\frac{\Psi'}{\Psi(\Psi-1)^{1/2}} = \frac{C}{\phi^2 + a\phi + b}. \tag{29}$$

It is now a differential equation relating Ψ with ϕ and yields on integration

$$\Psi = \sec^2 \ln C_1 \left(\frac{(2\phi+a) - (a^2-4b)^{1/2}}{(2\phi+a) + (a^2-4b)^{1/2}}\right)^{C/2(a^2-4b)^{1/2}}, \quad \text{for } a^2 > 4b,$$

$$\Psi = \sec^2\left(C_2 - \frac{c}{(2\phi+a)}\right), \quad \text{for } a^2 = 4b, \tag{30}$$

and

$$\Psi = \sec^2\left(C_3 + \frac{C}{(4b-a^2)^{1/2}} \tan^{-1}\frac{(2\phi+a)}{(4b-a^2)^{1/2}}\right), \quad \text{for } a^2 < 4b,$$

where C_1, C_2, and C_3 are constants of integration.

The relations (30) express Ψ as a function of the electric potential and one can now obtain a straightforward way the value of g_{00} as a function of ϕ by using Eq. (24).

ACKNOWLEDGMENT

One of the authors (A. Banerjee) would like to thank F.I.N.E.P. for the financial support.

[1] K. Nordtvedt, Astrophys. J. **161**, 1059 (1970).
[2] B.M. Barker, Astrophys. J. **219**, 5 (1978).
[3] A.K. Raychaudhuri and N. Bandyopadhyaya, Prog. Theor. Phys. **59**, 414 (1978).
[4] C. Brans and R.H. Dicke, Phys. Rev. **124**, 925 (1961).
[5] A. Banerjee and D. Bhattacharya, J. Math. Phys. **20**, 1908 (1979).
[6] A.H. Taub, Ann. Math. **53**, 472 (1951).

Dynamics in nonglobally hyperbolic, static space-times [a)]

Robert M. Wald [b)]
Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637

(Received 2 October 1979; accepted for publication 1 February 1980)

Ordinary Cauchy evolution determines a solution of a partial differential equation only within the domain of dependence of the initial data surface. Hence, in a nonglobally hyperbolic space-time, one does not have fully deterministic dynamics. We show here that for the case of a Klein–Gordon scalar field propagating in an arbitrary static space-time, a physically sensible, fully deterministic dynamical evolution prescription can be given. If the cosmic censor hypothesis should be overthrown, a prescription of this sort could rescue deterministic physics.

1. INTRODUCTION

The cosmic censor hypothesis of classical general relativity states that all singularities of gravitational collapse are hidden within black holes; that no "naked singularities"—visible to a distant observer—can be produced. A stronger version of this hypothesis recently proposed by Penrose[1] asserts that any physically reasonable spacetime must be globally hyperbolic. The solid theoretical evidence in favor of even the weaker form of this conjecture is still rather meager, consisting mainly of the analysis of spherical collapse and perturbations of spherical collapse[2] together with proofs of the impossibility of obtaining certain types of counter examples.[3] Indeed, the unaesthetic aspect of adding objects other than black holes as possible endpoints of gravitational collapse is probably more responsible than the above solid evidence for the widespread belief in the validity of the cosmic censor hypothesis in its weak form.

One of the main unaesthetic features of the lack of global hyperbolicity is that by definition, there is no initial data surface whose domain of dependence is the entire space-time. If naked singularities are formed in gravitational collapse, even the distant, asymptotically flat region of the space-time fails to lie in the domain of dependence of an initial surface. Thus, in nonglobally hyperbolic space-times, the dynamical equations cannot predict from initial conditions what happens in certain regions of the space-time. Physically, the reason for this is that singularities are present and the dynamical equations say nothing about what can (or cannot) come out of a singularity. Unless some additional type of boundary conditions can be imposed upon the singularity, a complete breakdown of predictability occurs in any region of the space-time where the singularity can be seen. In specific examples, it may be possible to invent boundary conditions on a singularity which yield a sensible, deterministic, dynamical evolution. But given the infinite variety of pathologies of singularities, it might well seem a hopeless task to invent a sensible general prescription for dynamical evolution in the presence of arbitrary singularities.

The purpose of this paper is to show that this task may not be quite as hopeless as it may at first appear. We shall consider the evolution of a Klein–Gordon scalar field in an arbitrary static space-time (with arbitrary singularities consistent with staticity). We will show that the problem of defining the dynamics can be translated into the problem of finding self-adjoint extensions of the spatial part of the wave operator. But the problem of finding self-adjoint extensions (as opposed to the problem of defining boundary conditions on singularities) is a well studied problem and, since the operator considered here is positive, it is known that positive self-adjoint extensions exist. Indeed, a natural choice of extension—namely, the Friedrichs extension—can be defined. Thus, the problem of defining dynamics of a Klein–Gordon field in a nonglobally hyperbolic, static space-time can be solved by using the prescription defined below in Sec. 2, choosing the Friedrichs (or another) self-adjoint extension. The methods and results below are special to the Klein–Gordon field in static space-times. However, the results indicate that if it should become necessary to abandon the cosmic censor hypothesis, it may well be possible to retain well defined, deterministic, physically sensible laws of dynamical evolution.

In Sec. 2, the dynamical evolution prescription is defined and shown to satisfy the following properties: (1) Solutions are uniquely determined throughout the space-time by their initial data; (2) Where ordinary dynamical evolution is defined (i.e., in the usual domain of dependence of the initial surface) the results coincide with the evolution prescription given here; (3) For smooth initial data of compact support, the solution is smooth throughout the space-time. Thus, our prescription defines a physically reasonable dynamical evolution.

Finally, in Sec. 3 we attempt to gain some insight into the meaning of the prescription in terms of "boundary conditions on singularities." We show that if the singularity of the space-time is an artificial one, that is, for an extendible space-time (so that a smooth boundary can be attached to the original manifold), the prescription defined by using the Friedrichs extension corresponds to putting zero Dirichlet conditions on the boundary. Thus, one may view the dynamical prescription of Sec. 2 (using the Friedrichs extension) as a means of generalizing the notion of Dirichlet boundary conditions to arbitrary naked singularities in static space-times.

[a)]Supported in part by NSF Grant PHY 78-24275 and by the Alfred P. Sloan Foundation.
[b)]Sloan Foundation Fellow.

 0022-2488/80/122802-04$1.00

2. PRESCRIPTION FOR DYNAMICS

Let $(M, g_{\mu\nu})$ be a static space-time, i.e., one that possesses a one parameter group of isometries with everywhere timelike orbits which are hypersurface orthogonal. We wish to consider the propagation of a massless Klein–Gordon scalar field ϕ, satisfying

$$\nabla^\mu \nabla_\mu \phi = 0. \tag{1}$$

Suppose we specify initial data for ϕ on a hypersurface Σ orthogonal to the static Killing field ξ^μ. If the space-time is not globally hyperbolic, Σ will not be a Cauchy surface and data on Σ will determine ϕ only in the domain of dependence $D(\Sigma)$. Outside $D(\Sigma)$, the partial differential Eq. (1) does not determine ϕ. Our aim is to formulate a physically sensible prescription for determining ϕ everywhere.

To do so, we rewrite Eq. (1) in the form

$$\partial^2\phi/\partial t^2 = VD^a(VD_a\phi), \tag{2}$$

where $V^2 = -\xi^\mu\xi_\mu$, t denotes the Killing parameter, and D_a denotes the derivative operator on the hypersurface Σ. We may then view

$$A = -VD^a(VD_a) \tag{3}$$

as an operator on the Hilbert space $\mathscr{H}$ of square integrable functions on Σ. If we choose the volume element used to define $\mathscr{H}$ to be V^{-1} times the natural volume element on Σ and if we initially define the domain of A to be $C_0^\infty(\Sigma)$ (i.e., the smooth function of compact support on Σ), then A will be a positive, symmetric (but not self-adjoint) operator. In this way, we may reformulate the problem of solving the partial differential Eq. (1) into the problem of finding a one-parameter family ϕ_t of vectors in $\mathscr{H}$ satisfying

$$d^2\phi_t/dt^2 = -A\phi_t. \tag{4}$$

Our reformulation, Eq. (4), is not strictly equivalent to the original Eq. (1): the time derivative in Eq. (4) is a Hilbert space derivative rather than a partial derivative at fixed spatial position and with the present definition of the domain of A, ϕ_t must lie in C_0^∞. These modifications do not yet improve our ability to solve the dynamical equation. However, we are now in a position to further modify Eq. (4) to yield our dynamical prescription.

Let A_E denote a positive self-adjoint extension of A. Because of the positivity of A, at least one such extension—the Friedrichs extension A_F—always exists.[4] We replace Eq. (4) by

$$d^2\phi_t/dt^2 = -A_E\phi_t, \tag{5}$$

and, in turn, replace Eq. (5) by its solution in terms of its initial data ϕ_0 and $\dot\phi_0$,

$$\phi_t = \cos(A_E^{1/2}t)\phi_0 + A_E^{-1/2}\sin(A_E^{1/2}t)\dot\phi_0. \tag{6}$$

Here the operators $\cos A_E^{1/2}t$ and $A_E^{-1/2}\sin A_E^{1/2}t$ are defined using the functional calculus of self-adjoint operators.[5,6] They are bounded operators (with $\|\cos A_E^{1/2}t\| = 1$ and $\|A_E^{-1/2}\sin A_E^{1/2}t\| = t$), and thus can be defined to act on all vectors ϕ_0 and $\dot\phi_0$ in $\mathscr{H}$. For ϕ_t defined by Eq. (6), it follows from the type of argument used in the proof of Stone's theorem[6] that for ϕ_0 and $\dot\phi_0$ in the domain of A_E, $d^2\phi_t/dt^2$ indeed exists (in the strong limit sense) and satisfies Eq. (5). Finally, it is clear that at $t = 0$, ϕ_t reduces to ϕ_0, while $d\phi/dt$ reduces to $\dot\phi_0$, so ϕ_0 and $\dot\phi_0$ are correctly identified as initial data for ϕ.

Our prescription for defining the dynamics of ϕ is thus the following: We prescribe a positive, self-adjoint extension A_E of A. The allowed initial data consists of all ϕ_0 and $\dot\phi_0$ lying in the domain of A_E. (Actually, $\dot\phi_0$ need only lie in $\mathrm{dom}A_E^{1/2}$.) In particular, since $\mathrm{dom}A_E \supset \mathrm{dom}A$, all C_0^∞ specifications of ϕ_0 and $\dot\phi_0$ are permitted. The solution corresponding to this initial data is then defined *everywhere* [not just in $D(\Sigma)$] by Eq. (6).

Having defined our prescription, we now turn to showing that it is physically sensible. Specifically, we first shall show that our solution Eq. (6) reproduces the solution of Eq. (1) determined by ordinary Cauchy evolution in the region $D(\Sigma)$ where Cauchy evolution is defined. Then we show that for initial data in C_0^∞ (or, more generally, for initial data in $\mathrm{dom}A_E^k$ for all k) our solution, Eq. (6), is smooth throughout the space-time.

Let ψ denote the solution obtained by ordinary Cauchy evolution in $D(\Sigma_0)$ of Eq. (1) with, say, smooth data $(\phi_0, \dot\phi_0)$ in $\mathrm{dom}A_E$ specified on the initial surface Σ_0. Suppose ϕ_t differed from ψ in $D(\Sigma_0)$. Then there would be a static hypersurface Σ_1 (corresponding to time $t = t_1$) such that, viewed as L^2-vectors on $\Sigma_1 \cap D(\Sigma_0)$, we have $\psi_{t_1} \neq \phi_{t_1}$. Let S be a Cauchy surface for $D(\Sigma_0)$ which coincides with Σ_1 on an open region where $\psi_{t_1} \neq \phi_{t_1}$ (see Fig. 1). Let $\dot f_{t_1}$ be a smooth function on S with compact support contained within $S \cap \Sigma_1$ such that

$$\int_{S\cap\Sigma_1} V^{-1}\dot f_{t_1}(\psi_{t_1} - \phi_{t_1}) \neq 0, \tag{7}$$

where here and in the following the natural volume element is understood in all integrals and the factor of V^{-1} is explicitly put in to yield the volume element used in defining $\mathscr{H}$. Define f throughout $D(\Sigma_0)$ to be the ordinary Cauchy evolution of the initial data ($f = 0$, $\dot f = \dot f_{t_1}$) on S; set $f = 0$ outside $D(\Sigma_0)$ in the region between Σ_0 and Σ_1. Then f satisfies Eq. (1) throughout the region between Σ_0 and Σ_1 and the restriction f_t of f to any hypersurface Σ_t lying between Σ_0 and Σ_1 lies in $C_0^\infty(\Sigma_t)$.

Consider, now the quantity

$$c(t) = \int_{\Sigma_t} V^{-1}\left\{f^*\left(\frac{\partial\psi}{\partial t} - \frac{d\phi_t}{dt}\right) - \frac{\partial f^*}{\partial t}(\psi - \phi_t)\right\}. \tag{8}$$

A simple calculation yields

$$\frac{dc}{dt} = \int_{\Sigma_t} V^{-1}\left\{f^*\left(\frac{\partial^2\psi}{\partial t^2} - \frac{d^2\phi_t}{dt^2}\right) - \frac{\partial^2 f^*}{\partial t^2}(\psi - \phi_t)\right\}. \tag{9}$$

But, since f and ψ are smooth solutions of Eq. (1) and f is of compact spatial support, a straightforward substitution of Eq. (2) to get rid of the time derivatives followed by integration by parts shows that the terms in ψ cancel. On the other hand, using Eq. (5), together with the fact that $\partial^2 f/\partial t^2 = -Af_t = -A_E f_t$, the terms in ϕ_t can be expressed as,

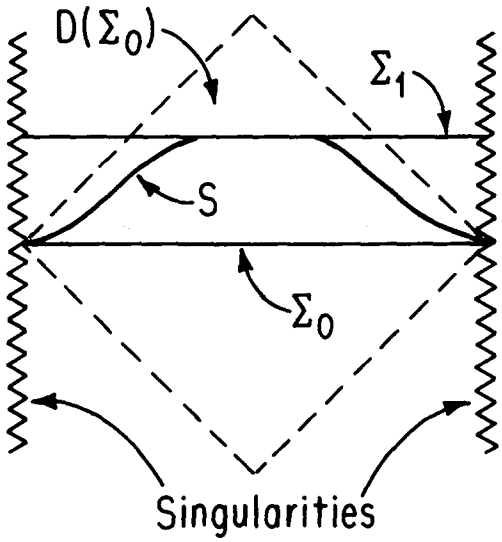


FIG. 1. The construction used in the proof that ϕ_t agrees with usual Cauchy evolution in $D(\Sigma_0)$.

$$\int_{\Sigma_t} V^{-1}\left\{-f^*\frac{d^2\phi_t}{dt^2}+\frac{\partial^2 f^*}{\partial t^2}\phi_t\right\}$$
$$=(f_t,A_E\phi_t)-(A_E f_t,\phi_t)$$
$$=0, \tag{10}$$

since A_E is self-adjoint. Thus, we have

$$dc/dt=0. \tag{11}$$

However, by construction $c(t_1)\neq 0$ and from the definition of $c(t)$ together with the assumption that ψ and ϕ_t have the same initial data on Σ_0, we have $c(t_0)=0$. This contradiction proves that ψ and ϕ_t must agree in $D(\Sigma_0)$.

Next, we show that our evolution prescription yields a smooth solution if the initial data is in $C_0^\infty(\Sigma_0)$ or, more generally, if ϕ_0 and $\dot\phi_0$ lie in $\mathrm{dom}A_E^k$ for all positive integers k (If ϕ and $\dot\phi_0$ are in $C_0^\infty(\Sigma_0)$, they clearly lie in $\mathrm{dom}A^k \subset \mathrm{dom}A_E^k$.) If ϕ_0 and $\dot\phi_0$ are C^∞ but not in $\mathrm{dom}A_E^k$ for all k, it is possible that our solution will still be smooth but our method of proof fails.

From the definition of ϕ_t, Eq. (6), it follows immediately that if $\phi_0,\dot\phi_0\in \mathrm{dom}A_E^k$, then $\phi_t\in\mathrm{dom}A_E^k$ and indeed,

$$A_E^k\phi_t=\cos(A_E^{1/2}t)A_E^k\phi_0+A_E^{-1/2}\sin(A_E^{1/2}t)A_E^k\dot\phi_0. \tag{12}$$

Thus, letting χ denote the vector on the right-hand side of Eq. (12), we see that for all $g\in C_0^\infty(\Sigma_t)$ we have,

$$(\phi_t,A^k g)=(\chi,g). \tag{13}$$

Equation (13) states that ϕ_t, viewed now as a distribution, is a weak solution of the partial differential equation.

$$A^k\phi_t=\chi. \tag{14}$$

But, on any open set $\Omega\subset\Sigma_t$ with compact closure, A^k is a strongly elliptic partial differential operator of order $2k$. Furthermore, since χ is in $\mathscr{H}$ it is certainly in $W_0(\Omega)$, where $W_m(\Omega)$ denotes the mth local Sobolev space[4] of Ω. Consequently, it follows from an elliptic regularity theorem of Friedrichs[7] that $\phi_t\in W_{2k}(\Omega)$. But Sobolev's lemma[4,8] then implies (for Σ three-dimensional) that $\phi_t\in C^{2k-2}(\Omega)$. Since both k and Ω are arbitrary, this implies $\phi_t\in C^\infty(\Sigma_t)$.

Thus, we have shown that for fixed t our solution is a smooth function of the spatial variables. Differentiability of ϕ_t with respect to t in the (strong) Hilbert space sense follows from the type of argument used in the proof of Stone's theorem.[6] Smoothness of the t-derivatives in the spatial variables then follows from a repetition of the above argument. However, as already noted above, the existence of the Hilbert space derivative with respect to t is not equivalent to the existence of the partial derivative with respect to t at fixed value of the spatial variable. Fortunately, we can prove space-time smoothness of our solution ϕ at an arbitrary point p as follows. Let Σ_t denote the static hypersurface passing through p. We have already shown that on Σ_t, ϕ_t and $d\phi_t/dt$ are smooth functions. Therefore, the ordinary solution ψ to the partial differential Eq. (1) with this initial data will be smooth throughout $D(\Sigma_t)$. But, by our previous result ϕ agrees with ψ in $D(\Sigma_t)$. Since $D(\Sigma_t)$ certainly includes p, this shows that ϕ is smooth at p in the space-time sense.

3. THE FRIEDRICHS EXTENSION AND DIRICHLET BOUNDARY CONDITIONS

In the previous section our dynamical evolution prescription was defined and shown to satisfy a number of reasonable conditions. However, these results do not shed light on the physical meaning of the prescription in terms of "boundary conditions on the singularity." In this section we shall attempt to gain insight into this issue by studying our prescription—using the Friedrichs extension[4] A_F of A—in the rather trivial case where the "singularities" are produced by "cutting out holes" from the space-time. More precisely, we consider the case where the given static space-time M is extendible to a larger static space-time M' and the boundary in M' of each static hypersurface Σ is a smooth two-dimensional manifold. We shall show that our requirement that ϕ_0 and $\dot\phi_0$ (and thus that our solution ϕ_t) lie in the domain of A_F implies "Dirichlet boundary conditions" for ϕ_t, i.e., that ϕ_t vanish on the boundary of Σ in M'. In other words, our solution is the one that would arise physically by putting a grounded conductor at the boundary of Σ. For the case of true singularities (i.e., an inextendible space-time) this result, of course, is not applicable. However, it does indicate that we may think of our dynamical prescription (using the Friedrichs extension A_F) as defining a generalized notion of Dirichlet boundary conditions applicable to true singularities.

We first demonstrate our result in the case of a two-dimensional space-time, taking the static hypersurface Σ to be simply a finite interval (a,b). The operator A in this case is simply

$$A=-V(x)\frac{d}{dx}\left(V(x)\frac{d}{dx}\right), \tag{15}$$

where, by the extendibility hypothesis, V approaches a finite, nonzero limit at the endpoints of the interval. Now, the domain of the Friedrichs extension, A_F, of the operator A [defined on the initial domain $C_0^\infty(\Sigma)$] is contained within the closure of $C_0^\infty(\Sigma)$ under the norm

$$\|f\|^2=(f,f)+(f,Af). \tag{16}$$

(Indeed, the Friedrichs extension is the unique self-adjoint

extension of A satisfying this property.[4]) But, using the explicit form of A, Eq. (15), for $f \in C_0^\infty$ we have

$$(f,Af) = \int_a^b V(x) \left| \frac{df}{dx} \right|^2 dx. \tag{17}$$

Consequently, the norm defined by Eq. (16) is equivalent to the first Sobolev norm of f. Since Σ is one-dimensional, convergence of a sequence of C_0^∞ functions in the first Sobolev norm implies uniform convergence of these functions.[8] Hence, every function in the domain of A_F is continuous (since it is the uniform limit of a sequence of continuous functions) and can be continuously extended to a function that vanishes at the endpoints of the interval (since every function in the sequence can be so extended). Hence, in the case where Σ is one-dimensional, every solution ϕ_t defined by our prescription satisfies Dirichlet conditions on the boundary of Σ.

The argument is similar for the case where Σ is three-dimensional and has a smooth two-dimensional boundary in the extended space-time. Again, the domain condition, Eq. (16), of the Friedrichs extension implies that ϕ_t is locally in the first Sobolev space. However, in three dimensions this does not imply that ϕ_t is continuous so we cannot necessarily even speak of the numerical values of ϕ_t as one approaches the boundary. However, it does imply that the restriction of ϕ_t to a two-dimensional hypersurface in Σ defines a locally L^2-function.[4] Our aim is to prove that for $\phi_t \in \mathrm{dom} A_F$ the restriction of ϕ_t to two-dimensional surfaces varies continuously and (viewed as an L^2-vector) vanishes as one goes to the boundary. To do so, we pick an open set on the boundary with compact closure and in a neighorhood of this portion of the boundary construct geodesic normal coordinates, thus obtaining a one-parameter family of two-dimensional surfaces σ_s which approach this part of the boundary as $s \to 0$. Fix a smooth function g with support contained within a compact region Γ where the geodesic normal construction is valid. (g is *not* required to vanish on the boundary.) Let $f \in C_0^\infty(\Sigma)$ and define

$$h(s) = \int_{\sigma_s} g f, \tag{18}$$

where the natural volume element induced on σ_s is used in the integral. Then, using the Schwarz inequality, we find

$$|h(s)|^2 \leqslant C_1 \int_{\sigma_s} |f|^2, \tag{19}$$

$$\left| \frac{dh}{ds} \right|^2 \leqslant C_2 \int_{\sigma_s} \left(|f|^2 + \left| \frac{\partial f}{\partial s} \right|^2 \right). \tag{20}$$

where C_1 and C_2 depend on g. Hence, we have

$$\begin{aligned} \int ds \left(|h|^2 + \left| \frac{dh}{ds} \right|^2 \right) &\leqslant C_3 \int_\Gamma \left(|f|^2 + \left| \frac{\partial f}{\partial s} \right|^2 \right) \\ &\leqslant C_3 \int_\Gamma (|f|^2 + |D^a f|^2) \\ &\leqslant C_4 \int_\Gamma (V^{-1}|f|^2 + V|D^a f|^2) \\ &\leqslant C_4 \int_\Sigma (V^{-1}|f|^2 + V|D^a f|^2) \\ &= C_4\{(f,f) + (f,Af)\}. \end{aligned} \tag{21}$$

Let $\{f_n\}$ be a sequence in $C_0^\infty(\Sigma)$ which approaches ϕ_t in the norm, Eq. (16). By Eq. (21) $h_n(s)$ will converge in the first Sobolev norm. Hence, as in the one-dimensional case, its limit

$$H(s) = \int_{\sigma_s} g \phi_t, \tag{22}$$

will be a continuous function which vanishes as $s \to 0$. Since g is an arbitrary smooth function, this yields the desired result that ϕ_t (viewed as a locally L^2-function on σ_s) varies continuously with s and goes to zero on the boundary. Thus, in the three-dimensional case, our prescription also yields Dirichlet boundary conditions.

[1] R. Penrose, in *General Relativity: An Einstein Centenary Survey*, edited by S.W. Hawking and W. Israel (Cambridge U.P., Cambridge, England, 1979).

[2] See, e.g., R. Price, Phys. Rev. D **5**, 2419 (1972); R.M. Wald, J. Math. Phys. **20**, 1056 (1979), Erratum **21**, 218 (1980).

[3] See, e.g., P.S. Jang and R.M. Wald, J. Math. Phys. **18**, 41 (1977).

[4] M. Reed and B. Simon, *Fourier Analysis, Self-Adjointness* (Academic, New York, 1975).

[5] M. Reed and B. Simon, *Functional Analysis* (Academic, New York, 1972).

[6] F. Riesz and B. Sz-Nagy, *Functional Analysis* (Ungar, New York, 1955).

[7] K.O. Friedrichs, Commun. Pure Appl. Math. **6**, 299 (1953). (We use the theorem in the form quoted on p. 112 of Ref. 4).

[8] See, e.g., P. Gilkey, *The Index Theorem and the Heat Equation* (Publish or Perish, Boston, 1975).

The use of anticommuting variable integrals in statistical mechanics. I. The computation of partition functions [a)]

Stuart Samuel [b)]
Lawrence Berkeley Laboratory, Unviersity of California, Berkeley, California 94720
and Institute for Advanced Study, Princeton, New Jersey 08540

(Received 16 July 1980; accepted for publication 30 July 1980)

Integrals over anticommuting variables are use to rewrite partition functions as fermionic field theories. The method is used to solve the two-dimensional Ising model, the planar close-packed dimer problems, and the free-fermion eight vertex model.

I. INTRODUCTION

The interplay of field theory and statistical mechanics is important. Many complicated field theories have simple underlying statistical mechanics analogues.[1] This supplies physical insight into these complicated field theoretic structures and allows the extraction of the key concepts. On the other hand, when a statistical mechanics model is expressed as a field theory, various field theory techniques can be used such as perturbation theory, operator methods, variational methods, functional methods, etc. These are powerful avenues of attack, especially for extracting numbers. In short, the statistical mechanics point of view allows one physical insight whereas the field theory point of view supplies the powerful mathematical tools. It is therefore important to establish connections between statistical mechanics and field theory. It is in this direction that these papers are written.

We will write statistical mechanics systems as fermionic field theories. This is done to systems which *a priori* have no vestige of fermionic character. What is involved is a mathematical rewriting of the degrees of freedom. A functional integral approach is used. This involves integrals over anticommuting variables (Grassmann integrals). It has been known for a long time that anticommuting variables are necessary for a fermionic path integral formulation.[2] Previously, however, such integrals were used only in formal ways,[3] rarely being employed in actual calculations. In this paper and the following ones they will be used in a practical manner to obtain numbers. They are, without a doubt, powerful mathematical tools.

In short, new mathematical methods are introduced to attack statistical mechanics problems by expressing partition functions as fermionic field theories via Grassmann integrals.

The new anticommuting variable techniques are important for two reasons. First, models solvable by previous methods are more easily solved using anticommuting variables. For example, the two-dimensional Ising model is solved in one line [Eq. (3.3)], a page of algebra yields the partition function [Eq. (3.12)] [later on graphical methods are introduced which solve the model by drawing one picture (see Sec. IV and Fig. (13)], and in a few more pages all correlation functions are computed (see Sec. III of paper II). This is the best way to solve the Ising model and compute physical quantities. The above statement applies to other two-dimensional models (free-fermion ferroelectric vertex models, planar closed-packed dimer problems, etc.). The only two-dimensional partition functions not yet computed via anticommuting variables are those solved by the Bethe ansatz.

Second and most important, anticommuting variables are useful in treating unsolved models. Most physical systems are not exactly solvable. Therefore, methods which exactly solve models but which cannot be adapted to unsolved models are not nearly as useful as those which can handle both. The anticommuting variables are in the latter class. Papers I and II show that they can solve the solvable models with ease. Paper III will show how they can generate viable approximation schemes. Although many models have been treated,[4,5,6,7] the contents of paper III are restricted for reasons for space to one unsolvable class of models, the dimer-monomer mixing problems. From the anticommuting variable viewpoint they are the simplest models in which to apply approximation methods. Paper III, in fact, numerically solves the monomer-dimer mixing models. In effect, an unsolvable model is solved. The point is that anticommuting variables yield good techniques for unsolved models.

Our method is completely new. There are other techniques with which anticommuting variables might be confused. These other techniques are different. There is the operator formalism[8,9,10] of Lieb, Schultz, and Mattis which solves the Ising model. Their basic objects are fermionic creation and destruction operators, $b_i, b_j^\dagger$, which satisfy canonical commutation relations, $b_i b_j^\dagger + b_j^\dagger b_i = \delta_{ij}$. The anticommuting variables $\eta_i, \eta_i^\dagger$, completely commute: $\eta_i \eta_j^\dagger + \eta_j^\dagger \eta_i = 0$. Unlike this paper, Ref. 8 used a transfer matrix method. The two methods are different and anticommuting variables are much more powerful. Pfaffian methods[9,11] have also been used to solve various two-dimensional models. Whenever the anticommuting variable action is quadratic, it is a Pfaffian according to Eq. (2.7) and, in principle, can be solved using Pfaffian methods. In this sense and for solvable models Pfaffian techniques come closest to anticommuting variable techniques. However, these two methods are different; many simplifications occur when using anticommuting variables, and, of course, Pfaffian methods cannot handle unsolved models.

[a)]Supported by the High Energy Physics Division of the United States Department of Energy.

 0022-2488/80/122806-09$1.00

What are the advantages of anticommuting variables over previous techniques? (A) Anticommuting variables are more natural. Grassmann integrals immediately present the problem in a powerful familiar form: as a fermionic field theory. Standard field theory method become applicable. (B) It is easy to express systems in integral form. There are brute force methods of doing this. Almost any lattice model with a graphical representation is expressible as a fermionic (albeit interacting) field theory. The same model often has several representations. This is where ingenuity is required. It is important to be efficient and elegant. Actions involving many types or large products of variables are useless. Approximation schemes will yield poor results and a proliferation of variables makes manipulations difficult. (C) Point (B) implies a wide range of applicability. Anticommuting variables have been applied to a large number of problems in two, three, and more dimensions. Pfaffian technqiues are restricted to two dimensions. (D) Technical problems are easier to handle. With Pfaffian methods every site on the lattice must be ordered to determine the overall sign. With anticommuting variables the minus sign problem can be treated locally (see Fig. 5 for the rules). Thus, extra minus signs are easily determined. Anticommuting variables are simple to manipulate. Given a string of fermionic creation and destruction operators a proliferation of terms is generated in getting destruction operators to the right of creation operators. Because anticommuting variables completely anticommute there is only one term. Anticommuting variables are more like ordinary numbers. It is easier to compute partition functions and correlation functions. The graphic methods of Sec. IV greatly simplify the task. (E) The big disadvantage of Pfaffian methods is their inability to handle "interacting" theories. Pfaffians are too ackward to treat unsolved models. Anticommuting variables, however, can handle such systems and do generate good approximation methods. All the techniques of many-body theory are available. This is by far their most important advantage.

Several models have quadratic action representations. Among these are the two-dimensional Ising model and the two-dimensional close-packed planar dimer problem. They are free theories and are exactly solvable. In this paper, these two partition functions are explicitly computed (Secs. III and IV). This is a straightforward calculation: one transforms to momentum space just as one would do with a free field theory. This partically diagonalizes the problem; it breaks up into a product of 4×4 determinants. Next, graphical methods are introduced to organize the algebra (Sec. IV). They are useful because they are systematic and pictorial. Section IV considers the general class of solvable 2-dimensional close-packed dimer problems on various lattices. A set of rules is derived which quickly computes partition functions. These rules are illustrated using the square lattice. Next, the rules are extended to general free theories. This means that, given any quadratic action, there is a simple and quick calculational procedure.

II. INTEGRALS OVER ANTICOMMUTING VARIABLES

This section will review[12] needed properties of integrals over Grassmann variables. More details may be found in Ref. 12. A set of N Grassmann (or anticommuting) variables are objects, η_α ($\alpha = 1,2,...,N$), satisfying

$$\eta_\alpha\eta_\beta + \eta_\beta\eta_\alpha = 0. \tag{2.1}$$

In particular, $\eta_\alpha^2 = 0$. Taking sums and products the most general construct is

$$f = a_0 + \sum_\alpha a_\alpha\eta_\alpha + \sum_{\alpha<\beta} a_{\alpha\beta}\eta_\alpha\eta_\beta + \cdots + a_{123\cdots N}\eta_1\eta_2\cdots\eta_N, \tag{2.2}$$

with the a's real or complex numbers. Functions of these variables are defined via Taylor series, which because of Eq. (2.1) terminate at the N th order. Equation (2.2) is the most general function, an N th order polynomial.

The anticommuting variable integral of a function, f, of the form of Eq. (2.2) is defined by

$$\int d\eta f \equiv \int d\eta_1\, d\eta_2\cdots d\eta_N f \equiv a_{123\cdots N}. \tag{2.3}$$

The only term which contributes is the one where each η occurs precisely once, the sign being determined by the order (for example, $\int d\eta_1\, d\eta_2\, \eta_2\, \eta_1 = -1$). Often η's are associated in pairs (or conjugates), one of which will have a dagger (i.e., η_α and $\eta_\alpha^\dagger$). This is convenient for determining the sign of an integral. For these the measure is defined as $\int d\eta\, d\eta^\dagger \equiv \int d\eta_1\, d\eta_1^\dagger\cdots d\eta_N d\eta_N^\dagger$.

Statistical mechanics problems will involve spins, atoms, bonds, etc. at sites, $\mathbf{x}$, to which anticommuting variables will be assigned. The variable, $\mathbf{x}$, will range over the region of interest; for a cubic crystal this might be a three-dimensional lattice so that $\mathbf{x} = (\alpha,\beta,\gamma)$ has integer coordinates. Often several variables are needed at a site, in which case, an additional label, r, is required, and the η's will appear as $\eta_{\mathbf{x}}^r$, $\eta_{\mathbf{x}}^{r\dagger}$ $\{r = 1,2,...,T\}$ for T types. Graphically $\eta_{\mathbf{x}}$ and $\eta_{\mathbf{x}}^\dagger$ may be represented by an o and an x at $\mathbf{x}$. Different types may be distinguished by using different colors. The important point to remember is that a contribution to an integral occurs only if each site is covered by one o and one x of each color (type).

Key properties of these integrals which are consequences of Eq. (2.3) are the following:

1. *Shift of variable*: Given J_α which anticommute with themselves and with all the η's,

$$\int d\eta f(\{\eta_\alpha\}) = \int d\eta f(\{\eta_\alpha + J_\alpha\}). \tag{2.4}$$

2. *Change of variables*: Let $\psi_\alpha = \Sigma_\beta A_{\alpha\beta}\eta_\beta$ (with A invertible) be linear combinations of η's and hence an equivalent set of anticommuting variables. Then

$$\int d\eta f(\eta) = (\det A)\int d\psi f(A^{-1}\psi). \tag{2.5}$$

Contrast this with normal (i.e., Riemann) integration where there is a factor $(\det A)^{-1}$ rather than $(\det A)$ in Eq. (2.5).

3. *Quadratic and quadratic-like actions*:

$$\int d\eta\, d\eta^\dagger \exp\Big(\sum_{\alpha\beta}\eta_\alpha A_{\alpha\beta}\eta_\beta^\dagger\Big) = \det A, \tag{2.6}$$

$$\int d\eta \exp\Big(\frac{1}{2}\sum_{\alpha\beta}\eta_\alpha A_{\alpha\beta}\eta_\beta\Big) = \mathrm{Pf}A, \tag{2.7}$$

$$\int d\eta\, d\eta^\dagger \int d\psi\, d\psi^\dagger \exp\left(\sum_{\alpha\beta} \eta_\alpha \eta_\alpha^\dagger A_{\alpha\beta} \psi_\beta \psi_\beta^\dagger\right) = \mathrm{perm} A, \tag{2.8}$$

$$\int d\eta\, d\eta^\dagger \exp\left(\frac{1}{2}\sum_{\alpha\beta} \eta_\alpha \eta_\alpha^\dagger A_{\alpha\beta} \eta_\beta \eta_\beta^\dagger\right) = \mathrm{hf} A. \tag{2.9}$$

These are respectively the determinant, Pfaffian,[13] permanent, and hfaffian of A. Permanents and hfaffians are determinants and Pfaffians without the sign of the permutation factor. In Eqs. (2.7) and (2.9) A must be even-dimensional. In Eq. (2.7) A may be chosen to be antisymmetric. In Eq. (2.9) it may be chosen to be symmetric, but must have zero's along the diagonal. These equations are easily proved by expanding the exponents: permutations of products of $A_{\alpha\beta}$ are obtained with the appropriate combinatorial and sign factors. Equation (2.6), however, is easier to prove by transforming $\eta^\dagger \to A^{-1}\eta^\dagger$ and using Eq. (2.5).

Anticommuting variables are powerful objects. Let us prove the well-known result[14] that $(\mathrm{Pf} A)^2 = \det A$ for an antisymmetric even-dimensional matrix. Usual proofs are quite cumbersome. Using Eq. (2.6) and rewrite

$$\eta_\alpha = \sqrt{\tfrac{1}{2}}(\eta_\alpha^{(1)} + i\eta_\alpha^{(2)}), \quad \eta_\alpha^+ = \sqrt{\tfrac{1}{2}}(\eta_\alpha^{(1)} - i\eta_\alpha^{(2)}),$$

$d\eta_\alpha d\eta_\alpha^\dagger = i d\eta_\alpha^{(1)}\, d\eta_\alpha^{(2)}$. Since A is antisymmetric $\eta_\alpha A_{\alpha\beta}\eta_\beta^\dagger = (1/2)\eta_\alpha^{(1)} A_{\alpha\beta}\eta_\beta^{(1)} + (1/2)\eta_\alpha^{(2)} A_{\alpha\beta}\eta_\beta^{(2)}$ (the cross terms cancel). The exponent factors into two exponents and the integral factorizes into two integrals, each of the form of Eq. (2.7).

Finally, one may take derivatives of anticommuting variables. For example, $(d/d\eta_1)\eta_1 = 1$, $(d/d\eta_1)\eta_2 = 0$. All the usual rules of differentiation hold except for minus signs in the product rule due to anticommutation relations. Thus $(d/d\eta_1)(\eta_2\eta_1) = ((d/d\eta_1)\eta_2)\eta_1 - \eta_2(d/d\eta_1)\eta_1 = -\eta_2$. These derivatives act to the right. Derivatives acting to the left are defined analogously: $\eta_1 \overleftarrow{d}/d\eta_1 = 1$. A powerful tool is the following:

4. *Integration by parts*: Given two functions, f and g,

$$\int d\eta\, f \frac{\vec{d}}{d\eta} g = \int d\eta\, f \frac{\overleftarrow{d}}{d\eta} g. \tag{2.10}$$

In conclusion, anticommuting variables may be manipulated, integrated, and differentiated much like ordinary variables except that anticomutation must be taken into account.

III. THE SOLUTION OF THE TWO-DIMENSIONAL ISING MODEL AND THE FREE-FERMION 8-VERTEX MODEL

The partition functions of a large number of statistical mechanics systems have representations in terms of anticommuting variable integrals. There are brute force methods of doing this. They require many types of anticommuting variables and have actions containing many terms involving products of many variables. Space limitations prevent us from illustrating these methods.[4] Instead, I will focus on models having simple representations. This paper will consider (solvable) models with quadratic actions. The third paper will treat interacting models with quartic actions.

Many partition functions which have a graphic representation are expressible as anticommuting integrals. The d-dimensional Ising model[15] has such a graphical representation,[9,11,14] where one sum's over closed nonoverlapping but (posssibly) intersecting polygonal curves; in two dimensions this is obtained by drawing curves separating regions of up spin from down spin. There is a Boltzmann factor for each unit of "Bloch" wall. Alternatively, one may use bond variables[16] (which works in any dimension) for which there is a similar representation with different Bloch wall Boltzmann factors.

Let us consider $d = 2$. Then

$$Z_{\mathrm{Ising}}(J_\mathrm{h}, J_\mathrm{v}) = f Z_{\text{closed polygons}}(z_\mathrm{h}, z_\mathrm{v}), \tag{3.1}$$

where $Z_{\mathrm{Ising}}(J_\mathrm{h}, J_\mathrm{v})$ is the Ising model partition function, with horizontal and vertical spin couplings J_h and J_v, $Z_{\text{closed polygons}}(z_\mathrm{h}, z_\mathrm{v})$ is the partition function for closed nonoverlapping polygons with Boltzmann weights, z_h and z_v for horizontal and vertical Bloch walls, and f is a multiplicative factor:

$$\begin{aligned} f &= \exp[N(\beta J_\mathrm{v} + \beta J_\mathrm{h})], \\ z_\mathrm{h} &= \exp(-2\beta J_\mathrm{v}), \\ z_\mathrm{v} &= \exp(-2\beta J_\mathrm{h}), \end{aligned} \tag{3.2}$$

where N is the number of sites.

The closed polygons should be considered as particle trajectories (vacuum bubble loops). The particles should be fermions so that polygons cannot overlap.

I will use anticommuting variables to draw the polygonal configurations. Two sets of variables will be used at each (α,β) site: $\eta_{\alpha\beta}^{\mathrm{h}^o}, \eta_{\alpha\beta}^{\mathrm{h}^x}$, and $\eta_{\alpha\beta}^{\mathrm{v}^o}, \eta_{\alpha\beta}^{\mathrm{v}^x}$. The superscripts h and v stand for horizontal and vertical. Consider

$$Z_{\text{closed polygons}}(z_\mathrm{h}, z_\mathrm{v}) = (-1)^N \int d\eta^o\, d\eta^x \exp(A), \tag{3.3}$$

where N is the number of sites and

$$\begin{aligned} A &= A_{\text{Bloch wall}} + A_{\text{corner}} + A_{\text{monomer}}, \\ A_{\text{Bloch wall}} &= \sum_{\alpha\beta} (z_\mathrm{h} \eta_{\alpha\beta}^{\mathrm{h}^x} \eta_{\alpha+1\beta}^{\mathrm{h}^o} + z_\mathrm{v} \eta_{\alpha\beta}^{\mathrm{v}^x} \eta_{\alpha\beta+1}^{\mathrm{v}^o}), \\ A_{\text{corner}} &= \sum_{\alpha\beta} (a_1 \eta_{\alpha\beta}^{\mathrm{h}^x} \eta_{\alpha\beta}^{\mathrm{v}^o} + a_3 \eta_{\alpha\beta}^{\mathrm{v}^x} \eta_{\alpha\beta}^{\mathrm{h}^o} \\ &\quad + a_2 \eta_{\alpha\beta}^{\mathrm{v}^x} \eta_{\alpha\beta}^{\mathrm{h}^x} + a_4 \eta_{\alpha\beta}^{\mathrm{v}^o} \eta_{\alpha\beta}^{\mathrm{h}^o}), \\ A_{\text{monomer}} &= \sum_{\alpha\beta} (b_\mathrm{h} \eta_{\alpha\beta}^{\mathrm{h}^o} \eta_{\alpha\beta}^{\mathrm{h}^x} + b_\mathrm{v} \eta_{\alpha\beta}^{\mathrm{v}^o} \eta_{\alpha\beta}^{\mathrm{v}^x}). \end{aligned} \tag{3.4}$$

The Bloch wall action produces a unit of Bloch wall in either the horizontal or vertical direction (see Fig. 1) weighted by

FIG. 1. Block wall operators: (a) is the graphical representation of $\eta_{\alpha\beta}^{\mathrm{h}^x} \eta_{\alpha+1\beta}^{\mathrm{h}^o}$ which occurs in Eq. (3.4) and produces a horizontal Block wall; (b) is the vertical Bloch wall operator, $\eta_{\alpha\beta}^{\mathrm{v}^x} \eta_{\alpha\beta+1}^{\mathrm{v}^o}$.

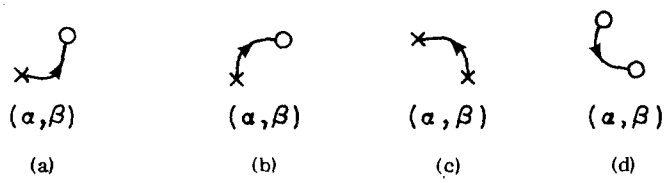


FIG. 2. The corner operators in Eq. (3.4): In all cases they occur at the (α,β) site, that is corner operators only change the direction of a curve; they do not connect neighboring sites. Although one could use labels to distinguish horizontal and vertical variables, it's easier to use the following convention: if an o or an x has a horizontal line coming into or out of it, it is a horizontal variable; on the other hand vertical variables have vertical lines flowing into or out of them. For example, (a) involves a horizontal x or $\eta^{h^x}_{\alpha\beta}$ and a vertical o or $\eta^{v^o}_{\alpha\beta}$. The arrow indicates the order, so that this term is $\eta^{h^x}_{\alpha\beta}\eta^{v^o}_{\alpha\beta}$, the first term in A_{corner} of Eq. (3.4). (b), (c), and (d) are the other three terms.

the appropriate Boltzmann factor. The term A_{corner} produces the four corners of Fig. (2) necessary to construct a ploygon.

The graphical notation in Figs. 1 and 2 is as follows: If a horizontal line is attached to a variable it is a horizontal variable. Likewise a line joins vertically to a vertical variable. Arrows denote the order of variables. The arrow originates from the first anticommuting variable and terminates on the second one. In this way Figs. 1(a) and 1(b) can precisely be associated with the terms in $A_{\text{Bloch wall}}$. Likewise for Fig. 2 and the corner action. Expand the exponent in Eq. (3.3). By the "golden rule" of Grassmann integrals, each site must have a horizontal x and o and a vertical x and o. The x's and o's link up to form precisely the Ising model polygons. The sides of polygons cannot overlap because the square of an anticommuting variable is zero. Likewise, the double corners of Fig. 3 do not occur; a single corner uses up both horizontal and vertical variables. Each polygonal configuration is included precisely once. Finally, A_{monomer} fills all unoccupied h and v sites with $o - x$ pairs (monomers). I have allowed for the most general quadratic form by weighting corners with a_i. This more general model is known as the free-fermion model. The eight possible configurations which can occur at a site are shown in Fig. 4 with their weights. There is an extra (-1) for each site because of the $(-1)^N$ in Eq. (3.3). For the Ising model set all $a_i = b_h = b_v = -1$. Although the action in Eq. (3.4) produces the polygonal configurations, it may not necessarily produce them with positive weight. This could be upset due to reordering of anticommuting variables. The Appendix deals with these kinds of minus signs. The result is the extra minus in Fig. 4(h). In general, it is quite easy to determine the overall sign using three rules. These are given and illustrated in Fig. 5.

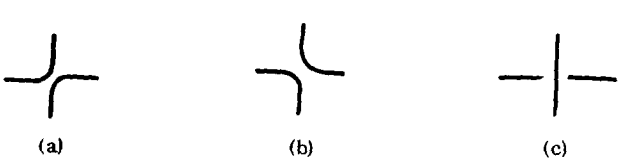


FIG. 3. Intersections. The double corners of Figs. (a) and (b) are not allowed by Eq. (3.3). When four lines meet at a site they must pass directly through as in Fig. (c).

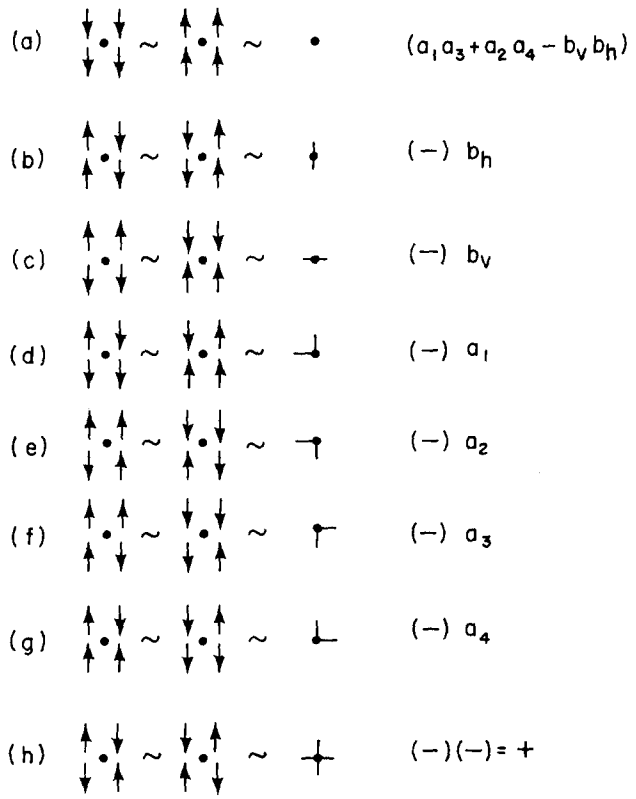


FIG. 4. The eight possible configurations that can occur at a site. When disorder variables are used [Eq. (3.2)], the first two columns represent corresponding spin configurations. In obtaining the weights of column 4 a (-1) factor has been included from the $(-1)^N$ of Eq. (3.3). The minus signs in (b) through (g) may be eliminated because i) there are always an even number of (b) and (c) configurations and ii) corners (d) and (f) as well as (e) and (g) occur in pairs. Alternatively, one could redefine the b's and a's in Eq. (3.4) to have minus signs. Configuration (h) has an extra minus sign due to reordering of anticommuting variables as described in Appendix B. The numbers in column 4 are easily obtained: For example, the b_h of (b) is obtained because a vertical bond enters and exits the vertical site and a horizontal monomer with b_h must fill the empty horizontal site.

The Ising (and free-fermion) model has been solved. Eq. (3.4) represents the solution. It is trivial to compute the partition function (and correlation functions). Equation (3.4) is a translationally invariant quadratic action. One treats it as one does with any free field theory: go to momentum space via Fourier transform. This diagonalizes the problem. Going

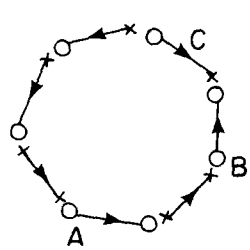


FIG. 5. Sign rules: The rules for evaluating the sign of a "dimer loop" are as follows: Pick an initial o or x (here, o is chosen at A) and proceed around the loop (here, counterclockwise). There is a) a minus sign for each x occurring before an o (the point, B), b) a minus sign for each arrow in the opposite direction (the bond, C), and finally c) a minus sign if one begins with an x. In this figure the sign is positive.

to momentum space means writing

$$\eta^{r}_{\alpha\beta} = \sum_{s,t} \frac{1}{(2M+1)^{1/2}} \frac{1}{(2N+1)^{1/2}} \times \exp\left(\frac{2\pi i \alpha s}{2M+1} + \frac{2\pi i \beta t}{2N+1}\right) a^{r}_{st}. \tag{3.5}$$

I will always choose α to range from $-M$ to M and β to range from $-N$ to N, sot that there are $(2N+1)$ rows and $(2M+1)$ columns. In the Ising model there are $(2N+1)(2M+1)$ sites. In Eq. (3.5) a^{r}_{st} are an equivalent set of anticommuting variables; s ranges from $-M$ to M and t ranges from $-N$ to N. The determinant of this transformation is one. One should think in terms of the correspondence:

$$(\alpha,\beta) \leftrightarrow (x,y), \tag{3.6}$$

$$\left(\frac{2\pi s}{2M+1}, \frac{2\pi t}{2N+1}\right) \leftrightarrow (p_x, p_y).$$

The variables s and t are simply momentum variables. Equation (3.5) implies periodic boundary conditons. These conditions will always be chosen, so that one is working on a torus.[17]

In momentum space the action of Eq. (3.4) becomes

$$\begin{aligned} A_{\text{free fermion}} = \sum_{s,t} \Big[& z_{\mathrm{h}} a^{\mathrm{h}^x}_{st} a^{\mathrm{h}''}_{st} \exp\left(\frac{2\pi i s}{2M+1}\right) \\ & + z_{\mathrm{v}} a^{\mathrm{v}^x}_{st} a^{\mathrm{v}^c}_{st} \exp\left(\frac{2\pi i t}{2N+1}\right) \\ & + a_1 a^{\mathrm{h}^x}_{st} a^{\mathrm{v}''}_{st} + a_3 a^{\mathrm{v}^x}_{st} a^{\mathrm{h}''}_{st} \\ & + a_2 a^{\mathrm{v}^x}_{st} a^{\mathrm{h}^x}_{-s-t} + a_4 a^{\mathrm{v}''}_{st} a^{\mathrm{h}''}_{-s-t} \\ & + b_{\mathrm{h}} a^{\mathrm{h}''}_{st} a^{\mathrm{h}^x}_{st} + b_{\mathrm{v}} a^{\mathrm{v}''}_{st} a^{\mathrm{v}^x}_{st} \Big]. \end{aligned} \tag{3.7}$$

Only (s,t) and $(-s,-t)$ variables are coupled. The integrations can explicitly be done using the definition in Section II:

$$\begin{aligned} & L\left(\frac{2\pi s}{2M+1}, \frac{2\pi t}{2N+1}\right) \\ & \quad = \mathrm{h}_s \mathrm{h}_{-s} \mathrm{v}_t \mathrm{v}_{-t} - a_1 a_3 (\mathrm{h}_s \mathrm{v}_t + \mathrm{h}_{-s} \mathrm{v}_{-t}) \\ & \qquad - a_2 a_4 (\mathrm{h}_s \mathrm{v}_{-t} + \mathrm{h}_{-s} \mathrm{v}_t) + (a_1 a_3 + a_2 a_4)^2, \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} \mathrm{h}_s &= b_{\mathrm{h}} - z_{\mathrm{h}} \exp\left(\frac{2\pi i s}{2M+1}\right), \\ \mathrm{v}_t &= \mathrm{h}_{\mathrm{v}} - z_{\mathrm{v}} \exp\left(\frac{2\pi i t}{2N+1}\right). \end{aligned} \tag{3.9}$$

The partition function is

$$Z_{\text{free fermion}} = \left(\prod_{st} L(s,t)\right)^{1/2}, \tag{3.10}$$

which becomes in the thermodynamic limit

$$-\beta f_{\text{free fermion}} = \frac{1}{2} \int_{-\pi}^{\pi} \frac{dp_x}{2\pi} \int_{-\pi}^{\pi} \frac{dp_y}{2\pi} \ln L(p_x, p_y), \tag{3.11}$$

where L is given by Eq. (3.8). The factor of 1/2 is due to double counting of (s,t) and $(-s,-t)$. Equations (3.8) and (3.11) agree with the known result.[9,18]

For the Ising model set $a_i = b_{\mathrm{v}} = b_{\mathrm{h}} = -1$ to get the famous Onsager result[19]:

$$\begin{aligned} & -\beta f_{\text{Ising}} \\ & = \frac{1}{2} \int_{-\pi}^{\pi} \frac{dp_x}{2\pi} \int_{-\pi}^{\pi} \frac{dp_y}{2\pi} \ln 4 \left[\cosh 2\beta J_{\mathrm{v}} \cosh 2\beta J_{\mathrm{h}} \right. \\ & \quad \left. + \sinh 2\beta J_{\mathrm{h}} \cos p_x + \sinh 2\beta J_{\mathrm{v}} \cos p_y \right]. \end{aligned} \tag{3.12}$$

IV. SOLVABLE CLOSE-PACKED DIMER MODELS AND THE GRAPHICAL RULES

In a dimer problem[9,11,14] there are a set of sites and a set bonds connecting certain pairs of sites. The bonds may absorb dimers. There is a Boltzmann factor, z_b, associated with an absorption. A site may be used only once, so that no two dimers may overlap or even touch. Effectively any two dimers are infinitely repulsive. There are two kinds of problems: the close-packed problem in which every site must be covered exactly once, and the usual dimer problem (or dimer-monomer mixing) problem where some sites may be left uncovered. The statistical mechanics of these systems is determined by their partition functions. These partition functions may be represented as anticommuting variable integrals. In general, the action contains both quadratic and quartic terms meaning that the models are unsolvable interacting theories. The third paper attacks these unsolvable problems. This section considers solvable two-dimensional close-packed dimer problems. By solvable, I mean solvable by the usual Pfaffian methods.[11] The models will be translated into Grassmann integral form, from which a series of graphical rules will be derived. The treatment used here does not differ from the usual Pfaffian treatment. What is gained is a simple graphical approach which allows one to rapidly solve a dimer problem. Furthermore, the diagrammatic methods extend to any free-field-like theory. This section serves as an introduction to graphic methods.

I refer the reader to the standard method of solution.[11] There are two key points:

I. *Solvability Condition*: A planar dimer problem is solvable if its bonds may be oriented so that every elementary polygon is clockwise odd. Planar means it may be drawn on a piece of paper so that bonds do not cross. The bonds are then given an orientation. The direction is usually denoted by an arrow. A polygon is clockwise odd, if when traversing clockwise, one encounters an odd number of bonds oriented in the opposite direction. An elementary polygon is a non-self-intersecting polygon made up of bonds which has no bonds in its interior.

II. *The Method of Solution*: Fix a standard B configuration which covers the lattice. Each covering (these new ones will be called A coverings) when combined with the B configuration results in a set of closed polygons and isolated dimer pairs, the partition function of which has a Pfaffian representation.

Condition I and Observation II make the problem solvable by Pfaffian methods.

For every model satisfying I, the Method of Solution II can be translated into Grassmann integral form: A bond oriented from point, P, to point, Q, upon which on A-dimer may be placed corresponds to a term $\eta_P \eta_Q$ in the action. A stan-

dard B-bond between P and Q corresponds to a term $\eta_Q^\dagger \eta_P^\dagger$, A-dimer operators are ordered with the graph orientations, whereas B-dimer operators are ordered oppositely to the graph orientations. The action is schematically of the form

$$A_{\text{dimer}} = \sum_{A\text{-dimers}} z_A \eta\eta + \sum_{B\text{-dimers}} \eta^\dagger \eta^\dagger. \tag{4.1}$$

The Boltzmann factors of A-dimers are z_a, whereas B-dimers have unit Boltzmann factors. It is not hard to see that this action produces the closed polygons and isolated dimer pairs used in the Method of Solution II. The signs are all positive because of Condition I. This may be proved by employing Kasteleyn's theorem[20] which is easily proved by induction on the length of a polygon and says that the above polygonal configurations are all clockwise odd. Associate an undaggered variable with an o and a daggered variable with an x and use the sign rules of Fig. 5. Let $2n$ be the number of edges (it must always be even). If B-dimers were oriented with the graphical orientation Kasteleyn's theorem would give a minus one for rule (b). Instead there is a $(-1)^{n+1}$ because the n B-dimers are oriented oppositely. If one begins with an initial o then there are $n-1$ x's which occur before o's. So rule (a) gives $(-1)^{n-1}$. Rules (a) and (b) combine to give plus for the overall sign.

Some dimer models satisfy

Simplying Condition (C): A graph satisfies Simplifying Condition C if vertices can be grouped into two sets (which I call odd and even) such that no two odd (or even) vertices have a bond in common.

When this condition is satisfied, transform $\eta \to \eta^\dagger$ and $\eta^\dagger \to \eta$ at all even sites. This makes the bilinears in the action of the form $\eta\eta^\dagger$, the partition function becomes a product of determinants rather than Pfaffians, the graphical rules simplify, and calculations are easier to do. The rules will be illustrated using the dimer model on a square quadratic lattice.

Graphical Rules When Condition C *Holds*: or *Rules When Bilinears Are of* $\eta\eta^\dagger$ *Form*:

1. Group vertices into repeating units that fill a square array. Use (α,β) to label the units and use $r = 1,2,3,\ldots,T$ to label the different vertices within a unit.

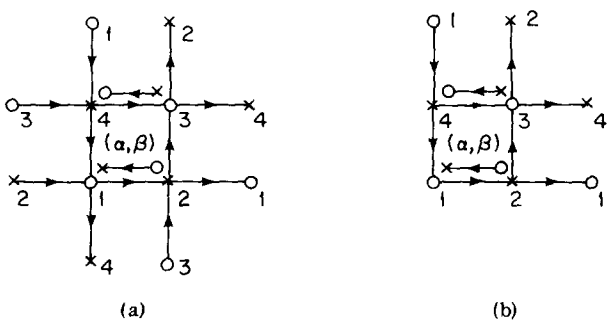


FIG. 6. Illustration of Rule 2: Figure (a) shows the (α,β) unit. There are two B-dimers and four A-dimers entirely contained in (α,β). There are eight A-dimers which connect sites in (α,β) to sites in nearby units. They occur in pairs. For example, the upper right A-dimer, $\eta^3_{\alpha\beta}\eta^{2\dagger}_{\alpha\beta+1}$, has a partner, the lower right A-dimer, $\eta^3_{\alpha\beta-1}\eta^{2\dagger}_{\alpha\beta}$. Rule 2 erases one bond from each pair. Figure (b) is an example of what results.

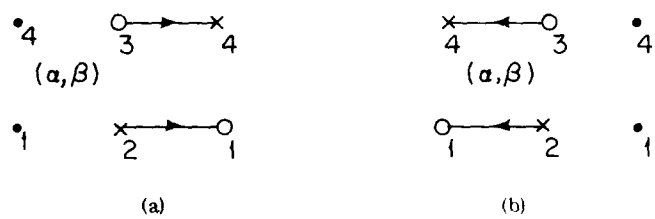


FIG. 7. Rule 3 for $\eta\eta\dagger$ products: Figure (a) shows the two dimers of Fig. 6(b) which start in the (α,β) unit at sites 2 and 3 and go to the sites 1 and 4 of the $(\alpha+1,\beta)$ unit. Rule 3 says to "fold" these back into the (α,β) unit as shown in (b). Let o and x correspond to the anticommuting variables a and $a\dagger$. The the $a_2^\dagger a_1$ bond weight gets multiplied by $\exp(ip_x)$ whereas the $a_3 a_4^\dagger$ weight gets multiplied by $\exp(-ip_x)$.

2. Consider one unit, U. There are two kinds of bonds: (a) those which are contained within U and (b) those which go from U to some other unit. Of the latter, [(b)], for every bond which goes from a type r vertex in U to type q vertex in another unit, there is one bond which goes from a type r vertex in another unit to a type q vertex in U. Thus, they occur in pairs. Half are to be included in U and the others ignored and erased. Figure 6 illustrates this for the square lattice.

3. Keep (a) type bonds as they are. For a (b) type bond which goes from an r in U to a q in another unit, "fold" it back into U, so that it goes from r to q within U (see Fig. 7). If q is an o located in a unit m horizontal spaces to the right and n spaces upward (m and n may be negative) multiply the bond weight by

$$\exp(imp_x + inp_y). \tag{4.2}$$

If q is an x multiply the bond weight by the complex conjugate of Eq. (4.2), that is

$$\exp(-imp_x - inp_y). \tag{4.3}$$

Figure 7 illustrates this. Figure 8 shows all the weights in the square lattice example after Rule 3 has been carried out.

4. Rules 1 through 3 result in a miniature dimer problem. Solve it by finding all coverings and their weights (see Fig. 9 for the square lattice). Call the sum of the diagrams $L(p_x,p_y)$. The free energy per site, f, is

$$-\beta f = \frac{1}{T}\int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\ln L(p_x,p_y). \tag{4.4}$$

The factor of $1/T$ occurs because there are T sites per unit.

Graphical Rules When Condition C *Fails*: or *Rules When Bilinears Are of* $\eta\eta$ *and* $\eta^\dagger\eta^\dagger$ *Form*: These rules will be exemplified by treating the square lattice dimer problem.

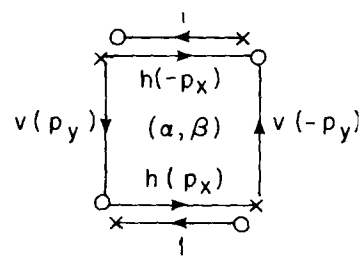


FIG. 8. The weights for the square lattice: Rule 3 applied to Fig. 6(b) results in this figure. The weights of th B-dimers remains 1 as indicated. The A-dimer weights have contributions from (a) type bonds as well as (b) types. When added they result in the factors $h(p_x) = z_h[1-\exp(ip_x)]$, $v(p_y) = z_v[1-\exp(ip_y)]$, etc.

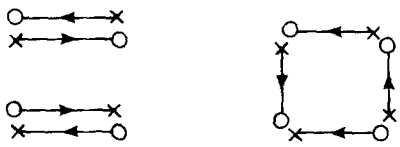


FIG. 9. The two coverings of Fig. 8: The value of (a) is $h(p_x)h(-p_x) = z_h^2(2-2\cos p_x)$. The value of (b) $v(p_y)v(-p_y) = z_v^2(2-2\cos p_y)$. The sum of these is $L(p_x,p_y)$. When put into Eq. (4.4), the free energy per site is obtained.

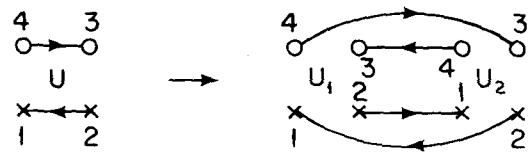


FIG. 10. The (a)-type bonds: In Fig. (a), there is an A-dimer and a B-dimer. Each of these result in two dimers, one from U_1 to U_2 and one from U_2 and U_1 as (b) indicates. The orientation remains the same, so that the A-dimer in U which goes from 4 to 3, still goes from 4 to 3 in both cases in Fig. (b).

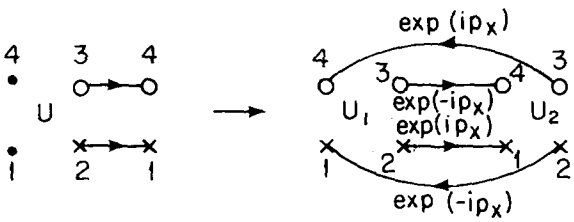


FIG. 11. The (b)-type bonds: Fig. (a) shows one $\eta\eta$ (b)-type bond and one $\eta^+\eta^+$ (b)-type bond. If U is the (α,β) unit then the two bonds go from the (α,β) unit to the $(\alpha+1,\beta)$ unit. Both give rise to two dimers in (b) the weights of which get multiplied by the indicated phase factors.

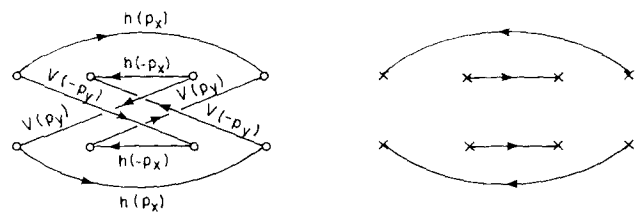


FIG. 12. The resulting bond weights: Figure (a) shows the resulting A-dimers and their bond weights. Figure (b) shows the B-dimers. Their weights are all unity. Here, $h(p_x) = z_h[1-\exp(ip_x)]$ and $v(p_y) = z_v[1-\exp(ip_y)]$. When superimposed (a) and (b) give rise to a miniature dimer problem.

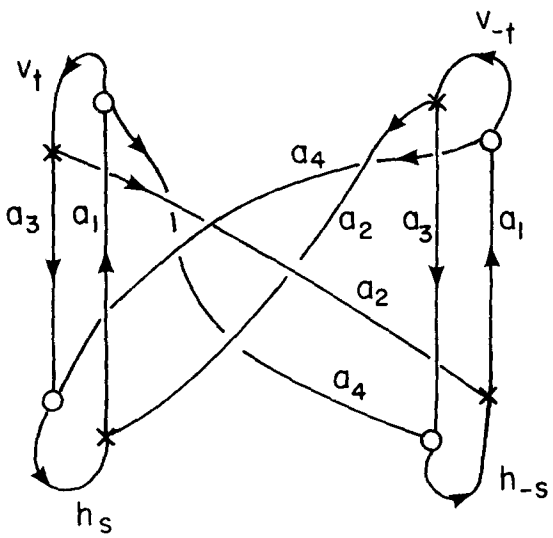


FIG. 13. The miniature dimer problem for the free-Fermion model: The upper left o and x are $a^{v'''}_{xt}$, $a^{v'}_{xt}$; the lower left are $a^{h'}_{xt}$, $a^{h''}_{xt}$; the upper right are $a^{v'''}_{x\ -t}$, $a^{v'}_{x\ -t}$; the lower right are $a^{h'''}_{x\ -t}$, $a^{h'}_{x\ -t}$. The weights of bonds are as indicated with h_x and v_t given by Eq. (3.9).

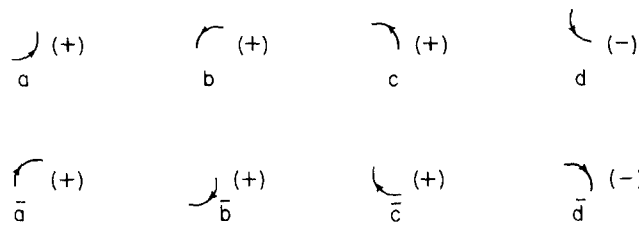


FIG. 14. The eight oriented corners and the minus sign factors associated with them.

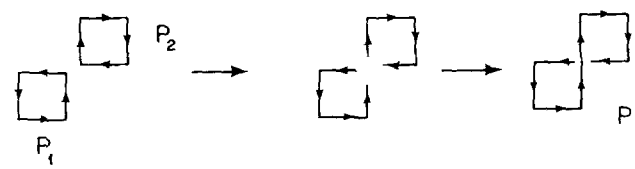


FIG. 15. The pasting construction: Polygon, P, may be obtained from two (possibly self-intersecting) polygons, P_1 and P_2, by cutting open the corners and rejoining. There are four (two different types of pairs of corners times two orientations) possible pasting constructions.

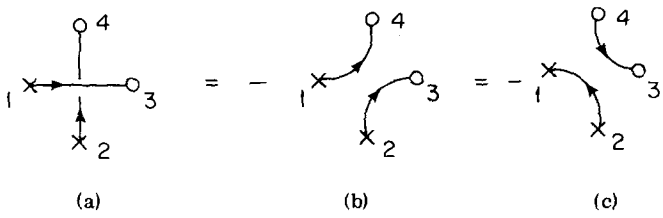


FIG. 16. How the minus sign arises: This is just a "fermion" statistics effect. The order of operators in an intersection of P is indicated in Fig. (a) and is $(\eta_1^x\eta_3^o)(\eta_2^x\eta_4^o)$. When P is decomposed into nonintersecting polygons are in Fig. 16, the order of the operators is that of (b) or (c). For case (b), $(\eta_1^x\eta_4^o)(\eta_2^x\eta_3^o) = -(\eta_1^x\eta_3^o)(\eta_2^x\eta_4^o)$, that is, there is a minus sign relative to (a). For case (c), $(\eta_2^x\eta_1^x)(\eta_4^o\eta_3^o)$ is also $-(\eta_1^x\eta_3^o)(\eta_2^x\eta_4^o)$.

Although Condition C is satisfied, the simplifying transformation will not be performed.

1. Same as above.

2. Same as above.

3. Draw two copies of U. Call them U_1 and U_2. For (a) type bonds going from r to q draw two lines: one from r in U_1 to q in U_2 and one from r in U_2 to q in U_1 (see Fig. 10). For $\eta\eta$ dimers (i.e., A-dimers) of (b) type originating at an r in U and terminating at a q in another unit, again draw two lines. First draw one from r in U_1 to q in U_2 and multiply its weight by $\exp(-imp_x - inp_y)$, then draw one from r in U_2 to q in U_1 and multiply its weight by $\exp(imp_x + inp_y)$ (see Fig. 11). For $\eta^\dagger\eta^\dagger$ dimers (i.e., B-dimers) do the same as for $\eta\eta$ dimers but multiply weights by the complex conjugated phase factors of the $\eta\eta$ case (see Fig. 11). In all cases, if bonds are oriented from r to q they remain so, regardless of whether they go from U_1 to U_2 or U_2 to U_1. Figure 12 shows the resulting weights for the square lattice.

4. Solve the miniature dimer problem and call the result $L(p_x,p_y)$. The free energy per unit site is

$$-\beta f = \frac{1}{2T}\int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\ln L(p_x,p_y). \tag{4.5}$$

Graphical Rules For A General Quadratic Theory: In general, there will be $\eta\eta^\dagger$, $\eta\eta$, and $\eta^\dagger\eta^\dagger$ products. Two copies, U_1 and U_2, of U are to be drawn. Follow the second set of rules, 1, 2, 3, for $\eta\eta$ and $\eta^\dagger\eta^\dagger$ products. For $\eta\eta^\dagger$ terms use Rule 3 of the first set for the U_1 copy of U but for U_2 use complex conjugated phase factors. Finally, use Eq. (4.5) and rule 4. Figure 13 shows the miniature dimer problem for the free-fermion action in Eq. (3.4). The coverings are easily summed to give the function in Eq. (3.8).

V. CONCLUSION

The novel approach of this paper provides the best means of solving the two-dimensional Ising model, the free-fermion eight vertex models, and the planar close-packed dimer problems.

APPENDIX

In this Appendix, I will analyze the sign problem associated with Eqs. (3.3) and (3.4).The conclusion will be that the sign of a configuration of polygons is equal to the number of intersections which occur. This explains the extra minus factor in the weight of Fig.4 (h),I will proceed in steps: first dealing with an isolated non-self-intersecting polygon, then with one that self-intersects, and finally dealing with a multipolygonal configuration.

Consider a closed polygon, P, which does not intersect itself. I will show that its sign is positive. Choose a horizontal bond of P and proceed to the right (and eventually around the polygon). Start at the x and use the rules of Fig. 5. When moving upward or to the right no minus signs result from rules (a) or (b) because arrows are in the correct direction and o's occur before x's. When moving downward or to the left, each site has a minus sign from rule (a) and a minus sign from rule (b). They cancel in pairs. Next consider what happens, when one goes around a corner. There are eight different types (see Fig. 14) (two orientations times the four basic corners of Fig. 2). They are oriented because we are moving around the polygon in a particular direction. Figure 14 summarizes the results: only corners of types d and $\bar{d}$ lead to a minus sign. Now use the following theorem (which is easily proved by induction on the area of P): Let m_a, m_b, etc. be the number of type a, type b, etc. corners occurring in an oriented non-self-intersecting polygon, P. If P is counterclockwise oriented then

$$\begin{aligned} m_a - m_{\bar{a}} &= 1, \\ m_b - m_{\bar{b}} &= 1, \\ m_c - m_{\bar{c}} &= 1, \\ m_d - m_{\bar{d}} &= 1. \end{aligned} \tag{A1}$$

This implies that the sign due to corners is $(-1)^{m_d}(-1)^{m_{\bar{d}}} = -1$. For clockwise oriented, P, the theorem holds with $a \leftrightarrow \bar{a}$, $b \leftrightarrow \bar{b}$, etc. Rules (a) and (b) therefore result in one minus sign which when combined with the minus sign of rule (c) gives an overall plus sign.

Now consider an oriented self-intersecting polygon, P. It may be constructed from nonintersecting ones by the pasting construction of Fig. 15. The order of the operators in P is indicated in Fig. 16(a). When they are regrouped into the forms occurring in the non-self-intersecting polygons [Figs. 16(b) and 16(c)] which "compose" P, a minus sign results for each intersection as Fig. 16 illustrates.

Finally, the result holds for multipolygonal configurations because pairs of polygons can only intersect an even number of times. Summarizing, an extra minus occurs for each intersection [Fig. 4(h)].

[1] S. Samuel, Phys. Rev. D **18**, 1916 (1978).
[2] D. J. Chandlin, Nuovo Cimento **4**, 231 (1956).
[3] An exception is S. Samuel, J. Math. Phys. **19**, 1438 (1978).
[4] S. Samuel, "The Use of Anticommuting Integrals in Statistical Mechanics III," LBL preprint 9347 (June 1979).
[5] S. Samuel, "The Pseudo-Free 128 Vertex Model," to be published in J. Phys. A.
[6] S. Samuel, "The Correlation Functions in the 32-Vertex Model, IAS preprint (March 1980).
[7] S. Samuel, papers in progress.
[8] T. D. Schultz, D. C. Mattis, and E. H. Lieb, Rev. Mod. Phys. **36**, 856 (1964).
[9] H. S. Green and C. A. Hurst, *Order-Disorder Phenomena* (Interscience,

New York, 1964).
[10]C. A. Hurst, J. Math. Phys. 7, 305 (1966).
[11]See, for example, E. W. Montroll, *Brandeis University Summer Institute in Theoretical Physics, 1966*, edited by M. Chretien, E. P. Gross, and S. Deser (Gordon and Breach, New York, 1968); E. W. Montroll, Chap. IV in *Applied Combinatorial Mathematics*, edited by E. F. Beckenbach (Wiley, New York, 1964).
[12]F. A. Berezin, *The Method of Second Quantization* (Academic, New York, 1966). See also the Appendices of Ref. 3.
[13]The definition of a Pfaffian is

$$\mathrm{Pf}A = \frac{1}{2^{N/2}} \frac{1}{(N/2)!} \sum_{\sigma} (\mathrm{sign}\sigma) A_{\sigma(1)\sigma(2)} A_{\sigma(3)\sigma(4)} \cdots A_{\sigma(N-1)\sigma(N)}.$$

The sum if over all permutations, σ.
[14]For example, B. M. McCoy and T. T. Wu, *The Two-Dimensional Ising Model* (Harvard U. P., Cambridge, 1973). The proof that $(\mathrm{Pf}A)^2 = \det A$ can be found on pages 47–51.
[15]Thre is an emormous amount of literature on the Ising model. See the references of Ref.14 for a partial list.
[16]See for example, pages 218–221 of Ref. 11.
[17]When this is done, some terms have the incorrect sign (those involving loops around the torus). This difficulty can be overcome by standard methods (see pages 61–67 of Ref. 14). However, in the thermodynamic limit, such configurations will be zero measure effect and can be ignored.
[18]C. Fan and F. Y. Wu, Phys. Rev. B **2**, 723 (1970).
[19]L. Onsager, Phys. Rev. **65**, 117 (1964). Textbook derivations are given in Refs. 9, 11, and 14.
[20]P. M. Kasteleyn, J. Math. Phys. **4**, 287 (1963). A pedagogical version given in Ref. 11.

The use of anticommuting variable integrals in statistical mechanics. II. The computation of correlation functions [a)]

Stuart Samuel
Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720 and Institute for Advanced Study, Princeton, New Jersey 08540

(Received 16 July 1980; accepted for publication 30 July 1980)

By using integrals over anticommuting variables all the correlation functions in the two-dimensional Ising model and free-fermion eight vertex model are computed. The method is quite general and applicable to other solvable systems.

I. INTRODUCTION

Paper I represented several models as fermionic functional integrals with quadratic actions.[1] As such they are exactly solvable by free-field-theory-like methods. Paper I computed the partition functions and established a simple set of computational rules. A simple extension of free field theory methods yields all correlation functions. The purpose of this paper is to do these computations. For the Ising model and for the free-fermion eight vertex model this is done in Secs. III and IV. For the latter model, this is the first time all correlation functions have been computed. This might seem quite a task since the model is quite general with six independent parameters. The method has actually been used in even more complicated models[2] and can be adapted to any model solvable via anticommuting variables. In fact, there is a simple three step procedure: First, determine anticommuting variable correlation functions in momentum space. This is done using free field theory methods. Next, determine them in coordinate space via Fourier transform. Finally, relate physical correlation functions to anticommuting variable ones. Steps two and three are in Sec. II. For the Ising model (respectively eight vertex model) step three is done in Sec. III (respectively Sec. IV). The result will always be a Pfaffian. If physical variables are related in a complicated way to anticommuting variables then the Pfaffian can become of cumbersome size. This happens with the Ising model. This is not the case for the eight vertex model, where, for example, all two point correlations are Pfaffians of 8×8 matrices. Only higher point correlations are Pfaffians of large order (n point is a $4n\times 4n$ Pfaffian).

II. ANTICOMMUTING VARIABLE CORRELATIONS

This section will compute the anticommuting variable correlations (or "propagators") for the free fermion model [Eq. (I. 3.4)]. The configurations and their weights were given in Fig. I. 4. In addition, there are z_h and z_v Boltzmann factors for each unit of horizontal and vertical Bloch wall.

The correlation functions will first be calculated in momentum space and then in coordinate space. This can be done using free field theory methods or it can be done graphically as was done with the partition function in Sec. IV of I. One obtains a miniature dimer problem with one fixed bond. Space limitation prevents us from describing the method.[3] The results are the following: The nonzero momentum space correlation functions are

$$\langle a^{h^o}_{st} a^{h^x}_{st}\rangle = (h_{-s} v_t v_{-t} - a_1 a_3 v_t - a_2 a_4 v_{-t})/D(s,t), \tag{2.1}$$

$$\langle a^{v^o}_{st} a^{v^x}_{st}\rangle = (h_s h_{-s} v_{-t} - a_1 a_3 h_s - a_2 a_4 h_{-s})/D(s,t), \tag{2.2}$$

$$\langle a^{h^o}_{st} a^{v^x}_{st}\rangle = a_1[h_{-s} v_{-t} - (a_1 a_3 + a_2 a_4)]/D(s,t), \tag{2.3}$$

$$\langle a^{v^o}_{st} a^{h^x}_{st}\rangle = a_3[h_{-s} v_{-t} - (a_1 a_3 + a_2 a_4)]/D(s,t), \tag{2.4}$$

$$\langle a^{h^o}_{st} a^{h^o}_{-s-t}\rangle = a_1 a_2 (v_t - v_{-t})/D(s,t), \tag{2.5}$$

$$\langle a^{v^o}_{st} a^{v^o}_{-s-t}\rangle = a_2 a_3 (h_{-s} - h_s)/D(s,t), \tag{2.6}$$

$$\langle a^{v^o}_{st} a^{h^o}_{-s-t}\rangle = a_2[(a_1 a_3 + a_2 a_4) - h_s v_{-t}]/D(s,t), \tag{2.7}$$

$$\langle a^{h^x}_{st} a^{h^x}_{-s-t}\rangle = a_3 a_4 (v_t - v_{-t})/D(s,t), \tag{2.8}$$

$$\langle a^{v^x}_{st} a^{v^x}_{-s-t}\rangle = a_1 a_4 (h_{-s} - h_s)/D(s,t), \tag{2.9}$$

$$\langle a^{v^x}_{st} a^{h^x}_{-s-t}\rangle = a_4[(a_1 a_3 + a_2 a_4) - h_s v_{-t}]/D(s,t), \tag{2.10}$$

where h_s, v_t, and $D(s,t) \equiv L(p_x,p_y)$ are given by Eqs. (I. 3.8) and (I. 3.9). Of course, correlations involving (s,t) and (s',t') variables vanish if neither $(s,t) \neq (s',t')$ nor $(s,t) \neq (-s',-t')$.

To obtain coordinate space correlations, use Eq. (I. 3.5) to express η's in terms of a's, and then use the results of Eqs. (2.1)–(2.10). The thermodynamic limit can be taken and the correlations are

$$\langle \eta^{h^o}_{\alpha\beta} \eta^{h^x}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi} \frac{dp_x}{2\pi} \int_{-\pi}^{\pi} \frac{dp_y}{2\pi} \exp[i(\alpha-\alpha')p_x + i(\beta-\beta')p_y]\,[h(-p_x)v(p_y)v(-p_y) - a_1 a_3 v(p_y) - a_2 a_4 v(-p_y)]/L(p_x,p_{yz}), \tag{2.11}$$

$$\langle \eta^{v^o}_{\alpha\beta} \eta^{v^x}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi} \frac{dp_x}{2\pi} \int_{-\pi}^{\pi} \frac{dp_y}{2\pi} \exp[i(\alpha-\alpha')p_x + i(\beta-\beta')p_y]\,[h(p_x)h(-p_x)v(-p_y) - a_1 a_3 h(p_x) - a_2 a_4 h(-p_x)]/L(p_x,p_y), \tag{2.12}$$

[a)]Work has been supported by the High Energy Division of the United States Department of Energy.

 0022-2488/80/122815-05$1.00

$$\langle \eta^{h^o}_{\alpha\beta}\eta^{v^x}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\exp[i(\alpha-\alpha')p_x + i(\beta-\beta')p_y]a_1[h(-p_x)v(-p_y) - (a_1a_3+a_2a_4)]/L(p_x,p_y), \tag{2.13}$$

$$\langle \eta^{v^o}_{\alpha\beta}\eta^{v^x}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\exp[i(\alpha-\alpha')p_x + i(\beta-\beta')p_y]a_3[h(-p_x)v(-p_y) - (a_1a_3+a_2a_4)]/L(p_x,p_y), \tag{2.14}$$

$$\langle \eta^{h^o}_{\alpha\beta}\eta^{h^o}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\exp[i(\alpha-\alpha')p_x + i(\beta-\beta')p_y]a_1a_2[v(p_y) - v(-p_y)]/L(p_x,p_y), \tag{2.15}$$

$$\langle \eta^{v^o}_{\alpha\beta}\eta^{v^o}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi} = \exp[i(\alpha-\alpha')p_x + i(\beta-\beta')p_y]a_2a_3[h(-p_x) - h(p_x)]/L(p_x,p_y), \tag{2.16}$$

$$\langle \eta^{v^o}_{\alpha\beta}\eta^{h^o}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\exp[i(\alpha-\alpha')p_x + i(\beta-\beta')p_y]a_2[(a_1a_3+a_2a_4) - h(p_x)v(-p_y)]/L(p_x,p_y), \tag{2.17}$$

$$\langle \eta^{h^x}_{\alpha\beta}\eta^{h^x}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\exp[i(\alpha'-\alpha)p_x + i(\beta'-\beta)p_y]a_3a_4[v(p_y) - v(-p_y)]/L(p_x,p_y), \tag{2.18}$$

$$\langle \eta^{v^x}_{\alpha\beta}\eta^{v^x}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\exp[i(\alpha'-\alpha)p_x + i(\beta'-\beta)p_y]a_1a_4[h(-p_x) - h(p_x)]/L(p_x,p_y), \tag{2.19}$$

$$\langle \eta^{v^x}_{\alpha\beta}\eta^{h^x}_{\alpha'\beta'}\rangle = \int_{-\pi}^{\pi}\frac{dp_x}{2\pi}\int_{-\pi}^{\pi}\frac{dp_y}{2\pi}\exp[i(\alpha'-\alpha)p_x + i(\beta'-\beta)p_y]a_4[(a_1a_3+a_2a_4) - h(p_x)v(-p_y)]/L(p_x,p_y), \tag{2.20}$$

where

$$h(p_x) = b_h - z_h\exp(ip_x),$$
$$v(p_y) = b_v - z_v\exp(ip_y), \tag{2.21}$$

and L is given by Eq. (I.3.8). Equations (2.11)–(2.20) are respectively obtained from Eqs. (2.1)–(2.10) by replacing h_s and v_t by the corresponding momentum valued functions of Eq. (2.21). The factors $\exp[i(\alpha-\alpha')p_x]$ and $\exp[i(\beta-\beta')p_y]$ in Eqs. (2.11)–(2.20) are translation operators. Equations (2.18)–(2.20) have conjugated translation factors.

Equations (2.11)–(2.20) are the coordinate-space anticommuting variable correlation functions for the free-fermion model.

III. THE ISING MODEL CORRELATION FUNCTIONS

This section will calculate the correlation function of two spin variables in the same row. It will be compared to the known result as a check on anticommuting variable techniques. Two horizontal spins are chosen for illustrative purposes only. The approach extends to an arbitrary pair; in fact, the vacuum expectation value of several σ's can be computed. The only drawback is the cumbersome form of the answer: a Pfaffian of (in general) large size. In short, everything you ever wanted to know about the Ising model is expressible as a Pfaffian.

We will need the free fermion anticommuting variable correlations [Eqs. (2.11)–(2.20)]. Bond variables will be used, in which case the Ising model is related to the free-fermion (or closed-polygon) partition function

$$z_h = \tanh\beta J_h, \quad z_v = \tanh\beta J_v,$$
$$a_1 = a_2 = a_3 = a_4 = b_v = b_h = -1. \tag{3.1}$$

The weights of configurations are given in Fig. I.4. These values must be used (as opposed to the less restrictive conditions $a_1a_3 = a_2a_4 = b_v^2 = b_h^2 = 1$) because correlation functions, unlike the partition function, need not have the same number of a_1 and a_3 type corners, a_2 and a_4 type corners, etc.

This is obvious from Eqs. (2.11)–(2.20) where correlations are not simply functions of a_1a_3, a_2a_4, etc.

Spin variable correlation functions can be considered as partition functions on a defective lattice.[4,5] I refer the reader to Ref. 5, p. 248–257. This means that spin correlations are (up to multiplicative constants) the partition functions of Ising models with modified Bloch wall Boltzmann factors along selected paths. For example, $Z_{\text{Ising}}\langle\sigma_{1,0}\sigma_{m+1,0}\rangle$ is z_h^m times the Ising model with the usual z_h and z_v Boltzmann factors for all Bloch walls except for the horizontal ones between (1,0) and $(m+1,0)$, where z_h^{-1} is the Boltzmann factor. This defective lattice partition function is obtained by replacing

$$\exp\Big(\sum_{\alpha=1}^{m} z_h\eta^{h^x}_{\alpha0}\eta^{h^o}_{\alpha+10}\Big)$$

by

$$\exp\Big[\sum_{\alpha=1}^{m} z_h\eta^{h^x}_{\alpha0}\eta^{h^o}_{\alpha+10} + \sum_{\alpha=1}^{m}(z_h^{-1}-z_h)\eta^{h^x}_{\alpha0}\eta^{h^o}_{\alpha+10}\Big]$$
$$= \exp\Big(\sum_{\alpha=1}^{m} z_h\eta^{h^x}_{\alpha0}\eta^{h^o}_{\alpha+10}\Big)$$
$$\times\prod_{\alpha=1}^{m}[1+(z_h^{-1}-z_h)\eta^{h^x}_{\alpha0}\eta^{h^o}_{\alpha+10}],$$

so that

$$\langle\sigma_{1,0}\sigma_{m+1,0}\rangle = \Big\langle\prod_{\alpha=1}^{m}[z_h + (1-z_h^2)\eta^{h^x}_{\alpha0}\eta^{h^o}_{\alpha+10}]\Big\rangle. \tag{3.2}$$

Equation (3.2) typifies how spin variable correlations are related to anticommuting variable correlations. Equation (3.2) can be generalized to the case when the left-hand side is the vacuum expectation value of several σ's.

For free theories, the following formulas are useful:

$$\langle\eta_1\eta_2\cdots\eta_m\rangle = \mathrm{Pf}M_{ij} \quad (\text{for } m \text{ even}), \tag{3.3}$$

where

$$M_{ij} = \langle\eta_i\eta_j\rangle. \tag{3.4}$$

If $\langle\eta_i\eta_j\rangle = \langle\eta_i^\dagger\eta_j^\dagger\rangle = 0$, then

$$\langle \eta_1^\dagger \eta_1 \eta_2^\dagger \eta_2 \cdots \eta_m^\dagger \eta_m \rangle = \det M_{ij}, \tag{3.5}$$

where

$$M_{ij} = \langle \eta_i^\dagger \eta_j \rangle. \tag{3.6}$$

These formulas are the analogs of Wick's expansion. In Eq. (3.3) one sums over all pairings of η's, the sign of which is determined by how many permutations are required to get the η's in paired form.

The vacuum expectation value of an arbitrary product of spins is expressible as a linear combination of anticommuting variable correlations. These vacuum expectation values can be computed using Eqs. (2.11)–(2.20) and Eq. (3.3). I will demonstrate this for two horizontal spins.

Equations (2.15) and (2.18) imply $\langle \eta_{\alpha 0}^{h''} \eta_{\beta 0}^{h''} \rangle = \langle \eta_{\alpha 0}^{h^x} \eta_{\beta 0}^{h^x} \rangle = 0$ for all α and β. Apply Eq. (3.5) to (3.2). The z_h term of $|z_h + (1 - z_h^2)\eta_{\alpha 0}^{h^x} \eta_{\alpha+1\,0}^{h''}|$ in Eq. (3.2) multiplies the same factor as the term in the Wick expansion obtained by contracting $\eta_{\alpha 0}^{h^x}$ with $\eta_{\alpha+1\,0}^{h''}$. Therefore

$$\langle \sigma_{1,0} \sigma_{m+1,0} \rangle = \det M_{ij}, \tag{3.7}$$

where

$$\begin{aligned} M_{ij} &= z_h \delta_{ij} + (1 - z_h^2)\langle \eta_{i0}^{h^x} \eta_{j+1\,0}^{h^o} \rangle \\ &= \int_{-\pi}^{\pi} \frac{dp_x}{2\pi} \int_{-\pi}^{\pi} \frac{dp_y}{2\pi} \exp[ip_x(j-i)] \\ &\quad \times \{z_h - (1 - z_h^2)\exp(ip_x)[h(-p_x)\, v(p_y) \\ &\quad \times v(-p_y) - v(p_y) - v(-p_y)]\}/L(p_x, p_y). \end{aligned} \tag{3.8}$$

In obtaining Eq. (3.8), Eq. (2.11) has been used. Equations (3.7) and (3.8) express the correlation function of two horizontal spins as a Toeplitz determinant, as is usually done and yields the correct result.[4,5]

To calculate the vacuum expectation value of a product of spin variables, proceed analogously. It will be equivalent to an Ising model on a defective lattice. When expressed in terms of anticommuting variables, it will result in an expression of the form

$$\begin{aligned} \langle \Pi \sigma\text{'s} \rangle = \langle (c_{12} + d_{12}\eta_1\eta_2)(c_{34} + d_{34}\eta_3\eta_4) \\ \cdots (c_{2m-1\,2m} + d_{2m-1\,2m}\eta_{2m-1}\eta_{2m}) \rangle. \end{aligned} \tag{3.9}$$

In Eq. (3.9) η_i denotes an anticommuting variable such as $\eta_{\alpha\beta}^{h''}$, $\eta_{\alpha\beta}^{h^x}$, $\eta_{\alpha\beta}^{v^o}$, or $\eta_{\alpha\beta}^{v^x}$. The variables $c_{i\,i+1}$ and $d_{i\,i+1}$ are constants determined by the defective lattice. For convenience write $d_{i\,i+1} = d_i d_{i+1}$; any values of d_i satisfying this will do. Wick's expansion along with Eq. (3.3) tells us that Eq. (3.9) is

$$\langle \Pi \sigma\text{'s} \rangle = \mathrm{Pf} M_{ij}, \tag{3.10}$$

where

$$M_{ij} = \begin{cases} d_i d_j \langle \eta_i \eta_j \rangle + \delta_{i+1\,j} c_{i\,i+1}, & i \text{ odd} \\ d_i d_j \langle \eta_i \eta_j \rangle - \delta_{i-1\,j} c_{i-1\,i}, & i \text{ even} \end{cases}. \tag{3.11}$$

The $\langle \eta\eta \rangle$ correlations are given in Eqs. (2.11)–(2.20).

All Ising model spin correlations may be easily calculated using the above method. The reason they result in such cumbersome expressions is the following: The variables which solve The Ising model are the η's. They might be called the mathematical variables because they represent it as a free field theory. Correlation functions of anticommuting variables are simple to compute. Contrast this with the spin variables. They are the physical variables. They are, however, complicated functions of the mathematical variables, the η's, which means that spin variable computations result in cumbersome expressions. In conclusion, there are two types of variables, spin variables which have a simple physical interpretation but are mathematically awkward to work with and η variables which do not have as simple a physical interpretation but are easy to work with mathematically.

IV. THE FREE-FERMION EIGHT VERTEX MODEL CORRELATION FUNCTIONS

Once a model is solved via the anticommuting variable method, it is straightforward to compute anticommuting variable correlation functions. It is then possible (in practically all cases) to compute physical correlation functions. This was demonstrated for the Ising model in Sec. III.

This section calculates all the vertex correlation functions for the free-fermion model described by Eq. (I.3.4) and Fig. I.4 of paper I. It is just a simple extension of the methods used in Sec. III. The answer is expressed in terms of a Pfaffian of (in general) a large matrix. A few simple examples are worked out [see Eqs. (4.2), (4.4), (4.14), (4.16), and (4.25)]. The main result is a set of computational rules. By blindly following them, all vertex correlation functions can be calculated.

In Sec. III Ising model spin correlation functions were calculated. It is just as easy to calculate vertex correlation functions in the free-fermion model [Eq. (I.3.4) and Fig. I.4]. Equations (2.11)–(2.20) are all that is needed.

Let

$$\begin{aligned} B_{\alpha+(1/2)\beta} &= z_h \eta_{\alpha\beta}^{h^x} \eta_{\alpha+1\beta}^{h^o}, \\ B_{\alpha\beta+1/2} &= z_v \eta_{\alpha\beta}^{v^x} \eta_{\alpha\beta+1}^{v^o}. \end{aligned} \tag{4.1}$$

$B_{\alpha+(1/2)\beta}$ represents the operator which produces a unit of horizontal wall between (α,β) and $(\alpha+1,\beta)$. Likewise $B_{\alpha\beta+(1/2)}$ produces a unit of vertical wall between (α,β) and $(\alpha,\beta+1)$. If B's are inserted in the integral of Eq. (I.3.3) then walls must occur where B's operate. A closed polygon partition function with contraints that walls be in certain places is obtained. Hence

$$\begin{aligned} \langle B_{\alpha+(1/2)\beta} \rangle &= \text{the probability that a wall occurs at } (\alpha+\tfrac{1}{2},\beta) \\ &= z_h \langle \eta_{\alpha\beta}^{h^x} \eta_{\alpha+1\beta}^{h^o} \rangle, \\ \langle B_{\alpha\beta+1/2} \rangle &= \text{the probability that a wall occurs at } (\alpha,\beta+\tfrac{1}{2}) \\ &= z_v \langle \eta_{\alpha\beta}^{v^x} \eta_{\alpha\beta+1}^{v^o} \rangle. \end{aligned} \tag{4.2}$$

Because Eqs. (2.11) and (2.12) have computed these anticommuting variable correlations, these probabilities are explicitly known. In general

$$\begin{aligned} &\langle B_{\alpha_1\beta_1} B_{\alpha_2\beta_2} \cdots B_{\alpha_m\beta_m} \rangle \\ &= \text{the probability that walls simultaneously occur} \\ &\text{at } (\alpha_1,\beta_1), (\alpha_2,\beta_2), \cdots, (\alpha_m,\beta_m). \end{aligned} \tag{4.3}$$

In Eq. (4.3) one of the indices α_i or β_i is half integer. To calculate (4.3) insert the definitions in Eq. (4.1), factor out the

z_h's and z_v's to obtain the expectation value of a product of $2m$ η's. Use Eq. (3.3) to express this as a Pfaffian of a matrix M. The elements of M are the anticommuting variable correlations given in Eqs. (2.11)–(2.20). The answer is just a $2m\times 2m$ dimensional Pfaffian (or an $2m\times 2m$ determinant since $(\mathrm{Pf}M)^2 = \det M$). For example, the probability that horizontal walls simultaneously occur at $(\alpha+\frac{1}{2},\beta)$ and $(\alpha'+\frac{1}{2},\beta')$ is

$$\begin{aligned}\langle B_{\alpha+(1/2)\beta}B_{\alpha'+(1/2)\beta'}\rangle = z_h^2\big[&\langle\eta^{h^x}_{\alpha\beta}\eta^{h^o}_{\alpha+1\beta}\rangle\langle\eta^{h^x}_{\alpha'\beta'}\eta^{h^o}_{\alpha'+1\beta'}\rangle\\ &+\langle\eta^{h^x}_{\alpha\beta}\eta^{h^o}_{\alpha'+1\beta'}\rangle\langle\eta^{h^o}_{\alpha+1\beta}\eta^{h^x}_{\alpha'\beta'}\rangle\\ &-\langle\eta^{h^x}_{\alpha\beta}\eta^{h^x}_{\alpha'\beta'}\rangle\langle\eta^{h^o}_{\alpha+1\beta}\eta^{h^o}_{\alpha'+1\beta'}\rangle\big].\end{aligned} \tag{4.4}$$

The quantities on the right-hand side of Eq. (4.4) are given in Eqs. (2.11), (2.15), and (2.18).

A different set of questions can be asked, such as what is the probability that one of the configurations in Fig. I.4 occurs at (α,β). Define

$$O^{(a)}_{\alpha\beta} = (b_h b_v - a_1a_3 - a_2a_4)\eta^{h''}_{\alpha\beta}\eta^{h^x}_{\alpha\beta}\eta^{v''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta}, \tag{4.5}$$

$$O^{(b)}_{\alpha\beta} = b_h\eta^{h''}_{\alpha\beta}\eta^{h^x}_{\alpha\beta}(1-b_v\eta^{v''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta}), \tag{4.6}$$

$$O^{(c)}_{\alpha\beta} = b_v\eta^{v''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta}(1-b_h\eta^{h''}_{\alpha\beta}\eta^{h^x}_{\alpha\beta}), \tag{4.7}$$

$$O^{(d)}_{\alpha\beta} = a_1\eta^{h^x}_{\alpha\beta}\eta^{v''}_{\alpha\beta}(1-a_3\eta^{v^x}_{\alpha\beta}\eta^{h''}_{\alpha\beta}), \tag{4.8}$$

$$O^{(e)}_{\alpha\beta} = a_2\eta^{v^x}_{\alpha\beta}\eta^{h^x}_{\alpha\beta}(1-a_4\eta^{v''}_{\alpha\beta}\eta^{h''}_{\alpha\beta}), \tag{4.9}$$

$$O^{(f)}_{\alpha\beta} = a_3\eta^{v^x}_{\alpha\beta}\eta^{h''}_{\alpha\beta}(1-a_1\eta^{h^x}_{\alpha\beta}\eta^{v''}_{\alpha\beta}), \tag{4.10}$$

$$O^{(g)}_{\alpha\beta} = a_4\eta^{v''}_{\alpha\beta}\eta^{h''}_{\alpha\beta}(1-a_2\eta^{v^x}_{\alpha\beta}\eta^{h^x}_{\alpha\beta}), \tag{4.11}$$

$$\begin{aligned}O^{(h)}_{\alpha\beta} &= 1-\sum_{(j)=(a)}^{(g)} O^{(j)}_{\alpha\beta}\\ &= (1-b_h\eta^{h''}_{\alpha\beta}\eta^{h^x}_{\alpha\beta})(1-b_v\eta^{v''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta})\\ &\quad\times(1-a_1\eta^{h^x}_{\alpha\beta}\eta^{v''}_{\alpha\beta})(1-a_2\eta^{v^x}_{\alpha\beta}\eta^{h^x}_{\alpha\beta})\\ &\quad\times(1-a_3\eta^{v^x}_{\alpha\beta}\eta^{h''}_{\alpha\beta})(1-a_4\eta^{v''}_{\alpha\beta}\eta^{h''}_{\alpha\beta}).\end{aligned} \tag{4.12}$$

In Eq. (4.12) the sum lets j be a through g. The superscripts (a), (b),...,(h) refer to the configurations in Fig. I.4, i.e. $O^{(a)}_{\alpha\beta}$ should be associated with Fig. I.4a. The probability that configuration (a) occurs at (α,β) is

$$\langle O^{(a)}_{\alpha\beta}\rangle = \text{Prob. that conf. (a) occurs at } (\alpha,\beta). \tag{4.13}$$

The reason for this is simple: when $O^{(a)}_{\alpha\beta}$ operates all the anticommuting variables are used up; no walls can enter the (α,β) site so that nothing can happen (which is exactly what is depicted in Fig. I.4a). The factor of $(b_h b_v - a_1a_3 - a_2a_4)$ assures that this site has the appropriate weight of configuration (a). A similar conclusion is reached for the other $O_{\alpha\beta}$'s. The probability in Eq. (4.13) is easily calculated

$$\begin{aligned}\langle O^{(a)}_{\alpha\beta}\rangle = (b_h b_v - a_1a_3 - a_2a_4)\big[&\langle\eta^{h''}_{\alpha\beta}\eta^{h^x}_{\alpha\beta}\rangle\langle\eta^{v''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta}\rangle\\ &-\langle\eta^{h''}_{\alpha\beta}\eta^{v''}_{\alpha\beta}\rangle\langle\eta^{h^x}_{\alpha\beta}\eta^{v^x}_{\alpha\beta}\rangle+\langle\eta^{h''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta}\rangle\langle\eta^{h^x}_{\alpha\beta}\eta^{v''}_{\alpha\beta}\rangle\big].\end{aligned} \tag{4.14}$$

In general

$$\langle O^{(c_1)}_{\alpha_1\beta_1}O^{(c_2)}_{\alpha_2\beta_2}\cdots O^{(c_m)}_{\alpha_m\beta_m}\rangle = \text{the probability that sites,} \tag{4.15}$$

(α_1,β_1) through (α_m,β_m) have configurations (c_1) through (c_m).

Equation (4.15) is calculated using Eqs. (4.5)–(4.12), Eq. (3.3), and Eqs. (2.11)–(2.20). The result would be a sum of Pfaffians. A similar sum was encountered in Sec. III in computing Ising model spin correlations [see Eqs. (3.9)–(3.11)]. There, it was possible to rewrite this sum as a single Pfaffian. The same trick works here. For example,

$$\begin{aligned}\langle O^{(b)}_{\alpha\beta}\rangle &= \langle b_h\eta^{h''}_{\alpha\beta}\eta^{h^x}_{\alpha\beta}(1-b_v\eta^{v''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta})\rangle\\ &= (-b_h b_v)(\mathrm{Pf}M),\end{aligned} \tag{4.16}$$

where

$$\begin{aligned}&M_{34} = -M_{43} = \langle\eta_3\eta_4\rangle - 1/b_v,\\ &\text{all other } M_{ij} = \langle\eta_i\eta_j\rangle,\end{aligned} \tag{4.17}$$

and the abbreviations

$$\eta_1 = \eta^{h''}_{\alpha\beta}\quad \eta_2 = \eta^{h^x}_{\alpha\beta}\quad \eta_3 = \eta^{v''}_{\alpha\beta}\quad \eta_4 = \eta^{v^x}_{\alpha\beta}, \tag{4.18}$$

have been used. In other words the contraction between $\eta^{v''}_{\alpha\beta}\eta^{v^x}_{\alpha\beta}$ in calculating Pf M (via Wicks theorem or Gaussian integration) gets an extra contribution of $-1/b_v$. A systematic set of rules can be developed to calculate Eq. (4.15) as a $4m\times 4m$ Pfaffian.

Rules for Calculating Equation (4.15) the Vertex Correlations:

1. Using the following abbreviations for anticommuting variables

$$\begin{aligned}&\eta_{4l-3} = \eta^{h^o}_{\alpha\beta_l},\\ &\eta_{4l-2} = \eta^{h^o}_{\alpha\beta_l},\\ &\eta_{4l-1} = \eta^{v^o}_{\alpha\beta_l},\\ &\eta_{4l} = \eta^{v^x}_{\alpha\beta_l},\end{aligned}\qquad \text{for } l = 1,2,\ldots,m. \tag{4.19}$$

2. Equation (4.15) is

$$\langle O^{(c_1)}_{\alpha_1\beta_1}\cdots O^{(c_m)}_{\alpha_m\beta_m}\rangle = \left(\prod_{i=1}^{m} f^{(c_i)}\right)\mathrm{Pf}\,M_{ij}, \tag{4.20}$$

where

$$M_{ij} = \langle\eta_i\eta_j\rangle + \Delta_{ij}. \tag{4.21}$$

It remains to define the f's and Δ_{ij}'s:

3. The f's are

$$\begin{aligned}&f^{(a)} = f^{(h)} = (b_h b_v - a_1a_3 - a_2a_4)\equiv f,\\ &f^{(b)} = f^{(c)} = -b_h b_v,\\ &f^{(d)} = f^{(f)} = +a_1a_3,\\ &f^{(e)} = f^{(g)} = +a_2a_4.\end{aligned} \tag{4.22}$$

4. The Δ_{ij}'s are somewhat more awkward to define. Δ_{ij} is antisymmetric in i and j, that is $\Delta_{ij} = -\Delta_{ji}$, so that M in Eq. (4.21) is an antisymmetric matrix. Each of the m operators, $O^{(c_l)}_{\alpha_l\beta_l}$, involve the four anticommuting variables at (α_l,β_l). It is useful to group these into "clusters". The cluster associated with $O^{(c_1)}_{\alpha_1\beta_1}$ is η_1, η_2, η_3, and η_4, with $O^{(c_2)}_{\alpha_2\beta_2}$ it is η_5, η_6, η_7, η_8, etc. If η_i and η_j are from different clusters then $\Delta_{ij} = 0$ (in fact most Δ_{ij} are zero). It is thus sufficient to define Δ_{ij} for i and j within the l th cluster. This depends on (c_l), the configuration associated with the l th cluster. The results are tabulated as follows:

Configuration	Nonzero Δ_{ij}	i,j Values
(a)	all $\Delta_{ij}=0$	
(b)	$\Delta_{ij}=-1/b_v$	for $i=4l-1, j=4l$
(c)	$\Delta_{ij}=-1/b_h$	for $i=4l-3, j=4l-2$,
(d)	$\Delta_{ij}=1/a_3$	for $i=4l-3, j=4l$,
(e)	$\Delta_{ij}=1/a_4$	for $i=4l-3, j=4l-1$,
(f)	$\Delta_{ij}=-1/a_1$	for $i=4l-2, j=4l-1$,
(g)	$\Delta_{ij}=1/a_2$	for $i=4l-2, j=4l$,
(h)	$\Delta_{ij}=-b_h/f$	for $i=4l-1, j=4l$,
	$\Delta_{ij}=-b_v/f$	for $i=4l-3, j=4l-2$,
	$\Delta_{ij}=-a_1/f$	for $i=4l-3, j=4l$,
	$\Delta_{ij}=-a_2/f$	for $i=4l-3, j=4l-1$,
	$\Delta_{ij}=a_3/f$	for $i=4l-2, j=4l-1$,
	$\Delta_{ij}=-a_4/f$	for $i=4l-2, j=4l$,

(4.23)

where f in defined in Eq. (4.22). Equation (4.23) defines Δ_{ij} for $i<j$. For $i>j$, $\Delta_{ij}=-\Delta_{ji}$. All other Δ_{ij} are zero.

Equations (4.16)–(4.18) form a simple example of these rules. As a more complicated example let us calculate the probability, $P^{(a)\,(c)\,(h)}_{\alpha_1\beta_1,\alpha_2\beta_2,\alpha_3\beta_3}$, of having simultaneously configurations (a), (c), and (h) at sites (α_1,β_1), (α_2,β_2), and (α_3,β_3). Set

$$\begin{aligned}
M_{5,6} &= \langle \eta_5\eta_6\rangle - 1/b_h,\\
M_{9,10} &= \langle \eta_9\eta_{10}\rangle - b_v/f,\\
M_{9,11} &= \langle \eta_9\eta_{11}\rangle - a_2/f,\\
M_{9,12} &= \langle \eta_9\eta_{12}\rangle - a_1/f,\\
M_{10,11} &= \langle \eta_{10}\eta_{11}\rangle + a_3/f,\\
M_{10,12} &= \langle \eta_{10}\eta_{12}\rangle - a_4/f,\\
M_{11,12} &= \langle \eta_{11}\eta_{12}\rangle - b_h/f,
\end{aligned} \tag{4.24}$$

all other

$$M_{ij}=\langle \eta_i\eta_j\rangle,\ (i,j=1 \text{ to } 12).$$

The η_i's are defined via Eq. (4.19) for $l=1,2$, and 3. The answer is

$$P^{(a)\,(c)\,(h)}_{\alpha_1\beta_1,\alpha_2\beta_2,\alpha_3\beta_3}=(-b_hb_v)(f)^2\mathrm{Pf}M_{ij}, \tag{4.25}$$

where f is given in Eq. (4.22). It is easy to calculate free-fermion vertex correlations using the above rules. If m configurations are specified the answer is a Pfaffian of a $4m\times 4m$ matrix.

I conclude this section with a few remarks:

Remark (a): It is trivial to adapt the formalism to handle walls and vertex configurations simultaneously. Everything is calculable in terms of a Pfaffian. The probability of having a wall as (α,β) and a (b) vertex configuration at (α',β') is easily calculated and would be a Pfaffian of a 6×6 matrix.

Remark (b): When vacuum expectation values are taken, other operators work equally well. For example

$$\langle z_h b_v \eta^{h'}_{\alpha-1\beta}\eta^{h''}_{\alpha\beta}\eta^{v''}_{\alpha\beta}\eta^{v'}_{\alpha\beta}\rangle=\langle O^{(c)}_{\alpha\beta}\rangle. \tag{4.26}$$

The reason for this is simple. $B_{\alpha-(1/2)\beta}$ which is $z_h\eta^{h'}_{\alpha-1\beta}\eta^{h''}_{\alpha\beta}$ produces a unit of wall which enters the (α,β) site from the left. Because $\eta^{v''}_{\alpha\beta}\eta^{v'}_{\alpha\beta}$ uses up the vertical variables at (α,β) this wall must continue straight through thus yielding configuration (c); it is impossible to use any of the corners at (α,β).

In a sense the $O_{\alpha\beta}$'s are not unique; many operators will work. Those defined in Eqs. (4.5)–(4.12), however, have the advantage of using only those anticommuting variables at one site.

Remark (c): The matrix elements of M in Eq. (4.21) involve the anticommuting variable correlations. These, in turn, are given in integral form in Eqs. (2.11)–(2.20). In principle, the integrals in Eqs. (2.11)–(2.20) can be done in terms of elliptic functions.

V. SUMMARY

Here is a summary of these first two papers. The focus of attention was solvable two-dimensional statistical mechanics models, in particular, the Ising model, the free-fermion model, and the close-packed dimer problems. The partition functions were expressed as integrals over anticommuting variables. In this form they resembled fermionic field theories. The solvable models had quadratic actions, which were computed by using free field theory techniques.

What else has been accomplished?

(a) The methods of derivation were new. This was the first time Grassmann integrals have been used to obtain physical concrete results. These are powerful new techniques.

(b) In a novel and concise manner the Ising model partition function was computed. Using the formulas in Secs. II and III, any spin correlation function can be computed in a page of algebra. This includes the vacuum expectation value on any arbitrary product of spin variables. This work presented the simplest and shortest derivation of these results.

(c) For the first time all correlation functions were computed in the free-fermion eight vertex model.

(d) New graphical methods were developed which allowed one to compute partition functions and anticommuting variable correlation functions by solving miniature dimer problems. This provided a quick and simple graphical calculational approach. Many models can be solved by drawing a few diagrams.

These two papers show that the best approach to solving these two-dimensional models is through anticommuting variables and functional integrals.

[1] S. Samuel, J. Math. Phys. **21**,2806(1980). References to equations and figures in this paper will be prefixed by a I, e.g. Eq. (I.1.1) and Fig. I.1 refer to Eq. (1.1) and Figs. 1 of Ref. 1.

[2] S. Samuel, "The Correlation Functions in the 32 Vertex Model," IAS preprint (March, 1980).

[3] S. Samuel, "The Use of Anticommuting Integrals in Statistical Mechanics II," LBL preprint 8300 (Oct. 1978), which can be found in S. Samuel, Ph.D. thesis, Berkeley (1979).

[4] E. W. Montroll, R. B. Potts, and J. C. Ward, J. Math. Phys. **4**, 308 (1963); B. M. McCoy and C. A. Tracy; Phys. Rev. Lett. **38**, 793 (1977).

[5] See, for example, E. W. Montroll, *Brandeis University Summer Institute in Theoretical Physics. 1966*, edited by M. Chrétien, E. P. Gross, and S. Deser (Gordon and Breach, New York, 1968).

The use of anticommuting variable integrals in statistical mechanics. III. Unsolved models [a)]

Stuart Samuel [b)]
Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720
Institute For Advanced Study, Princeton, New Jersey 08540

(Received 16 July 1980; accepted for publication 30 July 1980)

The Ising model in three dimensions is fermionized by using integrals over anticommuting variables. The result is generalized to the Ising model in arbitrary dimensions and in a magnetic field. Approximation methods are developed to attack unsolved statistical mechanics models. Perturbation theory and the Hartree approximation are applied to the unsolved monomer-dimer problems. The result is a numerical solution to this unsolved class of problems. Anticommuting variables appear to be a powerful approach to unsolved problems.

I. INTRODUCTION

Two previous papers[1-3] have applied anticommuting variable integrals to statistical mechanics problems. They provided a direct and simple way of writing statistical systems as fermionic field theories. They considered two-dimensional solvable models: the Ising model, the free-fermion ferroelectric eight vertex models, and the planar close-packed dimer models. Anticommuting variables are the best way of completely solving these models. All quantities were trivially calculable including partition and correlation functions. These models had quadratic actions and were like free field theories.

This paper continues to represent via Grassmann integrals partition functions as fermionic field theories. Now, however, only models resulting in interacting theories are considered. These models are not exactly solvable, although they are amenable to approximation schemes. Because they are in field theory form all the techniques of many body theory are applicable. This is one big advantage of anticommuting variables. There are several (Pfaffian and fermionic operator) methods which exactly solve certain models in two dimensions. Most physical systems, though, are "interacting" models. These other methods neither go beyond two dimensions nor are able to treat unsolvable systems. They have limited applicability. Anticommuting variables can handle both solvable and unsolvable problems. In the former case they efficiently solved the model and in the latter case they generate viable approximation schemes. They can perturb about a solvable model to obtain results for an unsolvable system. This paper will show how this is done.

This is just the beginning. This paper uses only two of the many possible approximation techniques available. This will certainly be an active area of future research: to establish new techniques as well as adapting many-body theory techniques. The number of models to which anticommuting variables can be applied seems limitless. This paper considers dimer and Ising models in two, three, and more dimensions.

Nontrivial models can be represented in fermionic functional integral form. To demonstrate this the Ising model in three dimensions is considered. A four fermion field theory is obtained. Next the two-dimensional Ising model in magnetic field is treated, and, it, too, is a four fermion field theory. In addition, it is represented as a Z_2 gauge theory coupled to a fermion. Although, these anticommuting variable representations are somewhat complicated they demonstrate that non-trivial models can be handled. The approximation methods developed in this and future papers[4] can be applied to these models, however to illustrate the methods the dimer-monomer mixing problem is considered. It is the simplest nontrivial model representable in Grassmann integral form. As such it is quite amenable to approximation schemes. It is also of particular interest: many problems can be mapped into a dimer-monomer mixing model.

This paper contains several new results. The following list of results, which might be part of the conclusion, serves to indicate the contents of this paper.

In Sec. II the three-dimensional Ising model is expressed as an integral over anticommuting variables.[5] In this form it is equivalent to an interacting "fermionic" field theory. This is an important result because this paper's approximation schemes become applicable to the Ising model. This section will form the foundation of future work. The higher dimensional Ising models are also written as anticommuting variable integrals. This is the first fermionic representation of the three-dimensional Ising model.

In Sec. III the integral representation in Sec. II is adapted to the two-dimensional Ising model in magnetic field. This is also an interacting "fermionic" field theory. Presented next is a representation as a Z_2 lattice gauge theory coupled to a "fermion". The representations are again extendable to high dimensions. Again, this is the first time Ising models in magnetic field have been fermionized.

Section IV deals with dimer models in the abstract, that is, the most general dimer model is considered. They are expressed in anticommuting variable form and many-body field theory methods are applied. Feynman graph rules are presented. Perturbation theory turns out to be equivalent to the low temperature expansion. The self-consistent Hartree approximation is calculated. The Feynman rules are adapted so that corrections to the Hartree approximation can be

[a)]This work was supported by the High Energy Physics Division of the U. S. Department of Energy under contract No. W-7405-ENG-48.
[b)]Present address: Institute for Advanced Study, Princeton, New Jersey 08540.

 0022-2488/80/122820-14$1.00

calculated. These are computed to sixth order. No specific model is considered. The results of Sec. IV are valid for the most general dimer model. This is a new expansion.

In Sec. V the methods of Sec. IV are applied to specific dimer models. The lattices include the d-dimensional hypercubic lattices (such as the simple quadratic and simple cubic lattices), the planar triangular, the tetrahedral, the bcc and the fcc lattices. A special set of Feynman rules are derived for translationally invariant lattices. Embedding graphs (and their weights) were found to 5th order for close-packed lattices and to 6th order for loose-packed lattices. This allowed a rapid computation of the Hartree-improved expansion to 5 or 6 orders. The method is carried out to sixth order first for the two-dimensional dimer problem and then for the d-dimensional hypercubic lattices. This is for the nonisotropic case in which Boltzmann factors in different directions need not be equal. In the isotropic case, it is found that the d-dimensional hypercubic dimer problem is exactly solvable as d becomes large as long as the temperature is high enough. A $1/d$ expansion is presented. A similar analysis is applied to lattices with large coordination number, q. All dimer models become exactly solvable as $q \to \infty$ and a $1/q$ expansion is presented. For lattices with q varying from 4–12, molecular freedoms are computed in the pure dimer limit. Even at such small q values results are good to several per cent. For the isotropic case, previously established low temperature expansions are combined with Hartree methods to obtain the Hartree expansion from 8–16 orders on six lattices. These new series expansions accurately represent the six models in the entire physical region. In the region where the approximation method is expected to be the worst, that is, at close-packing, molecular freedoms are computed to an accuracy of a fraction of a percent. Next the density and entropy are calculated. At a density of about 90% maximum density the density and entropy are calculated to an accuracy of from four to seven decimal places. At 50% density the accuracy is from six to nine decimal places and at 10% density the accuracy ranges from 11 to 19 decimal places. These new series expansions are as good as any in the literature.

The Hartree series represents a numerical solution to an interesting class of unsolvable models. The extreme accuracy achieved is beyond the requirements of physical or theoretical demands. In effect an unsolvable model has been solved.

Anticommtuing variables have been used to obtain many, many new results. Space restrictions prevent us from presenting all of them but here is a list of what else has been accomplished.[6] These results will be published elsewhere.[4]

1. Other complicated free-fermion vertex models have been solved.[7]

2. Correlation functions have been computed in the free-fermion 32 vertex model.[8]

3. The $1/N$ expansion and random phase approximation have been applied to dimer models. A dimer model with a local U(N) symmetry has been solved in a $1/N$ limit.

4. Partition and correlation functions have been computed for the general one-dimensional polymer system. These results have been used to compute transfer matrix elements for two-dimensional dimer and polymer systems. The two-dimensional extremely anisotropic dimer and polymer models have been "perturbatively solved" using these results.

5. Bosonization of fermionic systems has been discussed.

6. The free-fermion model has been used to obtain results for the general unsolved eight vertex model via perturbation theory. Feynman rules have been presented and caluclations to second order have been done.

7. Hartree-Fock equations for the unsolved eight vertex model have been derived.

8. Integral transformations on various models have been performed. These include partial integration, real space renormalization, rescaling and canonical transformations. The Schwinger–Dyson equations have been used to establish relations among correlation functions.

9. Rigorous upper bounds on the free energies have been obtained for several systems.

The anticommuting variable methods developed in these three papers are the foundation for all these new results. They will also be the foundation for future work.

II. THE THREE-DIMENSIONAL ISING MODEL AS AN INTERACTING FERMIONIC FIELD THEORY

This section expresses the partition function for the d-dimensional Ising model as an anticommuting variable integral over an action, that is, a lattice fermionic field theory. Unlike the two-dimensional Ising model where the action was quadratic,[1] the action of the three-dimensional model involves quartic as well as bilinear terms. For the d-dimensional model there is a product of $2(d-1)$ anticommuting variables. Therefore the d-dimensional model is not of the solvable free-fermion form but represents an interacting field theory. The construction for the three-dimensional case will be explained but the formula will be written for the d-dimensional case.

The partition function has a well-known geometrical low temperature expansion similar to the two-dimensional model except that one must sum over closed polyhedrons instead of closed polygons. What kinds of configurations are allowed? First, any number of nonoverlapping polyhedrons can occur. They may intersect in the manner of Fig. 1a but they may not overlap as in Fig. 1b. The configuration of Fig. 1b would be drawn as in Fig. 1c. The fact that overlap is not permitted makes the use of anticommuting variables ideal. Polyhedrons, constructed out of anticommuting variables, cannot overlap because the square of a variable is zero.

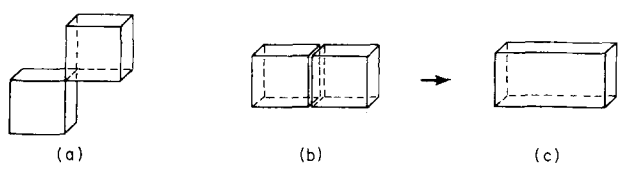


FIG. 1. (a) Intersecting polyhedrons: Such intersections are allowed. (b) Overlapping Polyhedrons. Such overlaps are forbidden. This configuration would be drawn as in (c).

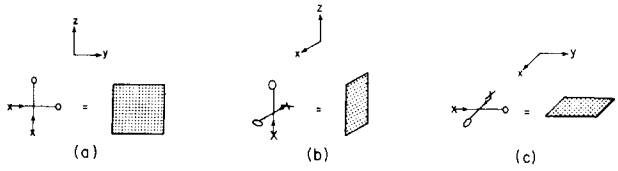


FIG. 2. A_{face}.

$$Z_{\text{3-d Ising}}(J_1, J_2, J_3) \propto Z_{\text{closed polyhedrons}}(z_1, z_2, z_3), \quad (2.1)$$

where $Z_{\text{closed polyhedrons}}(z_1, z_2, z_3)$ is the partition function for nonoverlapping but possibly intersecting polyhedrons in which the three types of faces are weighted by z_1, z_2, and z_3 and

$$z_i \equiv \exp(-2\beta J_i). \quad (2.2)$$

In the two-dimensional model, the anticommuting variable action that generated $Z_{\text{closed polygons}}$ consisted of three pieces: A_{wall}, A_{corner}, and $A_{\text{monomer}} \cdot A_{\text{wall}}$ drew the sides of polygons, A_{corner} formed corners, and A_{monomer} filled unfilled sites. Similarly, in three dimensions the action consists of three pieces, A_{face}, A_{corner}, and $A_{\text{monomer}} \cdot A_{\text{face}}$ draws the faces of the polyhedrons and A_{corner} joins the faces together.

The expression for $Z_{\text{closed polyhedron}}$ in terms of anticommuting variables is first presented and subsequently explained.

$$Z_{\text{closed polyhedrons}}(z_1, z_2, z_3) = \int d\eta\, d\eta^{\dagger} \exp A, \quad (2.3)$$

where

$$A = A_{\text{face}} + A_{\text{corner}} + A_{\text{monomer}}. \quad (2.4)$$

Only two out of the three types of anticommuting variables occur at a particular edge midpoint. If it is an x edge, they are the other two types, namely, η^2, $\eta^{2\dagger}$, η^3, $\eta^{3\dagger}$. Likewise for y and z edges.

A_{face} has three terms. Each draws one of the faces of Fig. 2. A_{face} involves a product of four anticommuting variables. Together they span a square unit of surface area as Fig. 2 illustrates.

These quartic terms have two arrows. These arrows determine the ordering of each of the two bilinears making up the quartic. There is never any confusion determining the ordering of anticommuting variables from figures such as Fig. 2 because bilinears commute.

The faces in the x direction (for example) can link to form larger x directed surface areas (Fig. 3a) but faces in two different directions cannot (Fig. 3b). A_{corner} makes this possible by using bilinear "hooks". What is needed to link the two faces in Fig. 3b is the object in Fig. 3c. It is of the form, $\eta^1\eta^3$, and acts like a hinge. Such objects are needed at the midpoints of each of the three types edges. Thus, A_{corner} has three terms. Figure 4a shows an x edge, the possible anticommuting variables which could enter it, and the corners. The corners are exactly the same as for the two-dimensional Ising model.[1] Figures 4b and 4c show the analogous objects for y and z edges. Perhaps a better name for A_{corner} would be A_{hinge} because of the manner in which the edges are joined.

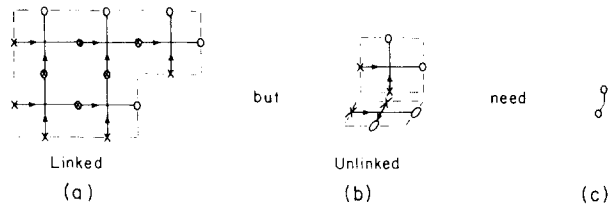


FIG. 3. Linking: (a) The faces of Fig. (a) can form larger area elements. Here five faces link. (b) But a face in the x direction is unable to link with a face in the z direction. The object in Fig. (c) is needed.

Finally, monomers are needed to fill empty sites. These monomers are similar to the two-dimensional case.

In short, the corner and face actions correspond to the simple pictures in Fig. 2 and 4. One should always think in terms of these pictures.

A moment's thought reveals that the action [Eq. (2.4)] generates closed polygons of the type needed in Eq. (2.1). If faces are weighted by the appropriate Boltzmann factors [Eq. (2.2)], then, up to a minus sign, the correct weights are obtained. A minus sign might be generated because of anticommuting variable reordering. The anticommuting variables must be put in $\eta\eta^{\dagger}$ form. This involves anticommuting operations, each of which yields a minus factor. Fortunately, all terms are indeed positive: the quartic terms can be broken up into the product of the two bilinears. The bilinears are only able to combine with corners in a two-dimensional plane. They generate planar closed polygons like the ones in the two-dimensional Ising model.[1] By choosing the same bi-

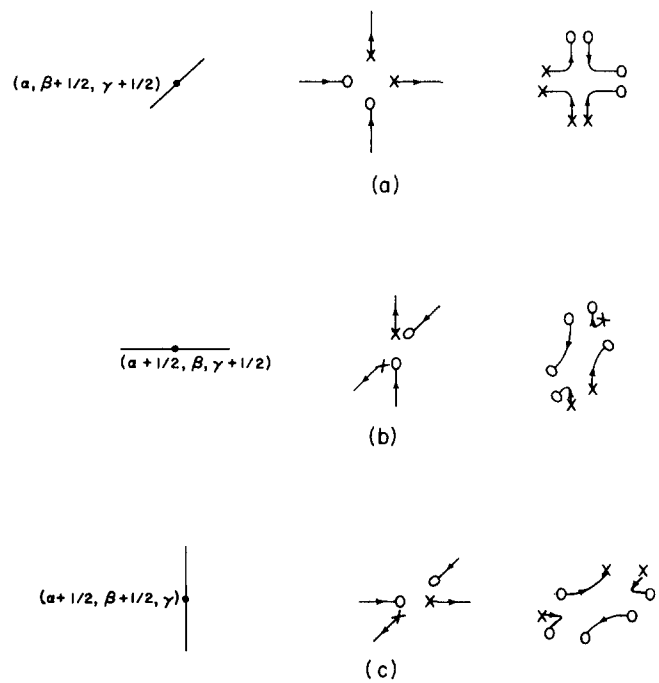


FIG. 4. A_{corner}. To the left is the edge and its coordinates. In the middle are the possible anticommuting variables which could enter. These variables come from A_{face}. To the right are the four types of corners needed to link faces.

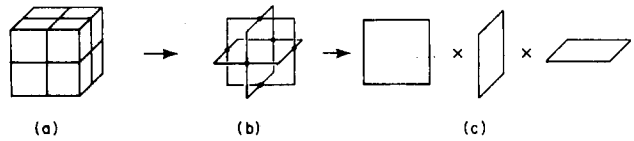


FIG. 5. The Minus Sign Problem: (a) A cube of polyhedron. (b) The anticommuting variables used to construct the cube trace out this object. (c) By breaking quartics into products of bilinears, the object factorizes into a product of three planar polygons. Reordering minus factors reduce to the planar case.

linear ordering as in a two-dimensional model, all terms are guaranteed to be positive. Effectively, the minus sign problem reduces to the two-dimensional case. Figure 5 illustrates this.

For the d-dimensional Ising model, use objects of dimension $d-1$ (the low temperature expansion). The action consists of bilinear terms plus interacting $2(d-1)$ products of anticommuting variables:

$$A_{\text{polycomplex}} = A_{(d-1)\text{face}} + A_{\text{corner}} + A_{\text{monomer}}, \tag{2.5}$$

$$A_{(d-1)\text{face}} = \sum_{\mathbf{x}} \sum_{\substack{i_1,i_2,\ldots,i_d \\ \text{cyclically}}} z_{i_1} (\eta^{i_2\dagger}_{\mathbf{x}+\mathbf{v}_{i_1,i_2}} \eta^{i_2}_{\mathbf{x}+\mathbf{u}_{i_1,i_2}} \times \eta^{i_3\dagger}_{\mathbf{x}+\mathbf{v}_{i_1,i_3}} \eta^{i_3}_{\mathbf{x}+\mathbf{u}_{i_1,i_3}} \cdots \times \eta^{i_d\dagger}_{\mathbf{x}+\mathbf{v}_{i_1,i_d}} \eta^{i_d}_{\mathbf{x}+\mathbf{u}_{i_1,i_d}}), \tag{2.6}$$

$$A_{\text{corner}} = \sum_{\mathbf{x}} \sum_{i<j} \{\eta^{i\dagger}_{\mathbf{x}+\mathbf{u}_{ij}} \eta^{j}_{\mathbf{x}+\mathbf{u}_{ij}} + \eta^{i\dagger}_{\mathbf{x}+\mathbf{u}_{ij}} \eta^{i\dagger}_{\mathbf{x}+\mathbf{u}_{ij}} + \eta^{j}_{\mathbf{x}+\mathbf{u}_{ij}} \eta^{i\dagger}_{\mathbf{x}+\mathbf{u}_{ij}} + \eta^{j}_{\mathbf{x}+\mathbf{u}_{ij}} \eta^{i}_{\mathbf{x}+\mathbf{u}_{ij}}\}, \tag{2.7}$$

$$A_{\text{monomer}} = \sum_{\mathbf{x}} \sum_{i<j} \{\eta^{i}_{\mathbf{x}+\mathbf{u}_{ij}} \eta^{i\dagger}_{\mathbf{x}+\mathbf{u}_{ij}} + \eta^{j}_{\mathbf{x}+\mathbf{u}_{ij}} \eta^{j\dagger}_{\mathbf{x}+\mathbf{u}_{ij}}\}, \tag{2.8}$$

$$\mathbf{u}_{ij} = \tfrac{1}{2}(\mathbf{e}_i + \mathbf{e}_j), \tag{2.9}$$

$$\mathbf{v}_{ij} = \tfrac{1}{2}(\mathbf{e}_i - \mathbf{e}_j).$$

The $\mathbf{e}_i$ are unit vectors in the ith direction.

The notation needs explaining. Begin with the spatial labels. When spins have integer cartesian coordinates, the polyhedrons, being drawn on the dual lattice, involve half-integer coordinates.

The anticommuting variables sit at edge midpoints. There are d types: η^i, $i = 1 \ldots d$ (along with their daggered partners), which refer to anticommuting variables associated with ith directions. Conventions used here are: o and x indicate undaggered and daggered variables; a line in the i^{th} direction attached to an anticommuting variable indicates that it is of the i^{th} type; the subscripts indicate an anticommuting variable's cartesian coordinates; and arrows denote the ordering of bilinears.

III. THE TWO-DIMENSIONAL ISING MODEL IN A MAGNETIC FIELD

This section expresses the partition function for the two-dimensional Ising model in background magnetic field in two ways. The first way uses anticommuting variables only. It has quartic terms in the action and hence is an interacting fermionic field theory. The second representation is of "mixed" form: using both anticommuting variables and bosonic variables. It is, in particle physics language, a Z_2 lattice gauge theory[9,10] coupled to a fermion. From a particle physicist's point of view this is an interesting representation: the four-dimensional counterpart is a model for quark confinement.

Let H be the magnetic field. In what follows it is necessary for H to be positive (or zero).

The two-dimensional Ising model in magnetic field is again equivalent to a closed polygon partition function. In addition to polygonal sides being weighted areas must also be weighted. There is a factor of

$$z_A = \exp(-2\beta H) \tag{3.1}$$

for each square unit of polygonal area. Treat the two-dimensional system as a three-dimensional system which is one unit thick in the z direction. Draw polyhedrons around regions of down spin in lieu of polygons. This transforms the problem into the $Z_{\text{closed polyhedron}}$ type of Sec. II. Take the action in Eq. (2.5) for $d = 3$ but restrict position sums to be in the $z = 0$ to $z = 1$ layer. Use z_1 and z_2 of Eq. (2.2) to weight faces in the x and y direction but use $z_A^{1/2}$ for z_3 (the square root of z_A appears because z_3 enters twice once in the $z = 1$ plane and once in the $z = 0$ plane, for each square unit of polygonal area). Thus the two-dimensional Ising model in magnetic field has been represented as a four-fermion interacting field theory. Of course, the construction works in d-dimensions by using Eq. (2.5) for the $(d+1)$-dimensional Ising model and restricting the $(d+1)$th direction to be one unit thick. The action involves bilinears and products of $2d$ anticommuting variables.

The task of weighting areas can also be done using a gauge field. Pretend, for the moment, that the polygons (or more precisely, the polygonal curves) are oriented. Think of such curves as charged particle trajectories, the orientation being associated with the direction of flow of charge. Coupling them to an Abelian gauge theory (as in quantum electrodynamics) would weight the polygon's area because $(1+1)$-dimensional QED has a linear potential. Unfortunately, the curves in $Z_{\text{closed polygon}}$ [Eq. (I3.1)] are not oriented and this trick fails. Fortunately, the difficulty can be overcome by using a Z_2 gauge field instead of a U(1) one. Being blind to the difference between positive and negative charges, a Z_2 gauge field works. The result is

$$Z_{\text{Ising}}(J_h, J_v, H) = f' \sum_{\substack{U_{\alpha+1/2\beta} = \pm 1 \\ U_{\alpha\beta+1/2} = \pm 1}} \int d\eta\, d\eta^\dagger \exp A, \tag{3.2}$$

where the action, A is

$$A = A_{\text{wall}} + A_{\text{corner}} + A_{\text{monomer}} + A_{Z_2}, \tag{3.3}$$

A_{corner}, and A_{monomer} are the same actions as in I Eq. (I.3.4) (with $a_1 = a_2 = a_3 = a_4 = b_v = b_h = -1$). A_{wall} is modified to

$$A_{\text{wall}} = \sum_{\alpha\beta} (\eta^{h\dagger}_{\alpha\beta} \eta^{h}_{\alpha+1\beta} z_h U_{\alpha+1/2\beta} \, \eta^{v\dagger}_{\alpha\beta} \eta^{v}_{\alpha\beta+1} z_v U_{\alpha\beta+1/2}), \tag{3.4}$$

and

$$A_{Z_2} = K \sum_{\alpha\beta} (U_{\alpha+1/2\beta} \times U_{\alpha+1\beta+1/2} U_{\alpha+1/2\beta+1} U_{\alpha\beta+1/2}). \quad (3.5)$$

In these formulas

$$z_h = \exp(-2\beta J_v),$$
$$z_v = \exp(-2\beta J_h), \quad (3.6)$$
$$\tanh K = \exp(-2\beta H),$$
$$f' = \frac{\exp(N\beta H)}{(4\cosh K)^N} \exp \beta N (J_v + J_h),$$
$$= \frac{1}{(8 \sinh 2K)^{N/2}} \exp \beta N (J_v + J_h).$$

Again, the method generalizes to higher dimensions.

IV. COORDINATE SPACE PERTURBATION THEORY FOR THE GENERAL DIMER PROBLEM

This section and the following section will deal with the dimer problem. This constitutes a whole class of problems since there are many lattices at one's disposal. The dimer problem is not only important because of its direct application to physical systems,[11] but also because of the large number of problems which can be mapped into dimer form. This enhances their importance. The only models which have been solved are the one-dimensional dimer model and two-dimensional close-packed models. Approximation methods are therefore of interest. My purpose will be twofold: First, the anticommuting variable technique will be used to obtain new dimer series expansions. These represent new approaches to the dimer system. Secondly, in the process of obtaining the expansions, various anticommuting variable approximation techniques will be illustrated. Dimer models are a good laboratory for testing these because of their simplicity and because of other existing approximation schemes to which they can be compared. The importance of these sections is that the approximation techniques are applicable to any model representable in fermionic-like field theory form (such as the models discussed in Secs. II and III). One merely mimics the methods illustrated here.

An extensive set of dimer references can be found in Ref. 12, to which the reader is referred. I would like, however, to mention the following: Previous approximation schemes fall into the following catagories: First, there are those[13] which solve exactly small finite lattices and then extrapolate to large lattices. This technique is known as the exact finite method: A close cousin is Monte Carlo.[14] There are also transfer matrix methods.[15] These give excellent numerical results. Next is the Bethe approximation.[16] It is of interest because of its simplicity both mathematically and physically and because of its accuracy which is reasonable. There are ways of calculating corrections to the Bethe approximation.[17,18] Rigorous mathematical dimer results also exist.[12,19] The importance of Ref. 12 should not be neglected. With reasonable assumptions Heilmann and Lieb have shown that no phase transition can occur as long as monomer Boltzmann factors are nonzero. The result is general. It is applicable to almost all dimer models. Phase transitions can only occur for pure dimer systems. Finally, there are the series expansions. The simplest is the low temperature expansion in powers of the dimer Boltzmann factor. This can be organized into a Mayer type expansion.[20,21,22] A great improvement is Nagle's series.[18] It starts with the Bethe approximation and generates a series using graphical methods. It systematically calculates corrections to the Bethe approximation, which, because it is a good starting point guarantees an excellent series. Nagle's series is presently the best in the literature. The Hartree series developed in this section equals Nagle's in accuracy. It is a new expansion. The field theoretic Hartree method is used after expressing the dimer problem as a fermionic field theory. Since dimers cannot overlap, fermions are natural variables: roughly speaking, dimers constructed out of fermions are unable to overlap because of the Pauli principle. The perturbative techniques developed here are easily extended to other systems such as trimers or more complicated polymers. Nagle's method has also been extended to trimers[23] although more complicated polymeric systems have not been treated. A final note: Ref. 12 has an important implication for this paper's Hartree series (and also Nagle's series). It guarantees convergence in the entire physical region.

This section will treat the dimer problem from a general point of view: A specific example will be considered in the next section. Key results are the Hartree approximation [Eq. (4.8)] and the Hartree-improved Feynman rules which generate the series in Eq. (4.12) and Fig. 10.

The general dimer model is an interacting fermionic field theory with a quartic interaction term,

$$V = \frac{1}{2} \sum_{\alpha\beta} z_{\alpha\beta} \eta_\alpha \eta_\alpha^\dagger \eta_\beta \eta_\beta^\dagger. \quad (4.1)$$

One sums over all sites α and all sites β allowing $z_{\alpha\beta}$ to be zero if no bond exists between α and β. The factor of $1/2$ compensates for the double counting in Eq. (4.1) ($z_{\alpha\beta} \equiv z_{\beta\alpha}$).

The interaction, V, is pictorially depicted in Fig. 6 and is of the same form as a two-body potential in a quantized many body theory.[24] This correspondence proves useful. The bare propagator, $G^0_{\alpha\beta}$, is determined by the quadratic piece, that is, $\Sigma_\alpha \eta_\alpha \eta_\alpha^\dagger$. It is

$$G^0_{\alpha\beta} \equiv \langle \eta_\alpha \eta_\beta^\dagger \rangle_0 = \delta_{\alpha\beta}. \quad (4.2)$$

Perturbation theory is an expansion in powers of V (or $z_{\alpha\beta}$). Since $z_{\alpha\beta} = \exp(-\beta E_{\alpha\beta})$, this is the standard low temperature expansion:

Perturbation Theory = Low Temperature Expansion. (4.3)

Feynman rules are similar to the usual many body theory ones.[24] One draws all graphs using the interaction of Fig. 6. Because of the nature of the bare propagator in Eq. (4.2), fermion loops occur at a particular site. It is convenient to

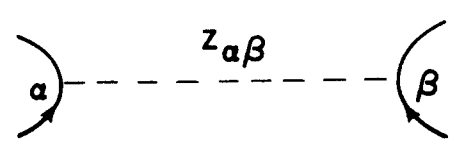
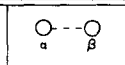
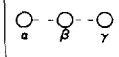
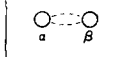
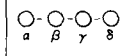
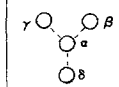
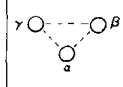
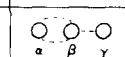
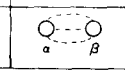


FIG. 6. The dimer potential.

	Graph	Contracted graph	Factors from rule (b)	Factor from rule (c)	Factor from rule (d)	Factor from rule (e)	Weight of graph
(a)			$\sum_{\alpha\beta} z_{\alpha\beta}$	$(-1)^2$	$\frac{1}{2}$	1	$\frac{1}{2}\sum_{\alpha\beta} z_{\alpha\beta}$
(b)			$\sum_{\alpha\beta\gamma} z_{\alpha\beta} z_{\beta\gamma}$	$(-1)^3$	$\frac{1}{2}$	1	$-\frac{1}{2}\sum_{\alpha\beta\gamma} z_{\alpha\beta} z_{\beta\gamma}$
(c)			$\sum_{\alpha\beta} z_{\alpha\beta}^2$	$(-1)^2$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}\sum_{\alpha\beta} z_{\alpha\beta}^2$
(d)			$\sum_{\alpha\beta\gamma\delta} z_{\alpha\beta} z_{\beta\gamma} z_{\gamma\delta}$	$(-1)^4$	$\frac{1}{2}$	1	$\frac{1}{2}\sum_{\alpha\beta\gamma\delta} z_{\alpha\beta} z_{\beta\gamma} z_{\gamma\delta}$
(e)			$\sum_{\alpha\beta\gamma\delta} z_{\beta\alpha} z_{\gamma\alpha} z_{\delta\alpha}$	$(-1)^3(-2)$	$\frac{1}{6}$	1	$\frac{1}{3}\sum_{\alpha\beta\gamma\delta} z_{\beta\alpha} z_{\gamma\alpha} z_{\delta\alpha}$
(f)			$\sum_{\alpha\beta\gamma} z_{\alpha\beta} z_{\beta\gamma} z_{\gamma\alpha}$	$(-1)^3$	$\frac{1}{6}$	1	$-\frac{1}{6}\sum_{\alpha\beta\gamma} z_{\alpha\beta} z_{\beta\gamma} z_{\gamma\alpha}$
(g)			$\sum_{\alpha\beta\gamma} z_{\alpha\beta}^2 z_{\beta\gamma}$	$(-1)^2(-2)$	1	$\frac{1}{2}$	$-\sum_{\alpha\beta\gamma} z_{\alpha\beta}^2 z_{\beta\gamma}$
(h)			$\sum_{\alpha\beta} z_{\alpha\beta}^3$	$(-2)^2$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}\sum_{\alpha\beta} z_{\alpha\beta}^3$

XBL 796-1723

FIG. 7. Simple perturbation theory to third order.

contract all fermion loops to a point. Figure 7 shows all the connected vacuum bubbles to third order, first in the usual way and then in the contracted form. The Feynman rules for contracted graphs are:

(a) Draw all topologically distinct graphs, consisting of any number of vertices. The vertices can have one or more lines attached to them. The vertices are assigned a site index, α. The empty graph is to be included and contributes one.

(b) For each edge associate a factor, $z_{\alpha\beta}$.

(c) For each vertex at α with l lines emanating from it (a vertex of degree l) put in a factor of $(-1)(l-1)!$.

(d) The graph may be topologically invariant under permutation of some of its vertices. Such permutations generate a symmetry group of the graph which is called the point symmetry group of the graph. Put in a factor of [order of the point symmetry group of the graph]$^{-1}$. The order of a group, G, is the number of elements in G.

(e) For each pair of vertices connected by l lines (Fig. 8) put in a factor of $1/l!$.

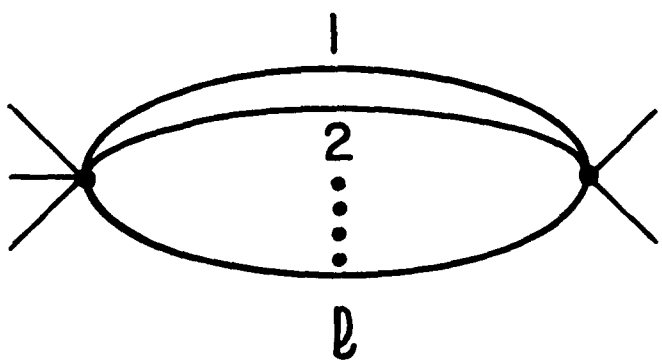


FIG. 8. Two vertices with l lines between them.

The (-1) in rule (c) arises because the vertex was originally a fermion loop for which Feynman rules assign a minus factor. The $(l-1)!$ is due to the fact that l lines entering a loop can be ordered in $(l-1)!$ ways.

If interchange of lines is considered a symmetry of a graph then rules (d) and (e) combine into one:

(de) Put in a factor of [the order of the total symmetry group of the graph]$^{-1}$.

Figure 7 shows the connected graphs through third order in $z_{\alpha\beta}$, along with the factors from rules (b), (c), (d), and (e). This illustrates how the Feynman rules work.

In rule (a) all topologically distinct graphs are to be considered including disconnected ones. It is well known in field theory that

$$Z = \sum_{\substack{\text{all graphs}\\ \text{connected or}\\ \text{disconnected}}} = \exp \sum_{\substack{\text{connected}\\ \text{graphs}}}, \tag{4.4}$$

that is, the connected graphs exponentiate. Therefore only connected graphs need be considered. Figure 7 thus gives

$$\begin{aligned} \ln Z = {} & \frac{1}{2}\sum_{\alpha\beta} z_{\alpha\beta} - \frac{1}{2}\sum_{\alpha\beta\gamma} z_{\alpha\beta} z_{\beta\gamma} \\ & + \frac{1}{4}\sum_{\alpha\beta} z_{\alpha\beta}^2 + \frac{1}{2}\sum_{\alpha\beta\gamma\delta} z_{\alpha\beta} z_{\beta\gamma} z_{\gamma\delta} \\ & + \frac{1}{3}\sum_{\alpha\beta\gamma\delta} z_{\beta\alpha} z_{\gamma\alpha} z_{\delta\alpha} - \frac{1}{6}\sum_{\alpha\beta\gamma} z_{\alpha\beta} z_{\beta\gamma} z_{\gamma\alpha} \\ & - \sum_{\alpha\beta\gamma} z_{\alpha\beta}^2 z_{\beta\gamma} + \frac{1}{3}\sum_{\alpha\beta} z_{\alpha\beta}^3 + \cdots. \end{aligned} \tag{4.5}$$

Equation (4.5) is generic in character: it is the low temperature dimer expansion to third order for *any* dimer problem.

Now that the dimer statistical system has been expressed in field theory language, standard field theory calculational methods are applicable. What has just been illustrated is simple coordinate space perturbation theory. Significant improvements can be made; for example, the self-consistent Hartree approximation.

It can be obtained by the replacement

$$\frac{1}{2}\sum_{\alpha\beta} z_{\alpha\beta}\eta_\alpha\eta_\alpha^\dagger\eta_\beta\eta_\beta^\dagger \to \frac{1}{2}\sum_{\alpha\beta} z_{\alpha\beta}\left[\eta_\alpha\eta_\alpha^\dagger\langle\eta_\beta\eta_\beta^\dagger\rangle_H + \langle\eta_\alpha\eta_\alpha^\dagger\rangle_H\eta_\beta\eta_\beta^\dagger - \langle\eta_\alpha\eta_\alpha^\dagger\rangle_H\langle\eta_\beta\eta_\beta^\dagger\rangle_H\right], \tag{4.6}$$

where $\langle\eta_\gamma\eta_\gamma^\dagger\rangle_H$, the Hartree propagator, is determined self-consistently:

$$\langle\eta_\gamma\eta_\gamma^\dagger\rangle_H = \left[1 + \sum_\beta z_{\gamma\beta}\langle\eta_\beta\eta_\beta^\dagger\rangle_H\right]^{-1}. \tag{4.7}$$

Equation (4.7) was obtained by calculating the propagator $\langle\eta_\gamma\eta_\gamma^\dagger\rangle$ with the quartic term in Eq. (4.1) replaced by Eq. (4.6). For a translationally invariant lattice Eq. (4.7) is simple to solve (this will be exemplified shortly).

The self-consistent Hartree approximation for Z is

$$\ln Z_H = \sum_\alpha \ln\left(1 + \sum_\beta z_{\alpha\beta}\langle\eta_\beta\eta_\beta^\dagger\rangle_H\right) - \frac{1}{2}\sum_{\alpha\beta} z_{\alpha\beta}\langle\eta_\alpha\eta_\alpha^\dagger\rangle_H\langle\eta_\beta\eta_\beta^\dagger\rangle_H. \tag{4.8}$$

Equation (4.8), the Hartree approximation to the partition function, is one of the results of this section.

For the $1-d$ dimer model, a numerical comparison of the Hartree approximation, Γ^H of Eq. (4.8), has been made to the exact result, Γ. Here, $\Gamma = (1/N)\ln Z$, is the grand potential per unit site. The Hartree approximation is, at most, off by 8.28% for the entire range of z. The z which yields the maximum error occurs near $z = 2.31$. It is particularly good for small z and large z. It is encouraging that such a simple technique yields a reasonably accurate approximation for all z.

For the d-dimensional dimer problem on a square lattice with weights, $z_1, z_2, \ldots, z_d$, in the first, second, ..., d th directions, the Hartree approximation is

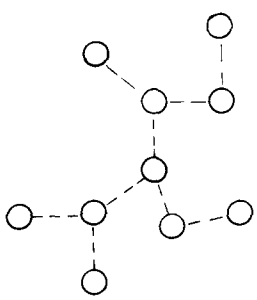


FIG. 9. A typical bubble tree graph included in the Hartree approximation.

Graph	(c) Vertex attaching factor	(d) Point symmetry factor	(e) Line symmetry factor	Graph	(c) Vertex attaching factor	(d) Point symmetry factor	(e) Line symmetry factor
1	1	$\frac{1}{2}$	$\frac{1}{2!}$	22	$-2!2!$	$\frac{1}{2}$	$\frac{1}{2!}$
2	-1	$\frac{1}{6}$	1	23	$-3!$	$\frac{1}{8}$	1
3	$2!2!$	$\frac{1}{2}$	$\frac{1}{3!}$	24	$-2!2!$	$\frac{1}{2}$	$\frac{1}{2!}\frac{1}{2!}$
4	$-2!2!$	$\frac{1}{2}$	$\frac{1}{2!}$	25	$3!3!$	$\frac{1}{2}$	$\frac{1}{3!}$
5	$-3!$	$\frac{1}{2}$	$\frac{1}{2!}\frac{1}{2!}$	26	$2!2!3!$	1	$\frac{1}{2!}$
6	1	$\frac{1}{8}$	1	27	$2!2!3!$	$\frac{1}{2}$	$\frac{1}{2!}\frac{1}{2!}$
7	$3!3!$	$\frac{1}{2}$	$\frac{1}{4!}$	28	$2!2!2!$	$\frac{1}{4}$	$\frac{1}{2!}\frac{1}{2!}$
8	-1	$\frac{1}{10}$	1	29	$3!3!$	$\frac{1}{4}$	$\frac{1}{2!}$
9	$-2!2!3!$	$\frac{1}{2}$	$\frac{1}{2!}\frac{1}{2!}$	30	$2!2!2!$	$\frac{1}{24}$	1
10	$-3!3!$	$\frac{1}{2}$	$\frac{1}{3!}$	31	$2!4!$	$\frac{1}{2}$	$\frac{1}{3!}$
11	$3!$	$\frac{1}{2}$	$\frac{1}{2!}$	32	$2!2!3!$	$\frac{1}{2}$	$\frac{1}{2!}\frac{1}{2!}$
12	$2!2!$	$\frac{1}{4}$	1	33	$2!4!$	1	$\frac{1}{2!}\frac{1}{2!}$
13	$2!2!$	$\frac{1}{2}$	$\frac{1}{2!}$	34	$3!3!$	$\frac{1}{2}$	$\frac{1}{2!}\frac{1}{2!}\frac{1}{2!}$
14	$2!2!$	$\frac{1}{2}$	$\frac{1}{2!}\frac{1}{2!}$	35	$2!2!3!$	1	$\frac{1}{2!}\frac{1}{3!}$
15	$4!4!$	$\frac{1}{2}$	$\frac{1}{5!}$	36	$5!$	$\frac{1}{6}$	$\frac{1}{2!}\frac{1}{2!}\frac{1}{2!}$
16	$-2!4!$	1	$\frac{1}{2!}\frac{1}{3!}$	37	$-4!4!$	$\frac{1}{2}$	$\frac{1}{4!}$
17	1	$\frac{1}{12}$	1	38	$-2!3!4!$	1	$\frac{1}{2!}\frac{1}{3!}$
18	$-2!2!$	$\frac{1}{2}$	$\frac{1}{2!}$	39	$-3!3!3!$	$\frac{1}{6}$	$\frac{1}{2!}\frac{1}{2!}\frac{1}{2!}$
19	$-2!2!$	$\frac{1}{2}$	1	40	$-3!5!$	1	$\frac{1}{2!}\frac{1}{4!}$
20	$-3!$	$\frac{1}{2}$	$\frac{1}{2!}$	41	$-2!2!5!$	$\frac{1}{2}$	$\frac{1}{3!}\frac{1}{3!}$
21	$-2!2!$	$\frac{1}{12}$	1	42	$5!5!$	$\frac{1}{2}$	$\frac{1}{6!}$

FIG. 10. The Hartree-improved perturbation theory graphs and their statistical weights to sixth order.

$$\Gamma^H_{d\text{-dimensional dimer}} = \ln\left[\frac{1}{2} + \frac{1}{2}\left(1 + 8\sum_{i=1}^d z_i\right)^{1/2}\right] - \frac{1}{16\Sigma_{i=1}^d z_i}\left[-1 + \left(1 + 8\sum_{i=1}^d z_i\right)^{1/2}\right]^2. \tag{4.9}$$

Unfortunately, the d-dimensional dimer problem is unsolved for $d > 1$, so that comparison with the exact result is impossible.

It is common knowledge that the Hartree approximation sums up the "tadpole" vacuum bubbles. A sample tadpole graph is shown in Fig. 9. In terms of contracted graphs (that is, with fermion loops contracted to points) the tadpole graphs are the tree graphs. Knowing this allows one to compute systematically the corrections to the Hartree approximation. Let

$$g_\gamma = \langle\eta_\gamma\eta_\gamma^\dagger\rangle_H, \tag{4.10}$$

be the solutions to Eqs. (4.7). Then

$$\ln Z = \ln Z_H + \sum G_H, \tag{4.11}$$

with Z_H given in Eq. (4.8), and ΣG_H is the sum over connected Feynman graphs with rule (c) modified to

(c′) Allow only graphs with vertex degree $\geqslant 2$, i.e., graphs with one line coming into a vertex are to be excluded.

For each vertex α and l lines enamating from it put in a factor of $\Sigma_\alpha(-1)g_\alpha^l(l-1)!$.

Feynman graph rules (a), (b), (d), and (e) remain unchanged.

Eliminating graphs of order one reduces the number of graphs to be considered. Not only is the Hartree expansion better than simple perturbation theory over an extended region of z but it is easier to calculate. Figure 10 displays the statistical factors due to rules (c'), (d), and (e) for the graphs in G_H to sixth order in edge weight. The graphs still need to be multiplied by the g_α and $z_{\alpha\beta}$ factors of rules (b) and (c'). The terms in Fig. 10 generate a result guaranteed to be correct to order $z_{\alpha\beta}^6$ when expanded in powers of $z_{\alpha\beta}$. Thus an answer correct to z^6 for the general dimer problem has been obtained. In addition, the effects of higher order (in z) graphs have been included in the Hartree-improved expansion, so that the result can be expected to have a wider range of validity than a simple low temperature expansion. The terms in Fig. 10 to fourth order are

$$\begin{aligned}\ln Z = \ln Z_H &+ \frac{1}{4}\sum_{\alpha\beta} z_{\alpha\beta}^2 g_\alpha^2 g_\beta^2 - \frac{1}{6}\sum_{\alpha\beta\gamma} z_{\alpha\beta}z_{\beta\gamma}z_{\gamma\alpha}g_\alpha^2 g_\beta^2 g_\gamma^2 \\ &+ \frac{1}{3}\sum_{\alpha\beta} z_{\alpha\beta}^3 g_\alpha^3 g_\beta^3 - \sum_{\alpha\beta\gamma} z_{\alpha\beta}^2 z_{\beta\gamma}z_{\alpha\gamma}g_\alpha^3 g_\beta^3 g_\gamma^2 \\ &- \frac{3}{4}\sum_{\alpha\beta\gamma} z_{\alpha\beta}^2 z_{\beta\gamma}^2 g_\alpha^2 g_\beta^4 g_\gamma^2 \\ &+ \frac{1}{8}\sum_{\alpha\beta\gamma\delta} z_{\alpha\beta}z_{\beta\gamma}z_{\gamma\delta}z_{\delta\alpha}g_\alpha^2 g_\beta^2 g_\gamma^2 g_\delta^2 \\ &+ \frac{3}{4}\sum_{\alpha\beta} z_{\alpha\beta}^4 g_\alpha^4 g_\beta^4 + \cdots. \end{aligned} \tag{4.12}$$

The terms of fifth and sixth order can easily be written down but for reasons of space are omitted. Equation (4.12) and Fig. 10 constitute an important result in this section.

It is clear that the g_α factors can be absorbed into the $z_{\alpha\beta}$ factors: Equivalent to rules (b) and (c') are rules (c) and (b') with

(b') for each edge (Fig. 6) associate a factor of $g_\alpha z_{\alpha\beta} g_\beta$. The Hartree improved expansion is in powers of $\omega_{\alpha\beta} \equiv g_\alpha z_{\alpha\beta} g_\beta$ in contrast to $z_{\alpha\beta}$ for simple perturbation theory. In general the factor $g_\alpha z_{\alpha\beta} g_\beta$ will be smaller than $z_{\alpha\beta}$ and for $z_{\alpha\beta}$ large it should be considerably smaller. The Hartree perturbation series represents a marked improvement over the simple low temperature one. To illustrate this consider the one-dimensional dimer problem again. For large z the Hartree expansion is considerably better than the low temperature expansion and for low temperatures the Hartree expansion is just as good. Furthermore, the Hartree expansion parameter is always less than z and never bigger than $\frac{1}{2}$. Therefore, ω is less than the simple perturbation theory expansion parameter. It is always bounded being guaranteed to be at most of intermediate strength. In contrast z can become arbitrarily large. For the d-dimensional hypercubic dimer problem the Hartree expansion parameters are

$$\omega_i = z_i\left(\frac{-1+(1+8\Sigma_{i=1}^d z_i)^{1/2}}{4\Sigma_{i=1}^d z_i}\right)^2, \tag{4.13}$$

and are small or at most of intermediate magnitude. In fact, the ω_i cannot be greater than $\frac{1}{2}$ and in the isotropic case $z_1 = z_2 = \cdots = z_d$, $\omega_1 = \omega_2 = \cdots = \omega_d \equiv \omega \leqslant 1/2d$. It appears as if the expansion parameter, ω, becomes smaller as the dimension is increased, a point that will be discussed later.

One may also treat the combined monomer-dimer system. The action is given by

$$\begin{aligned} A^{\text{dimer-monomer}}_{(z_\alpha, z_{\alpha\beta})} &= \sum_\alpha z_\alpha \eta_\alpha \eta_\alpha^\dagger \\ &\quad + \frac{1}{2}\sum_{\alpha\beta} z_{\alpha\beta} :\eta_\alpha \eta_\alpha^\dagger \eta_\beta \eta_\beta^\dagger: \end{aligned} \tag{4.14}$$

which differs from A^{dimer} in that z_α, the Boltzmann factors for monomers, are not unity. By rescaling $\eta_\alpha \to (1/z_\alpha)\eta_\alpha$ (or by using simple physical reasoning) $A^{\text{dimer-monomer}}$ can be related to A^{dimer} so that $A^{\text{dimer-monomer}}$ is not any more general than A^{dimer}. However, this is not quite true. In Eq. (4.14) some of the z_α may be set equal to zero (in which case corresponding sites must be occupied by a dimer). The rescaling transformation fails. Simple perturbation theory is impossible since certain propagators blow up. Nevertheless the Hartree expansion exists because a finite Hartree propagator is generated. Thus even pure dimer systems may be treated. Equations are easily modified to account for Eq. (4.14). For example, the 1's in Eqs. (4.7) and (4.8) become z_γ and z_α. The point is that the Hartree expansion can handle the situation of having some (or all) monomer Boltzmann factors zero, whereas the usual low temperature expansion cannot.

V. DIMER MODELS (SPECIFIC LATTICES)

This section tackles the dimer problems on various lattices via the methods of the last section. These models are unsolved (except in the pure dimer limit for two-dimensional planar lattices[25]). Unlike the generic expansion [Eq. (4.5)], a specific dimer problem has lattice embedding factors for which it is useful to derive rules. Each term in Eq. (4.5) will generate several terms as the indices α, β, γ, etc. range over sites. It is useful to group these terms into a new set of diagrams and define new rules. This is similar to the usual graph and embedding theory.[26]

This section obtains new series expansions and accurately calculates physical quantities such as molecular freedoms, densities, and entropies. Models in two, three, and higher dimensions are considered. These computations test the accuracy of the Hartree expansion. It is found that it works amazingly well.

Rules for a Dimer Problem on a Translationally Invariant Lattice: Rules (a) and (d) of Sec. IV get modified to

(a) Draw all diagrams on the lattice with different shapes. Two graphs which are translates of each other but have the same shape are considered equivalent.

(d) Treat vertices with different locations as being distinct; then there is a factor of [order of the point symmetry group of the diagram]$^{-1}$.

(f) $\Gamma = (1/N)\ln Z = \Sigma_{\text{connected diagrams}}$ (weight of diagram).

Consider now the Hartree-improved expansion. The diagrammatic rules are the same as in "Rules for a Dimer Problem on a Translationally Invariant Lattice" with the

above substitutions ($z_h \to \omega_h$ and $z_v \to \omega_v$) in rule (b) the diagrams with vertices of degree one are ignored. There is also a zeroth order contribution given in Eq. (4.8).

As an example, here is the Hartree expansion to sixth order for the two-dimensional lattice:

$$\Gamma_{\text{2-d dimer}} = \ln(\tfrac{1}{2} + \tfrac{1}{2}(1 + 8(z_h + z_v))^{1/2}) + \Gamma^{(1)}(\omega_h) + \Gamma^{(1)}(\omega_v) + \Gamma^{(2)}(\omega_h, \omega_v) + \cdots, \tag{5.1}$$

where, for later convenience, terms of one-dimensional character are grouped into $\Gamma^{(1)}$ and terms of two-dimensional character are grouped into $\Gamma^{(2)}$:

$$\Gamma^{(1)}(\omega_h) = -\omega_h + \tfrac{1}{2}\omega_h^2 + \tfrac{2}{3}\omega_h^3 - \tfrac{3}{4}\omega_h^4 - \tfrac{6}{5}\omega_h^5 + \tfrac{5}{3}\omega_h^6 + \cdots, \tag{5.2}$$

$$\Gamma^{(2)}(\omega_h, \omega_v) = -3\omega_h^2\omega_v^2 + 4(\omega_h\omega_v^4 + \omega_h^4\omega_v) + 4(\omega_h^2\omega_v^3 + \omega_h^3\omega_v^2) + 15(\omega_h^2\omega_v^4 + \omega_h^4\omega_v^2) + \cdots. \tag{5.3}$$

A piece, $-\omega_h$, from the Hartree approximation has been regrouped into $\Gamma^{(1)}(\omega_h)$.

When expanded in powers of z_h and z_v, the low temperature expansion is recovered. Equation (5.1) will reproduce correctly terms to sixth order in z's. Equation (5.1) will be very accurate at low temperatures. Since the Hartree expansion includes the effects of some higher order graphs, Eq. (5.1) is also expected to be good over a domain larger than the low temperature one. In fact, even though it is a modified low temperature low density expansion, the infinite temperature limit can be taken. This is because as $z_h \to \infty$ and $z_v \to \infty$, ω_h and ω_v approach constants. At infinite temperature the problem becomes the close-packed dimer model which has been solved.[25] In the isotropic case (when $z_h = z_v = z$) the answer is

$$\Gamma^{\text{close-packed}} = \tfrac{1}{2}\ln z + G/\Pi \approx \tfrac{1}{2}\ln z + .2916 \tag{5.4}$$

with G, Catalan's constant. The Hartree expansion in Eq. (5.1) gives

$$\Gamma \underset{z\to\infty}{\to} \tfrac{1}{2}\ln z + .2803. \tag{5.5}$$

It is reassuring that the Hartree improved expansion is accurate in a region so far from its range of validity (low temperatures). This indicates that Eq. (5.1) is probably reasonably good over the entire range of z_h and z_v.

The Hartree expansion to sixth order has also been obtained for the d-dimensional hypercubic lattice. Define

$$\Gamma^{(0)}(\{\omega_j\}) = \ln\left[\tfrac{1}{2} + \tfrac{1}{2}\left(1 + 8\sum_{i=1}^{d}\omega_i\right)^{1/2}\right],$$

$$\Gamma^{(3)}(\omega_1, \omega_2, \omega_3) = 8(\omega_1\omega_2^2\omega_3^2 + \omega_1^2\omega_2\omega_3^2 + \omega_1^2\omega_2^2\omega_3) + 8\omega_1^2\omega_2^2\omega_3^2 - 16(\omega_1\omega_2^2\omega_3^3 + \omega_1\omega_2^3\omega_3^2 + \omega_1^2\omega_2\omega_3^3 + \omega_1^3\omega_2\omega_3^2 + \omega_1^2\omega_2^3\omega_3 + \omega_1^3\omega_2^2\omega_3) - 16(\omega_1\omega_2\omega_3^4 + \omega_1\omega_2^4\omega_3 + \omega_1^4\omega_2\omega_3), \tag{5.6}$$

$$\Gamma^{(4)}(\omega_1, \omega_2, \omega_3, \omega_4) = -32(\omega_1\omega_2\omega_3^2\omega_4^2 + \omega_1\omega_2^2\omega_3\omega_4^2 + \omega_1\omega_2^2\omega_3^2\omega_4 + \omega_1^2\omega_2\omega_3\omega_4^2 + \omega_1^2\omega_2\omega_3^2\omega_4 + \omega_1^2\omega_2^2\omega_3\omega_4),$$

where $\Gamma^{(0)}$ is the first piece of the Hartree approximation [Eq. (4.9)] and the ω_i are defined in Eq. (4.13). Then the Hartree expansion in d-dimensions is

$$\Gamma(\{\omega_j\}) = \Gamma_0(\{\omega_j\}) + \sum_{i=1}^{d}\Gamma^{(1)}(\omega_i) + \sum_{i_1<i_2}^{d}\Gamma^{(2)}(\omega_{i_1}, \omega_{i_2}) + \sum_{i_1<i_2<i_3}^{d}\Gamma^{(3)}(\omega_{i_1}, \omega_{i_2}, \omega_{i_3}) + \sum_{i_1<i_2<i_3<i_4}^{d}\Gamma^{(4)}(\omega_{i_1}, \omega_{i_2}, \omega_{i_3}, \omega_{i_4}) + \cdots, \tag{5.7}$$

where $\Gamma^{(1)}$ and $\Gamma^{(2)}$ are given in Eqs. (5.2) and (5.3) and the other Γ's are given in Eq. (5.6). The superscript on the Γ's refers to the dimension of the subspace of the imbedded diagram. Thus $\Gamma^{(n)}$ refers to those diagrams which are imbedded in an n dimensional subspace of d-dimensional space.

In Sec. IV it was pointed out that, in the isotropic case, the expansion parameter, ω, gets smaller as the dimension of the lattice gets bigger. For the hypercubic lattice, $\omega = 1/2d + O(1/d^{3/2})$. This indicates that as d increases, the Hartree expansion works better and Eq. (5.7) will be an excellent approximation. The situation, however, is not so clear because the number of graphical embeddings increases with d. Let $d(G)$ be the dimension of the maximum space in which a graph can be embedded. A rough estimate of the number of embeddings of G is $(2d)^{d(G)} + O((2d)^{d(G)-1})$ for d large. The weight of G goes like $(1/2d)^b$ where b is the number of bonds, so that the total effect of G behaves like

$$\sim \frac{1}{(2d)^{(b-d(G))}}. \tag{5.8}$$

By inspection, it is found that $b - d(G) \geqslant 1$ for all graphs so that the effect of a graph is damped by a power of d. Graph 1 of Fig. 10 has the leading behavior, decreasing like $1/d$. There are many (an infinite number of) next-to-leading order graphs (i.e. graphs 3, 5, 6, 14, 24, etc. of Fig. 10) which behave as $1/d^2$. Thus as $d \to \infty$ the contribution of any given graph gets smaller. The Hartree expansion is better when d is bigger. Explicit examination of several series also seems to verify this. It appears that results in higher dimensions become more accurate.

Because of this, the hypercubic dimer model is exactly solvable in the $d \to \infty$ limit. Trivial algebra yields

$$\Gamma^{\text{d-dim}}(z)^{\text{dimer}} \underset{d\to\infty}{\to} \tfrac{1}{2}\ln d + \tfrac{1}{2}\ln 2z - \tfrac{1}{2} + \frac{2}{(8zd)^{1/2}} + (1 - 1/z)\frac{1}{8d} + (1/12z - 1)/((8zd)^{1/2}2d) + O(1/d^2). \tag{5.9}$$

Equation (5.9) was obtained by blindly expanding the Hartree improved series in powers of $1/d$. It is clear from Eq. (5.9) than not only must $d \gg 1$ but also $dz \gg 1$ so that z cannot be too small. Equation (5.9) is one of the interesting results in this section.

Since the Hartree approximation and graph 1 of Fig. 10 were the only inputs in Eq. (5.9), Eq. (5.9) will hold for any uniform loose-packed lattice for which the vertex degree (coordination number), q, is large. For hypercubic lattices $q = 2d$. In fact the result holds for lattices not containing a

triangle so that triangle graphs (graph 2 of Fig. 10) are absent. This triangle graph can potentially be of order $(1/q)$. Because $\omega \sim 1/q$ these dimer models are exactly solvable in the $q \to \infty$ limit:

$$\Gamma^{\text{dimer}}_{(z)} \underset{q\to\infty}{\to} \frac{1}{2}\ln\frac{q}{2} + \frac{1}{2}(2z) - \frac{1}{2} + \frac{2}{(4zq)^{1/2}} + \left(1 - \frac{1}{z}\right)\frac{1}{4q} + \frac{1}{(4zq)^{1/2}q}\left(\frac{1}{12z} - 1\right) + O\left(\frac{1}{q^2}\right). \tag{5.10}$$

For lattices with triangles Eq. (5.10) is valid to order $(1/q^{1/2})$:

$$\Gamma^{\text{dimer, lattice with triangles}}_{(z)} \underset{q\to\infty}{\to} \frac{1}{z}\ln\frac{q}{2} + \frac{1}{2}\ln(2z) - \frac{1}{2} + \frac{2}{(4zq)^{1/2}} + O\left(\frac{1}{q}\right). \tag{5.11}$$

Both Eqs. (5.10) and (5.11) are only valid if $zq \gg 1$ as well as $q \gg 1$. It is interesting that dimer models are exactly solvable in this limit. In the pure dimer limit, Eqs. (5.10) and (5.11) give rough approximations for the molecular freedom. A comparison with exact and estimated freedoms is presented in Table I for several models. The lattices are the one-dimensional (1-d), simple quadratic (sq), tetratrahedral (t), simple cubic (sc), body-centered cubic (bcc), planar triangular (pt), and face-centered cubic (fcc) lattices. The latter two contain triangles and the results are not expected to be as good as lattices without triangles. The results are accurate to several per cent, even though the q value is not that large.

Gaunt[21] has calculated low temperature expansions for several dimer models. These included both two- and three-dimensional systems. The expansions were for the isotropic case in which all z_i's are equal. The low temperature expansions were computed for various lattices to these orders: the simple quadratic lattice to 15 orders, the planar triangle lattice to 10 orders, the tetrahedral lattice to 16 orders, the simple cubic lattice to 12 orders, the body-centered cubic lattice to 12 orders, and the face-centered cubic lattice to 8 orders. When expanded in powers of z the Hartree expansion to order n is guaranteed to reproduce the low temperature expansion to order n. Hence n orders of low temperature expansion uniquely determine n orders of Hartree expansion and Gaunt's series can be used to obtain the Hartree series to many orders. The Hartree series in the isotropic case has been calculated this way for the above-mentioned lattices. The results are

$$\Gamma^{\text{sq}}(\omega) = \ln\left(\frac{1}{1-4\omega}\right) - 2\omega + \omega^2 + 1\tfrac{1}{3}\omega^3 - 4\tfrac{1}{2}\omega^4 + 13\tfrac{1}{5}\omega^5 + 33\tfrac{1}{3}\omega^6 - 106\tfrac{2}{7}\omega^7 + 273\tfrac{3}{4}\omega^8 + 1432\tfrac{4}{9}\omega^9 - 2816\tfrac{4}{5}\omega^{10} + 6197\tfrac{9}{11}\omega^{11} + 63602\omega^{12} - 93974\tfrac{2}{13}\omega^{13} - 446\tfrac{6}{7}\omega^{14} + 2667238\tfrac{14}{15}\omega^{15} + \cdots, \tag{5.12}$$

$$\Gamma^{\text{pt}}(\omega) = \ln\left(\frac{1}{1-6\omega}\right) - 3\omega + 1\tfrac{1}{2}\omega^2 - 23\tfrac{1}{4}\omega^4 + 92\tfrac{2}{5}\omega^5 - 8\omega^6 - 1743\tfrac{3}{7}\omega^7 + 8202\tfrac{3}{8}\omega^8 - 1478\omega^9 - 196618\tfrac{1}{5}\omega^{10} + \cdots, \tag{5.13}$$

$$\Gamma^{\text{t}}(\omega) = \ln\left(\frac{1}{1-4\omega}\right) - 2\omega + \omega^2 + 1\tfrac{1}{3}\omega^3 - 5\tfrac{1}{2}\omega^4 + 5\tfrac{3}{5}\omega^5 + 21\tfrac{1}{3}\omega^6 - 66\tfrac{2}{7}\omega^7 + 186\tfrac{1}{4}\omega^8 + 472\tfrac{4}{9}\omega^9 - 2744\tfrac{4}{5}\omega^{10} + 4493\tfrac{9}{11}\omega^{11} + 19074\tfrac{2}{3}\omega^{12} - 91614\tfrac{2}{13}\omega^{13} + 192537\tfrac{1}{7}\omega^{14} + 952636\tfrac{4}{15}\omega^{15} - 3910844\tfrac{7}{8}\omega^{16} + \cdots, \tag{5.14}$$

$$\Gamma^{\text{sc}}(\omega) = \ln\left(\frac{1}{1-6\omega}\right) - 3\omega + 1\tfrac{1}{2}\omega^2 + 2\omega^3 - 11\tfrac{1}{4}\omega^4 + 68\tfrac{2}{5}\omega^5 - 41\omega^6 - 279\tfrac{3}{7}\omega^7 + 5688\tfrac{3}{8}\omega^8 - 12695\tfrac{1}{3}\omega^9 + 10999\tfrac{4}{5}\omega^{10} + 543356\tfrac{8}{11}\omega^{11} - 2067458\tfrac{1}{2}\omega^{12} + \cdots, \tag{5.15}$$

$$\Gamma^{\text{bcc}}(\omega) = \ln\left(\frac{1}{1-8\omega}\right) - 4\omega + 2\omega^2 + 2\tfrac{2}{3}\omega^3 - 15\omega^4 + 235\tfrac{1}{5}\omega^5 - 645\tfrac{1}{3}\omega^6 + 1979\tfrac{3}{7}\omega^7 + 30390\tfrac{1}{2}\omega^8 - 189343\tfrac{1}{9}\omega^9 + 1370054\tfrac{2}{5}\omega^{10} + 1393387\tfrac{7}{11}\omega^{11} - 35573416\omega^{12} + \cdots, \tag{5.16}$$

$$\Gamma^{\text{fcc}}(\omega) = \ln\left(\frac{1}{1-12\omega}\right) - 6\omega + 3\omega^2 - 4\omega^3 - 79\tfrac{1}{2}\omega^4 + 1192\tfrac{4}{5}\omega^5 - 10232\omega^6 + 48353\tfrac{1}{7}\omega^7 + 166814\tfrac{1}{4}\omega^8 + \cdots, \tag{5.17}$$

where ω is defined in terms of z and the coordination number, q, by

$$\omega = z\left(\frac{-1 + (1+4qz)^{1/2}}{2qz}\right)^2, \tag{5.18}$$

or

$$z = \omega/(1 - q\omega)^2. \tag{5.19}$$

The coordination numbers of the various lattices can be found in Table I.

By taking $z \to \infty$ and using the truncated series in Eqs. (5.12)–(5.17) the molecular freedoms at close packing can be calculated. These along with a comparison to other methods are shown in Table II. Rough error estimates are also included. As expected, more accuracy is obtained for models with larger q's. As an indication of what is obtainable "by hand" (that is, without the use of computers) sixth order computations are also shown. Even at this order molecular freedoms are correct to 1% or 2% for lattices with small q and to less

TABLE I. Molecular freedoms at close-packing as computed in the $1/q$ expansion for various lattices.

Model	q	$\frac{1}{q}$ estimate	Exact or best estimate	% error
1-d	2	.94	1	6%
sq	4	1.67	1.79	6.5%
t	4	1.67	1.70	2%
sc	6	2.40	2.45	2%
bcc	8	3.13	3.19	2%
pt	6	2.21	2.36	6.5%
fcc	12	4.41	4.57	3.5%

TABLE II. Molecular freedoms at close-packing as computed by the Hartree series with a comparison to other methods.

	sq	pt	t	sc	bcc	fcc
Exact	1.7916	2.3565	----	----	----	----
Bethe approximation	1.69	2.41	1.69	2.41	3.14	4.61
Nagle's series truncated	1.769	2.352	1.701	2.442	----	4.564
Nagle's series extended by Gaunt, truncated	1.773	2.360	1.701	2.451	3.189	4.565
Gaunt's Padé improved	1.78 – 1.80	2.356	1.702	2.449	3.198	4.570
Hartree approximation	1.47	2.21	1.47	2.21	2.94	4.41
Hartree series at sixth order	1.75 ± 0.03	2.37 ± 0.06	1.70 ± 0.02	2.44 ± 0.01	3.17 ± 0.01	4.56 ± 0.03
Hartree series, truncated	1.776 ± 0.009	2.347 ± 0.015	1.700 ± 0.003	2.449 ± 0.005	3.187 ± 0.003	4.574 ± 0.004

than 1% for lattices with larger q. At maximum order results are correct to within 0.1% for large q lattices and within $\frac{1}{2}$% for the low q lattices with the exception of the simple quadratic lattice where the error persists at 1%.

The dimer density, ρ, normalized so that at close-packing $\rho = 1/q$, is

$$\rho = \frac{2}{q} z \frac{d\Gamma}{dz}. \tag{5.20}$$

The quantity $\frac{1}{2}q\rho$ is the number of dimers per site, whereas ρ is the number of dimers per bond. The entropy, S, and molecular freedom, ϕ, are

$$S = -\rho \ln z + (2/q)\Gamma,$$
$$\phi = \exp(qS). \tag{5.21}$$

Tables III, IV, and V show the numerical values of ρ and S as a function of $\omega/\omega_{\max}$ [ω is the Hartree expansion parameter [Eq. (5.18)] and

$$\omega_{\max} \equiv 1/q, \tag{5.22}$$

is the maximum physical value of ω]. These numerical values were computed from the truncated series in Eqs. (5.12)–(5.17). The subscripts on ρ and S in Tables III, IV, and V indicate the orders at which the series were truncated.

Notice that lattices with the same coordination number (Table III and Table IV) have almost identical entropies and almost identical densities. Only at extremely high temperatures do they begin to deviate for different models. Mathematically the reason for this is simple: Models have a universal (as far as q is concerned) Hartree expansion to order ω^2:

$$\Gamma^q_{(\omega)} = \ln\left(\frac{1}{1-q\omega}\right) - \tfrac{1}{2}q\omega + \tfrac{1}{4}q\omega^2 + \text{ (nonuniversal)}. \tag{5.23}$$

Because the Hartree series at second order is already a good approximation models with the same q have almost identical properties. Furthermore in higher orders, they will have many identical Feynman graphs. In fact for lattices without triangles subgraphs, Eq. (5.23) is universal to third order

$$\Gamma^q_{(\omega)} = \ln\left(\frac{1}{1-q\omega}\right) - \tfrac{1}{2}q\omega + \tfrac{1}{4}q\omega^2 + \tfrac{1}{3}q\omega^3 + \text{ (nonuniversal)}. \tag{5.24}$$

Next, notice that ω is a good approximation to the density, ρ. For the simple quadratic and tetrahedral lattices, for the planar triangular, simple cubic, and body-centered cubic lattices and for the face-centered cubic lattice, ρ and ω never

TABLE III. The density, ρ, and the entropy, S, of the simple quadratic and tetrahedral dimer lattice models.

		Simple Quadratic Lattice		Tetrahedral Lattice	
$\frac{w}{w_{\max}}$	w	ρ_{15}	S_{15}	ρ_{16}	S_{16}
0.1	0.025	0.025534345332921949 ± (15)	0.116814864054067036 ± (37)	0.02553535977813695063 ± (60)	0.1168118264806264350 ± (21)
0.2	0.050	0.05180380074377 ± (41)	0.1949643466614 ± (12)	0.051790983870680 ± (32)	0.194927214968692 ± (84)
0.3	0.075	0.07838101424 ± (14)	0.25349328917 ± (39)	0.078318571157 ± (17)	0.253349894577 ± (34)
0.4	0.100	0.1049369153 ± (85)	0.295353081 ± (12)	0.1047525671 ± (13)	0.2950219984 ± (19)
0.5	0.125	0.13122827 ± (19)	0.32135808 ± (17)	0.130819262 ± (37)	0.320810503 ± (33)
0.6	0.150	0.1570755 ± (22)	0.33132011 ± (73)	0.15632504 ± (52)	0.33064735 ± (16)
0.7	0.175	0.182330 ± (16)	0.3240767 ± (45)	0.1811370 ± (42)	0.3235364 ± (13)
0.8	0.200	0.206800 ± (73)	0.296955 ± (74)	0.205148 ± (23)	0.296918 ± (23)
0.9	0.225	0.22998 ± (20)	0.24426 ± (38)	0.228209 ± (71)	0.24440 ± (14)

TABLE IV. The density and entropy for the dimer models on the triangular and simple cubic lattices.

		Planar triangular lattice		Simple cubic lattice	
$\frac{w}{w_{max}}$	w	ρ_{10}	S_{10}	ρ_{12}	S_{12}
0.1	$\frac{1}{60}$	0.01689214299601 ± (89)	0.0841945595015 ± (36)	0.0169006865736503 ± (31)	0.084231124247639 ± (12)
0.2	$\frac{2}{60}$	0.03405266733 ± (74)	0.1422176561 ± (23)	0.034114158049 ± (10)	0.142428661957 ± (32)
0.3	$\frac{3}{60}$	0.051266315 ± (34)	0.187111775 ± (85)	0.0514487921 ± (10)	0.1876342661 ± (27)
0.4	$\frac{4}{60}$	0.06838822 ± (49)	0.22104537 ± (94)	0.068760241 ± (27)	0.221940351 ± (51)
0.5	$\frac{5}{60}$	0.0853296 ± (35)	0.2446703 ± (49)	0.08594040 ± (31)	0.24589635 ± (42)
0.6	$\frac{6}{60}$	0.102041 ± (16)	0.257883 ± (14)	0.1029054 ± (21)	0.25930907 ± (66)
0.7	$\frac{7}{60}$	0.118500 ± (54)	0.259825 ± (17)	0.1195810 ± (93)	0.2612807 ± (20)
0.8	$\frac{8}{60}$	0.13470 ± (13)	0.248409 ± (39)	0.135880 ± (29)	0.249809 ± (13)
0.9	$\frac{9}{60}$	0.15069 ± (20)	0.21841 ± (16)	0.151661 ± (57)	0.220254 ± (63)

differ by more than about 5%, 3%, and 1%. The reason for this is simple. Equation (5.20) implies that

$$\rho = \langle z\eta_{\mathbf{x}}\eta_{\mathbf{x}}^{\dagger}\eta_{\mathbf{x}'}\eta_{\mathbf{x}'}^{\dagger}\rangle, \tag{5.25}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are nearest neighbors. In the Hartree approximation

$$\rho \approx z(\langle \eta_{\mathbf{x}}\eta_{\mathbf{x}}^{\dagger}\rangle_H)^2 \equiv \omega. \tag{5.26}$$

In other words, ω is the Hartree approximation to the density. From this point of view the Hartree series has a more physical flavor: it is an expansion in a parameter which is approximately the density.

Tables II, III, IV, and V have error estimates. The uncertainty is in the last two figures, so that, for example, the sq lattice at $\omega = 0.225$ has $\rho_{15} = 0.22998 \pm 0.00020$. These errors are set to the contribution of the last order. Doing this work only when the numerical coefficient of the maximum power of ω is not unnaturally small. This turns our to be the case for all the models considered. Since the Hartree series seem to converge, this is a rough but reasonable measure of the error. As a check, the exactly solvable one-dimensional dimer model can be used. Its Hartree expansion to 16th order is

$$\Gamma_{(\omega)}^{1-d} = \ln(1/(1-2\omega)) - \omega + \tfrac{1}{2}\omega^2 + \tfrac{2}{3}\omega^3 - \tfrac{3}{4}\omega^4 - 1\tfrac{1}{5}\omega^5 + 1\tfrac{2}{3}\omega^6 + 2\tfrac{6}{7}\omega^7 - 4\tfrac{3}{8}\omega^8 - 7\tfrac{7}{9}\omega^9 + 12\tfrac{3}{5}\omega^{10} + 22\tfrac{10}{11}\omega^{11} - 38\tfrac{1}{2}\omega^{12} - 71\tfrac{1}{13}\omega^{13} + 122\tfrac{4}{7}\omega^{14} + 228\tfrac{4}{5}\omega^{15} - 402\tfrac{3}{16}\omega^{16} + \cdots.$$

Table VI displays the approximated ρ_{16}, the exact ρ, the approximated S_{16}, and the exact S. The same error estimate method was used. As can be seen, the exact results always fall within the "error bar" region. In fact, estimated errors are roughly five times actual errors. For the one-dimensional model this is a conservative method of estimating errors.

Tables III, IV, and V show excellent accuracy. In 90% of the physical region (as measured by ω) the density and entropy are at least computed to 0.1% for all models. For the bcc and fcc lattices the minimal accuracy is about five decimal places. It is only for dense systems (i.e. 90% maximal dimer density) that errors are even of the above stated size. For example, at 10% maximal dimer density, results for the six models are accurate to an estimated 17, 18, 11, 14, 14, and 11 decimal places. As expected at low dimer densities best accuracy is achieved for those models for which the series has been computed to the most orders whereas at high densities best accuracy occurs for models with the highest q.

The general dimer model is an unsolvable model; it is an interacting fermionic field theory. No analytic or exact mathematical expressions exist for the free energy, density, entropy, etc. In this section a "physicist's solution" has been

TABLE V. The density and entropy for the dimer models on the bcc and fcc lattices.

		Body-centered cubic lattice		Face-centered cubic lattice	
$\frac{w}{w_{max}}$	ρ_{12}	S_{12}	ρ_8	S_8	
0.1	0.0126308078545463 ± (13)	0.0666064413803685 ± (54)	0.0083890254360 ± (42)	0.047655678485 ± (20)	
0.2	0.0254353696746 ± (42)	0.113584413678 ± (14)	0.01684367298 ± (88)	0.0821056143 ± (34)	
0.3	0.03830861920 ± (44)	0.1508592102 ± (12)	0.025312649 ± (18)	0.110062846 ± (59)	
0.4	0.051173247 ± (11)	0.180052915 ± (24)	0.03376250 ± (14)	0.13266446 ± (39)	
0.5	0.06397071 ± (13)	0.20157738 ± (21)	0.04217243 ± (67)	0.1502187 ± (15)	
0.6	0.07665304 ± (75)	0.21524997 ± (92)	0.0505306 ± (22)	0.1626165 ± (36)	
0.7	0.0891737 ± (38)	0.2203287 ± (19)	0.0588309 ± (53)	0.1693450 ± (60)	
0.8	0.101475 ± (12)	0.2152042 ± (20)	0.067070 ± (10)	0.1692448 ± (59)	
0.9	0.113467 ± (23)	0.196182 ± (19)	0.075243 ± (12)	0.1595724 ± (42)	

TABLE VI. A comparison of the exact density and entropy to the Hartree estimated density and entropy for the one-dimensional dimer model.

$\frac{w}{w_{max}}$	w	ρ_{16}	ρ_{exact}	S_{16}	S_{exact}
0.1	0.05	0.0522332644055048897 ± (81)	0.0522332644055048890	0.202159043168350649 ± (23)	0.202159043168350647
0.2	0.10	0.10776772972370 ± (43)	0.10776772972363	0.32877414039132 ± (83)	0.32877414039119
0.3	0.15	0.16476080022 ± (23)	0.16476080017	0.41476557223 ± (30)	0.41476557217
0.4	0.20	0.221456997 ± (18)	0.221456993	0.464874051 ± (13)	0.464874048
0.5	0.25	0.27639331 ± (50)	0.27639320	0.481211845 ± (94)	0.481211825
0.6	0.30	0.3285028 ± (70)	0.3285014	0.4652904 ± (25)	0.4652909
0.7	0.35	0.377125 ± (58)	0.377115	0.417789 ± (58)	0.417798
0.8	0.40	0.42195 ± (31)	0.42191	0.33720 ± (53)	0.33726
0.9	0.45	0.46291 ± (97)	0.46284	0.2158 ± (25)	0.2159

obtained, that is, expressions accurate to four or five decimal places in the entire physical region. This is a significant achievement. In effect, the Hartree series has "solved" an important class of unsolvable models.

VI. CONCLUSION

For the dimer model the Hartree approximation has the following physical interpretation. Consider a particular dimer configuration. Erase the bonds. What remains is a collection of monomers. It is reasonable that a dimer system can be approximated by a monomer one. As seen from Eq. (4.6) the Hartree approximation is an attempt to find a good monomer approximation. This, in fact, is the basis for many approximation schemes: to find a quadratic action (or a solvable system) which approximates an unsolvable model. In general, it requires ingenuity to find the right perturbing model. The relevant degrees of freedom must be extracted. But once found, a few correction orders yields the physics of an unsolved model. This is what has been done with the dimer model.

These papers have demonstrated the power of anticommuting variables. Models, which are solvable, are trivially solved. For models which are unsolvable there are powerful approximation methods. I have chosen simple but interesting models to exemplify the techniques. However, the anticommuting variable method is applicable to a wide range of systems. Whenever there is a constraint that objects cannot overlap (be it polygons, polymers, or surfaces) the anticommuting variables will be useful.

Although new results have been obtained, much more can be done: the Ising model in three dimensions has been expressed in anticommuting variable form and is thus amenable to new approximation schemes. Results for general ferroelectric vertex models and Ising type models will be forthcoming.[4,6,7,8] Additional results for dimer and polymer systems will also be published.[4,6] This new research area is still in its incipience. Many more models can be treated. Many more techniques can be developed. The most important progress can be made in the area of critical phenomenon. What is needed is an adaptation of renormalization group methods. In short, this body of work is a small piece of what can be done with anticommuting variables.

ACKNOWLEDGMENTS

Concerning these three papers, I thank Harry Morrison and Korkut Bardakci for guidance and making useful suggestions, Eliot Lieb for advice concerning the format of these papers, Eliezer Rabinovici for some initial assistance with the computer work, Luanne Neumann for the typing, and Antoinette Czerwinski for assistance with the figures.

[1] S. Samuel, J. Math. Phys. **21**, 2806 (1980). References to the figures and equations of this paper are prefaced by a I.
[2] S. Samuel, J. Math. Phys. **21**, 2815 (1980). References to the figures and equations of this paper are prefaced by a II.
[3] A review of this work can be found in S. Samuel, in the proceedings of the 5th Workshop On Current Problems in High Energy Particle Theory, Bad Honnef, Germany.
[4] S. Samuel, papers in progress.
[5] Some time after the completion of the original version of this manuscript (LBL preprint 9347, June 1979) a paper by E. Fradkin, M. Srednicki, and L. Suskind (U. of Illinois preprint) appeared. They, too, have obtained this representation by using the results of Refs. 1 and 2. They also have obtained a fermionic Hamiltonian formulation of the three-dimensional Ising model.
[6] S. Samuel, "The Use of Anticommuting Integrals in Statistical Mechanics III," LBL preprint 9347 (June 1979).
[7] S. Samuel, "The Pseudo-Free 128 Vertex Model," to be published in J. Phys. A.
[8] S. Samuel, "The Correlation Functions in the 32-Vertex Model," [IAS preprint, March 1980).
[9] R. Balian, J. M. Drouffe, and C. Itzykson, Phys. Rev. D **11**, 2098 (1975); Phys. Rev. D **11** 2104 (1975); E. Fradkin and S. H. Shenker, Phys. Rev. D **19**, 3682 (1979).
[10] F. J. Wegner, J. Math. Phys. **12**, 2259 (1971).
[11] J. K. Roberts, Proc. Soc. (London) A **152**, 469 (1935); See also the references in Ref. 12.
[12] O. J. Heilmann and E. H. Lieb, Commun. Math. Phys. **25**, 190 (1972).
[13] R. H. Fowler and G. S. Rushbrooke, Trans. Faraday Soc. **33**, 1272 (1937); F. H. Ree and D. A. Chestnut, Phys. Rev. Lett. **18**, 5 (1967); A.Bellemans and R. K. Nigam, J. Chem. Phys. **46**, 2922 (1967); J. Orban and A. Bellemans, J. Chem. Phys. **49**, 363 (1968).
[14] J. K. Roberts, Proc. R. Soc. (London) A **161**, 141 (1937); Proc. Cambridge Philos. Soc. **34**, 399 (1938); P. A. Readhead, Trans. Faraday Soc. **57**, 641 (1961); D. R. Rossington and R. Bost, Surf. Sci. **3**, 202 (1965).
[15] E. H. Lieb, J. Math. Phys. **8**, 2339 (1967); R. J. Baxter, J. Math. Phys. **9**, 650 (1968); L. K. Runnels, J. Math. Phys. **11**, 842 (1970).
[16] T. S. Chang, Proc. Cambridge Philos. Soc. **35**, 265 (1939); J. K. Roberts and A. R. Miller, Proc. Cambridge Philos. Soc. **35**, 293 (1939).
[17] R. D. Kaye and D. M. Burley, Physica **87** A, 499 (1977).
[18] J. F. Nagle, Phys. Rev. **152**, 190 (1966).
[19] C. Gruber and H. Kunz, Commun. Math. Phys. **22**, 133 (1971).
[20] G. S. Rushbrooke, H. I. Scoins, and A. J. Wakefield, Discus. Faraday Soc. **15**, 57 (1953).
[21] D. S. Gaunt, Phys. Rev. **179**, 174 (1969).
[22] L. Degreve, Physica **66**, 395 (1973).
[23] J. Van Craen and A. Bellemans, J. Chem. Phys. **56**, 2041 (1972).

[24]A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems*, (McGraw–Hill, San Francisco, 1971); D. Pines, *The Many-Body Problem* (Benjamin, New York, 1961).
[25]H. N. V. Temperley and M. E. Fisher, Philos. Mag. 6, 1061 (1960); M. E. Fisher, Phys. Rev. 124, 1664 (1961).
[26]See the articles in Volume 3 of *Phase Transitions and Critical Phenomena*, edited by C. Domb and M. S. Green (Academic, New York, 1974).

Geometrical reinterpretation of Faddeev–Popov ghost particles and BRS transformations

Jean Thierry-Mieg
Groupe d'Astrophysique Relativiste, Observatoire de Meudon, Meudon, 92190, France

(Received 7 May 1979; accepted for publication 16 November 1979)

A classical geometrical interpretation of the ghosts fields is presented. BRS rules follow from the Cartan–Maurer fibration theorem. The statistics of ghosts are explained and the effective quantum Lagrangian is derived without factorizing the volume of the gauge group. Topologically nontrivial ghost configurations are defined.

1. INTRODUCTION

Because of gauge invariance, the classical Yang–Mills Lagrangian does not define a propagator for the gauge field.[1] Using the path integral formulation of quantum field theory, Faddeev and Popov[2] attributed this effect to the overcounting of gauge equivalent configurations. By fixing the gauge, Feynman diagrams are generated but unitarity is lost[3] unless additional quantum fields are introduced: the ghost particles. However, the effective Lagrangian still supports a global invariance of a new kind, the nilpotent BRS transformation,[4] which by itself implies the renormalizability of the theory.[5,6] The geometrical meaning of this symmetry has already been partly explained[7,8] but the picture is here completed.

Yang-Mills gauge theories are naturally described as geometrical theories over a principal bundle $\mathcal{F}$. Now, in Sec. 3, it is shown that the independent mathematical field of the theory, the connection 1 form ω, actually describes at the same time both the Yang–Mills gauge particle and the Faddeev–Popov ghost particle. With respect to a section, i.e., a gauge being chosen, the connection actually splits into the sum of two components: the gauge field φ which is horizontal and the ghost field χ which is normal to the section. By assumption, the ghost does not contribute to the description of motions tangent to the section. The exterior differential over $\mathcal{F}$ of a function also splits, and its component normal to the section is recognized as the BRS operator. Further, the Cartan–Maurer structural theorem, which states the compatibility of the connection with the fibration, implies the BRS transformation rules of the gauge and ghost fields. Moreover, the ghost does not contribute to the curvature 2 form (field strength) and may be thus eliminated from the description of the classical theory.

Section 4 is devoted to the study of gauge transformations. The identification of the infinitesimal active gauge transformations, generated by moving the section, with the passive gauge transformations, generated by relabeling the coordinates in the fiber, is shown possible only if the matter fields satisfy their own BRS transformation law as a constraint. Under those ordinary transformations the ghost is invariant. Another kind of gauge transformation may however be defined such that the ghost field becomes no longer trivial. By relaxing slightly the axiom of local triviality, modifications of the topology of the fiber bundle are then allowed in a way that seems adapted to the construction of the quantum theory and the study of soliton configurations.

In Sec. 5, the construction of the effective Lagrangian by using the generating functional is revisited. No infinite constant has to be extracted, as the differential of the volume element of the group is actually lifted into the effective Lagrangian in the form of the ghost. The nongeometric transformation of the antighost, a Lagrange multiplier, is not recovered. However, the proof of renormalizability is not altered by the noninvariance of the effective Lagrangian, as one usually cancels the antighost variation via its equation of motion. On the contrary, the renormalized BRS operator is shown, as geometry suggests, not to act on the antighost.

Despite its formal character, this study may have various applications. At the local level, the statistics of the ghost are simply those of a classical 1 form; and the relation of the BRS operator with gauge transformations is made explicit. At the classical global level, nontrivial topologies of the fiber bundle may be studied by including nontrivial configurations of the ghost field, or by working directly with the gauge

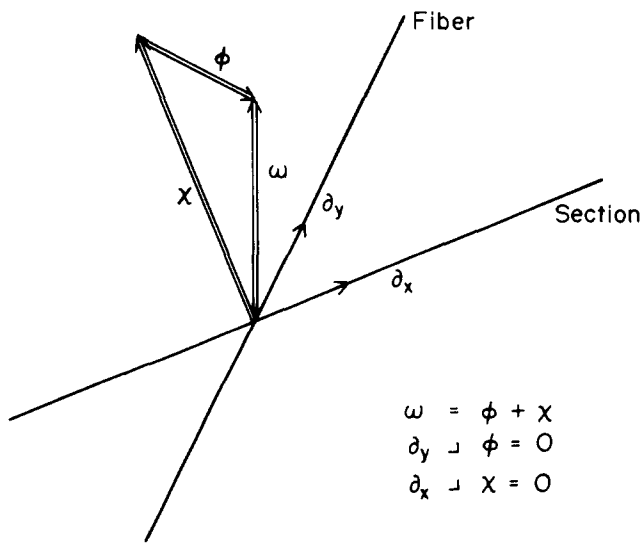


FIG. 1. The ghost and the gauge field: The single lines represent a local coordinate system of a principal fiber bundle of base space–time. The double lines are 1 forms. The connection of the principal bundle ω is assumed to be vertical. Its contravariant components φ and χ are recognized, respectively, as the Yang–Mills gauge field and the Faddeev–Popov ghost form.

 0022-2488/80/122834-05$1.00

independent globally defined connection ω whose interpretation as the sum of the ghost and gauge field is provided here, thus overcoming the obstruction to the use of a global section. Finally, the formulation of Yang–Mills theory over the fiber bundle itself and the precise understanding of the geometrical role of the ghost field provides a link towards the new promising approach to gravity: the soft group manifold.[9]

2. THE PRINCIPAL FIBER BUNDLE

A principal fiber bundle $(\mathcal{F}, \mathcal{B}, \Pi, \mathcal{G}, \cdot)$[10,11] is the true arena of a pure Yang–Mills theory.[12] The fiber bundle $\mathcal{F}$ and base space $\mathcal{B}$ are $\mathcal{C}^\infty$ manifolds. The projection Π is a $\mathcal{C}^\infty$ mapping of $\mathcal{F}$ onto $\mathcal{B}$. The point $\cdot$ denotes the action of the (graded) Lie group $\mathcal{G}$ in $\mathcal{F}$. The motions are assumed to preserve the fiber:

$$\Pi : \mathcal{F} \to \mathcal{B},$$
$$\cdot : \mathcal{F} \times \mathcal{G} \to \mathcal{F},$$
$$\forall u \in \mathcal{F},\ \forall a \in \mathcal{G},\ \Pi(u \cdot a) = \Pi(u).$$

The motions represent the group:

$$\forall u \in \mathcal{F},\ \forall a,b \in \mathcal{G},\ (u \cdot a) \cdot b = u \cdot (ab).$$

The last multiplication is the group operation. Further, the space is locally trivial, i.e., any point x of the base space (space–time) possesses a neighborhood V_x such that an isomorphism t exists between $\Pi^{-1}(V_x)$ and the direct product $V_x \times \mathcal{G}$:

$$t: \Pi^{-1}(V_x) \to V_x \times \mathcal{G},$$
$$u \to (\Pi(u), \tau(u)),$$
$$\tau(u \cdot a) = \tau(u) a.$$

The relevance of this axiom in physics will be further analyzed in Sec. 4. The point operation induces a map $\sim$ from the generators, y of the gauge group into Killing vector fields $\tilde{y}$ which span the space tangent to the fibers. The point being a representation of $\mathcal{G}$, $\sim$ is an isomorphism of Lie algebras, from the gauge algebra $\mathcal{A}$ onto the Killing algebra structured by the Poisson bracket:

$$\sim : \mathcal{A} \to \mathcal{F}_*$$
$$Y \to \tilde{Y}$$
$$[Y, Y']_{\mathrm{LB}} = [\tilde{Y}, \tilde{Y}']_{\mathrm{PB}}.$$

These Killing vector fields are called vertical. However, no horizontal vectors are yet specified. Rather than to give a metric on $\mathcal{F}$ (Kaluza–Klein theory), it is weaker to define an $\mathcal{A}$-valued vertical 1 form ω: the connection (Yang–Mills theory). ω maps vectors of $\mathcal{F}_*$ on $\mathcal{A}$:

$$\omega : \mathcal{F}_* \to \mathcal{A},$$
$$v \to v \lrcorner \omega = \omega(v) = \omega^i(v) Y_i.$$

The symbol $\lrcorner$ denotes the contraction of vectors with forms. The components ω^i are just ordinary 1 forms and are defined with respect to a basis y_i of $\mathcal{A}$. The kernel of ω then defines a subspace H of $\mathcal{F}_*$ called horizontal:

$$h \in H \subset \mathcal{F}_* \Leftrightarrow \omega(h) = 0.$$

The 2 form of curvature is defined as

$$\Omega = d\omega + \tfrac{1}{2}[\omega, \omega].$$

To be meaningful, this structure must be compatible with the vertical motions, and it is assumed that the restriction of the connection to the fiber is the pull back of the left invariant 1 forms of the gauge group:

$$\forall Y \in \mathcal{A},\ \tilde{Y} \lrcorner \omega = Y.$$

Moreover, the Lie derivative of the connection with respect to vertical vector fields is constrained by the equivariance condition

$$\mathcal{L}_{\tilde{Y}} \omega = -[Y, \omega] = +[\omega, Y].$$

We use the sign convention that, when contracting a p form with vectors, one must contract from the inside to get a plus sign (a convention adapted to supergroups):

$$v' \lrcorner v \lrcorner \omega \wedge \omega' = \omega(v)\omega'(v') - \omega(v')\omega'(v),$$
$$\tfrac{1}{2}\tilde{Y} \lrcorner [\omega, \omega] = [Y, \omega],$$
$$[\omega, \omega] = [\omega^i Y_i, \omega^j Y_j] = -\omega^j \wedge \omega^i f_{ij}^k Y_k.$$

The Lie derivative is a natural extension of the ordinary derivation; identical to the former when acting on functions, it is defined as the Poisson bracket when acting on vector fields:

$$\forall v, v' \in \mathcal{F}_*, \quad \mathcal{L}_v v' = [v, v']_{\mathrm{PB}}.$$

For Killing vector fields we get

$$\mathcal{L}_{\tilde{X}} \tilde{Y} = [\tilde{X}, \tilde{Y}]_{\mathrm{PB}} = [X, Y].$$

The Lie derivative obeys the Liebnitz rule

$$\forall \rho \in \mathcal{F}^*\ \mathcal{L}_v(v' \lrcorner \rho) = (\mathcal{L}_v v') \lrcorner \rho + v' \lrcorner \mathcal{L}_v \rho.$$

For Killing vector fields we get

$$\mathcal{L}_{\tilde{X}}(\tilde{Y} \lrcorner \omega) = [X, Y] \lrcorner \omega + \tilde{Y} \lrcorner \mathcal{L}_{\tilde{X}} \omega$$
$$= [X, Y] + [Y, X] = 0.$$

The left hand side of this equation, being the derivative of a constant, vanishes and the two axioms are indeed compatible. Further, Lie derivation, exterior differentiation, and contraction are related:

$$d(v \lrcorner \rho) = \mathcal{L}_v \rho - v \lrcorner d\rho,$$

yielding

$$d(\tilde{Y} \lrcorner \omega) = \mathcal{L}_{\tilde{Y}} \omega - \tilde{Y} \lrcorner d\omega.$$

The left hand side again vanished; thus,

$$\tilde{Y} \lrcorner d\omega = -[Y, \omega] = -\tfrac{1}{2}\tilde{Y} \lrcorner [\omega, \omega].$$

Accordingly, the 2 form of curvature Ω is purely horizontal:

$$\tilde{Y} \lrcorner \Omega = 0.$$

This very important theorem is known as the Cartan–Maurer structural condition. Over a Lie group, the curvature constructed out of the left invariant 1 forms identically vanishes, but over a fiber bundle, Ω is horizontal because the connection is only subject to the equivariance condition along each fiber, the fibers over different points being independent.

Because 1 forms anticommute, the connection fulfils the Jacobi identity in the form

$$[\omega, [\omega, \omega]] = 0.$$

The Bianchi identity follows:

$$D\Omega = d\Omega + [\omega,\Omega] = 0.$$

The efficiency of the notations of exterior calculus is apparent in the simple aspect of these two identities.

3. PHYSICAL INTERPRETATION

In Yang–Mills theory, the base space $\mathscr{B}$ is identified with space time and $\mathscr{G}$ is of course the gauge group. A gauge choice is a one to one map Σ, called a section, of $\mathscr{B}$ into $\mathscr{F}$:

$$\Sigma\mathscr{B}\to\mathscr{F},$$

$$\forall x\in\mathscr{B}\ \Pi(\Sigma(x)) = x.$$

No global section exists if the topology of $\mathscr{F}$ is nontrivial (monopoles). However, local triviality ensures the existence of a local section Σ, and it is possible to choose a local coordinate system in $\mathscr{F}$ adapted to the section as follows: Let y^i be coordinates in the fiber and x^μ be the lift in Σ of the coordinates of the base space. Thus, the vector ∂_{y^i} is tangent to the fiber and vertical, whereas the vector ∂_{x^μ} is tangent to the section but neither vertical nor horizontal. The 1 forms dy^i, dx^μ span $\mathscr{F}^*$ and one may decompose the connection1 form on this cobasis:

$$\omega = \chi_i\, dy^i + \varphi_\mu\, dx^\mu.$$

The vertical connection form ω splits into two components that will be later identified as the gauge and the ghost field of the quantum field theory. The gauge field

$$\varphi = \varphi_\mu\, dx^\mu$$
$$\Rightarrow \partial_{y^i} \lrcorner \varphi = 0$$

may be called horizontal, because all the vertical Killing vectors belong to its kernel. The ghost field

$$\chi = \chi_i\, dy^i$$
$$\Rightarrow \partial_{x^\mu} \lrcorner \chi = 0$$

may be called normal to Σ as the vectors tangent to Σ belong to its kernel. It is recalled that ω was assumed to be vertical, defining the horizonal vectors as those belonging to its kernel. The decomposition of ω is presented in the picture. In the same manner, the exterior differential df of a 0 form f may be written as

$$df = sf + bf,$$

where s and b are defined as

$$sf = \partial_{y^i} f\, dy^i, \quad bf = \partial_{x^\mu} f\, dx^\mu.$$

The fundamental rule of cohomology then implies

$$b^2 = sb + bs = s^2 = 0.$$

s, here defined as the exterior differential normal to the section, is nilpotent and will be identified with the BRS operator and the letter s stands for the name of Stora, whereas b is a horizontal operator and the letter b stands for base space or for the name of Becchi. (The author is sorry for A. Rouet.) b can also be viewed as the pullback of d onto the base space by the section Σ:

$$\Sigma^*(df) = \Sigma^*(bf),$$

$$\Sigma^*(sf) = 0.$$

On the other hand, choosing the local trivialization τ which maps the section Σ onto the identity of the group

$$\forall x\in\Sigma, \quad \tau(x) = 1_G,$$

The ghost form χ appears as the pull back onto $\mathscr{F}^*$ of the Cartan left invariant form of the Lie group.

With respect to the section Σ, the 2 form of curvature breaks into three pieces:

$$\Omega = \tfrac{1}{2}\Xi_{ij}\, dy^i{}_\wedge\, dy^j + \Psi_{i\mu} dy^i{}_\wedge\, dx^\mu + \tfrac{1}{2}\Phi_{\mu\nu}\, dx^\mu_\wedge\, dx^\nu.$$

Ξ is evaluated by expanding Ω in terms of its components and picking the terms with two dy:

$$\Xi = s\chi + \tfrac{1}{2}[\chi,\chi].$$

Ψ and Φ are the terms in $dy_\wedge\, dx$ and $dx_\wedge\, dx$ in the same expansion:

$$\Psi = s\varphi + b\chi + \tfrac{1}{2}([\chi,\varphi] + [\varphi,\chi])$$
$$= s\varphi + b\chi + [\varphi,\chi] = s\varphi + B\chi,$$
$$\Phi = b\varphi + \tfrac{1}{2}[\varphi,\varphi].$$

The evaluation of Ψ uses the fact that a skew Lie bracket acting on anticommuting 1 forms defines a symmetric operation. By the Cartan–Maurer structural theorem, which follows from the equivariance condition, curvature is purely horizontal. Thus,

$$\Xi = \Psi = 0.$$

The curvature is completely specified once φ and $b\varphi$ are known over a section Σ. χ is an auxiliary field which satisfies the constraints

$$\Xi = s\chi + \tfrac{1}{2}[\chi,\chi] = 0,$$
$$\Psi = s\varphi + B\chi = 0.$$

The equations may be recognized as the Becchi–Rouet–Stora transformations of the quantum field theory, justifying the identification of χ as the Faddeev–Popov ghost field. Accordingly, the ghost field has a classical meaning but may be excluded from local problems. It must not be considered as a genuine quantum entity. Its so-called wrong statistics are now explained. Customarily, one works with the components φ_μ of φ but does not decompose the 1 form χ which therefore anticommutes with itself and with the exterior differentials b and s. The ghost is not a Fermion; it is a Bose 1 form and commutes with any Fermi function. By example, in quantum supergravity, the ghost for the local translations commutes with the spin 3/2 gauge field of supersymmetry. The fields of this model are thus doubly graded[13] by ghost and Fermi number. The associated $\mathbb{Z}^2\times\mathbb{Z}^2$ sign rules for closed loops are necessary for proving unitarity.

4. GAUGE TRANSFORMATIONS

According to our definitions, an infinitesimal gauge transformation of parameter $\lambda^i(x^\mu)$ is induced by moving from one section Σ to a neighboring section Σ'. In the coordinates adapted to Σ, the equations of Σ and Σ' are

$$\Sigma : y^i = 0,$$

$$\Sigma' : y^i = -\lambda^i(x^\mu).$$

The x^μ coordinates, lifted from the base, are used both in Σ and Σ', and the y^i may be chosen such that over Σ, their

tangent vectors ∂y^i are the Killing vectors $\tilde{y}_i$. The parameters λ^i are defined only over the section Σ but we may extend them in an arbitrary way as fields over the whole of $\mathscr{F}$:

$$\lambda'^i = \lambda'^i(x^\mu, y^j),$$

$$\lambda'^i(x^\mu, 0) = \lambda^i(x^\mu).$$

Using the linear map $\sim$ from the Lie algebra $\mathscr{A}$ onto the Killing vector fields

$$\sim : Y_i \to \tilde{Y}_i(x^\mu, y^j),$$

we may use λ' to define a vertical vector field $\tilde{\Lambda}$ as

$$\tilde{\Lambda} = \lambda'^i(x^\mu, y^j)\, \tilde{Y}_i(x^\mu, y^j).$$

The connection over Σ' is deduced from its value over Σ by adding the Lie derivative:

$$\begin{aligned}\mathscr{L}_{\tilde{\Lambda}}\omega &= \tilde{\Lambda} \rfloor\, d\omega + d(\tilde{\Lambda} \rfloor \omega) \\ &= -[\tilde{\Lambda} \rfloor \omega, \omega] + d\Lambda \\ &= d\Lambda + [\omega, \Lambda] = D\Lambda.\end{aligned}$$

The component in $dx(dy)$ of this equation is the gauge transformation law of the gauge (ghost) field

$$\delta_\Lambda \varphi = B\Lambda = b\Lambda + [\varphi, \Lambda],$$

$$\delta_\Lambda \chi = S\Lambda = s\Lambda + [\chi, \Lambda].$$

Moving the section Σ, an active transformation, is indeed identical to relabeling the coordinates y^i in the fiber, a passive transformation. When the condition $S\Lambda = 0$ is met, the ghost which is the pullback of the left invariant Cartan form over the group is indeed invariant, and the gauge transformation parametrized by Λ corresponds to a passive left translation in the group. More generally, if a matter field belonging to any representation with matrices $\tau_i{}^A{}_B$ of the gauge group is defined by a set f^A of real valued functions over the principal bundle itself, the active and passive gauge transformations will coincide only if these functions satisfy the BRS constraint

$$Sf^A = 0,$$

$$\Leftrightarrow sf^A = -\chi^i \tau_i{}^A{}_B f^B.$$

This usual concept of a gauge transformation may be generalized. The gauge parameter Λ itself does not necessarily fulfill the above condition. When $S\Lambda$ does not vanish, we shall speak of a ghost transformation. χ remains a pure gauge field and the Cartan–Maurer–Becchi–Rouet–Stora conditions are not affected; however, the ghost is no longer invariant. Restricting our attention to a single fiber, we see that this active transformation can be compensated for by a general transformation of the group coordinates y^i. The ghost plays along the fiber the role of the vielbein of general relativity.

The restriction of a ghost transformation to a single fiber is a map from the gauge group onto itself. If this map is not diffeomorphic to the identity, the ghost acquires a topological charge and is no longer the global pullback of the group left invariant forms, but a trivialization isomorphism still holds locally from a neighborhood of any point in $\mathscr{F}$ to the product of a neighborhood in $\mathscr{B}$ of the projection of that point by a neighborhood of the identity in $\mathscr{G}$

$$\forall u \in \mathscr{F}, \quad \exists t, \quad V_u^{\mathscr{F}} \to V_{\Pi(u)}^{\mathscr{B}} \times V_e^{\mathscr{G}}.$$

This local version of the trivialization axiom seems, on the other hand, very well adapted to the construction of a quantum theory in which the local Maurer–Cartan BRS constraint is easily imposed, whereas it is very difficult to include the global topological condition on the ghost required by the usual axiom. (The same difficulty is met in Ref. 9 in which the dynamics impose spontaneous fibration of the group manifold, but only locally.) As a result, even if a global section Σ exists, the fiber bundle endowed with this restricted structure does not necessarily have the topology of the direct product $\Sigma \times \mathscr{G}$ and the quantum theory may include in a natural way a number of soliton configurations.[14]

In the framework of conventional fiber bundles, the Gribov–Singer problem may also be overcome by working directly with the connection ω, the sum of the ghost and gauge field, which is gauge independent and globally defined over $\mathscr{F}$.

5. THE FADDEEV–POPOV EFFECTIVE LAGRANGIAN

It remains to be shown that our definition of the ghost field coincides with the usual one.[2] The Lagrangian in Yang–Mills theory is defined in terms of the curvature 2 form as

$$\mathscr{L} = \Omega^i \wedge {}^*\Omega_i.$$

The trace is with respect to the Killing metric of the gauge group and the asterisk denotes the Hodge adjoint of the 2 form of curvature with respect to a given metric in the Base space:

$$\Omega^i = \tfrac{1}{2}\Omega^i{}_{\mu\nu}\, dx^\mu \wedge dx^\nu,$$

$${}^*\Omega^i = \tfrac{1}{4}\Omega^i{}_{\mu\nu}\epsilon^{\mu\nu}{}_{\rho\sigma}\, dx^\rho \wedge dx^\sigma.$$

Whenever the BRS conditions are satisfied, the Lagrangian is horizontal, does not depend on the ghost field, and is gauge invariant. By patching local sections it is thus possible to integrate $\mathscr{L}$ over the base space $\mathscr{B}$. The quantum theory is then constructed by summing over all configurations of the connection satisfying the BRS constraint and the generating functional of the Green's functions is defined as

$$W = \int \prod_{x,y} \mathscr{D}\omega(x,y)\delta(\mathrm{BRS}) e^{-i\int_{\mathscr{B}} \mathscr{L}}.$$

However, as noted by Faddeev and Popov, gauge equivalent configurations must not be overcounted. Thus, it is supposed that, at least locally, a set of constraints $\Sigma^i(\varphi)$ exists which is satisfied only once in every gauge equivalence class (ghost transformations are not involved here since Σ^i does not depend on χ):

$$\Sigma^i(\varphi) = 0.$$

As one integrates over all possible ω, one also integrates along the gauge classes. This is equivalent, however, to integrating over a moving section with fixed ω and this additional contribution is canceled by using the Dirac measure associated with the constraints:

$$\int \prod_i \delta(\Sigma^i) \wedge d\Sigma^i = 1.$$

The determinant of the constraints is hidden in the exterior product. $\Sigma^i = 0$ defines a section Σ in $\mathscr{F}$. $d\Sigma^i$ is normal to this section and may be expressed in the adapted coordi-

nates as $s\Sigma^i$:

$$d\Sigma^i|_\Sigma = s\Sigma^i|_\Sigma.$$

At this stage, our analysis will differ from the original work of Faddeev and Popov. The volume of the fiber will not be factorized out of the functional integral but the Dirac function and 1 form $s\Sigma^i$ will be lifted into the Lagrangian by use of two Lagrange multipliers: a Bose field σ_i and a Grassman multiplier η_i anticommuting with s:

$$\prod_i \delta(\Sigma^i) = \int \prod_i d\sigma_i e^{-i\sigma_i \Sigma^i},$$

$$s\Sigma^i = i\frac{\partial}{\partial \eta_i} e^{-i\eta_i s\Sigma^i}\Big|_{\eta=0} \triangleq \int \bar{d}\eta_i\, e^{-i\eta_i s\Sigma^i}.$$

This last expression is well defined despite the nonintegrated differential form appearing in the exponent as it may always be linearized by performing the Berezin integration over $\eta_i \cdot \eta_i$ behaves as a vertical vector but is not a vector because $\eta_i s\Sigma^i$ must not be considered as a scalar. According to Bernshtein and Leites,[15] it is an integral form. The effective functional integral may now be written as

$$W = \int \prod_{x\in\Sigma} \mathscr{D}(\varphi^i{}_\mu, \chi^i, \sigma^i, \eta^i)\delta(s\chi^i + \tfrac{1}{2}[\chi,\chi]^i)$$
$$\times \delta(s\varphi^i{}_\mu - B_\mu \chi^i) e^{-i\int \mathscr{L} + \sigma_i \Sigma^i + \eta_i s\Sigma^i}.$$

The last term in the Lagrangian may be transformed using the BRS structural condition

$$s\Sigma^i = \frac{\delta \Sigma^i}{\delta \varphi^j} s\varphi^j = -\frac{\delta \Sigma^i}{\delta \varphi^j} B\chi^j.$$

The Faddeev–Popov effective Lagrangian has thus been exactly recovered. Because η and χ anticommute, a minus sign must, in the perturbation expansion, be associated to each closed ghost antighost loop. In the usual approach[2], an additional integration over the ghost variables together with an integration over the volume of the group are generated. However, these 2 corrections cancel one another because the volume element of the fiber is simply the exterior products of the ghost forms,

$$\int \prod_i [dy] \bar{d}c^i = \int \prod_i c^i \bar{d}c^i = 1.$$

The exterior differential of the Lagrange multipliers naturally vanishes; thus,

$$s\chi = -\tfrac{1}{2}[\chi,\chi], \quad s\varphi = -B\chi,$$
$$s\sigma_i = s\eta_i = 0, \quad s\mathscr{L}_{\text{eff}} = \sigma_i s\Sigma^i.$$

In this approach the effective Lagrangian is not BRS invariant because we have not recovered the nongeometric variation of the antighost. The study of the renormalizability of the theory is however not affected. Indeed the BRS–Ward identity must usually be completed by the equation of motion of the antighost precisely in order to compensate for its nongeometrical contribution to the former.

Moreover, a detailed analysis[16] shows that the renormalized BRS operator follows the above prescriptions and does not transform the antighost.

6. CONCLUSION

Differential geometry has provided us with a better understanding of the nature of the quantum gauge theories. In Sec. 3, the Faddeev–Popov ghost has been reinterpreted as the component of the connection 1 form normal to the section in the principal fiber bundle, and the BRS operator as the corresponding part of the exterior differential. The BRS transformation rules of the ghost and gauge field then follow from the Cartan–Maurer structural theorem which states the existence of a fibration. Under ordinary gauge transformations, the ghost is shown in Sec. 4 to be invariant, but a more general type of transformation is defined which is related to solitons. In Sec. 5, the effective Lagrangian with ghosts and gauge fixing term is obtained without factorizing an infinite constant out of the generating functional. In this picture of the Lagrangian is not BRS invariant, but this does not spoil the discussion of renormalizability in which one usually uses the equation of motion of the antighost to cancel its nongeometric variation. This presentation should find applications in the study of solitons and the group manifold approach to quantum gravity.

ACKNOWLEDGMENTS

It is a pleasure to thank B. Carter, S. Nussinov, J.B. Zuber, and T. Ungar for many discussions and Y. Ne'eman for his suggestions and his hospitality in Tel Aviv where this work was completed.

[1]G.'t Hooft, and M., Veltman Diagrammar, CERN 73.9, Genéve, 1973.
[2]L. D. Faddeev and V. N., Popov Phys. Lett. B **25**,29 (1967).
[3]R. P. Feynman, Acta Phys. Pol. **26**, 697 (1963).
[4]C. Beechi, A. Rouet, and R. Stora, Int. School of Math. Phys., Erice, August, 1975.
[5]B. Lee, Int. School of Theor. Phys., Les Houches, August 1975.
[6]J. Zinn-Justin, Int. School for Theor. Phys., Bonn, 1976.
[7]R. Stora, Summer Inst. for Theor. Phys., Cargese, 1976.
[8]D. A. Popov, Teor. Mat. Fiz. **24**, 3, 347 (1975).
[9]Y. Ne'eman and T. Regge, "Gauge theory on the group manifold," Nuovo Cimento **5** (1978).
[10]M. Spivak, *Introduction to Differential Geometry* (Publish or Perish, Boston, Mass. 1970).
[11]Y. Choquet-Bruhat, C. Dewitt-Morette and M. Dillard-Bleick, *Analysis Manifolds and Physics* (North Holland, Amsterdam).
[12]C. N. Yang A.N.Y.A.A. **9**, 294 (1977).
[13]Y. Ne'eman, J. Thierry-Mieg, and Marcel Grossman edited by Conf. Trieste, 1979 (to be published).
[14]H. Pagels, Lectures at Coral Gables, 1978, edited by I. M. Singer.
[15]I. N. Bernshtein and D. A. Leites, Functional Anal. i Prilozen. **11**, 1 (1977).
[16]See Ref. 6 Eqs. (78b) and (79) and J. Thierry-Mieg, these de doctorat, p. 97–104, Orsay, 1978.

Gauge symmetry and its breakdown: The example of a BCS superconductor

S. K. Bose
Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556

(Recieved 3 June 1980; accepted for publication 20 August 1980)

The mathematical structure of an infinitely extended BCS super-conductor is re-examined in the light of the theory of bundle representations. The role of the homotopy group in the BCS model is clarified. The precise characterization of the constant gauge transformation in terms of a principal fiber bundle (with discrete fiber and group) is pointed out.

In the limit of infinite volume, the BCS theory of superconductivity[1] provides an exactly soluble[2] model wherein the phenomenon of spontaneous symmetry breakdown occurs explicitly; the symmetry that gets broken being the gauge invariance.[3] The concomitant degeneracy of the ground state is such that it can be labeled either (1) by a continuous parameter α, $0<\alpha<2\pi$ or (2) by a discrete integer-valued parameter n, $n=0, \pm 1, \pm 2\cdots$. The particle number (physically, the number of Cooper pairs) is unsharp in states corresponding to the first way of labeling the ground state and is sharp in states that correspond to the second way. The representation of the algebra generated by the (smeared) fields is irreducible in the unsharp states and is reducible in the sharp states. These facts were established in a classic analysis of the BCS model performed by Haag[4] in 1962.

Recent developments[5-7] in the theory of bundle representations have provided us with an elegant technique through which to describe the phenomenon of spontaneous breakdown of a continuous symmetry. It is well known that when spontaneous breakdown occurs, the symmetry operation *cannot* be implemented via (continuous) unitary operators in a Hilbert space.[8] In the Araki–Haag framework it means that the symmetry is locally, but not globally, unitarily implementable. The method of bundle representations goes a step further and gives us a precise prescription for globally implementing the broken symmetry operations[9]: the latter are implemented as *bundle maps* on a suitably constructed Hilbert bundle (a fiber bundle whose fiber is a Hilbert space) based on an appropriately chosen homogeneous space. This method was developed by Borchers and Sen[5] and by Sen[6] originally for the purpose of describing relativity groups in an infinite medium, where the boost operations "get broken". Later the method was applied to the breaking of internal symmetries.[7] In view of the foregoing developments it seems worthwhile to re-examine the BCS model in the framework of bundle representations. In the process, additional insight is gained on known results whose intuitive content becomes easily visualizable, and the *general* features of gauge symmetry breaking begin to emerge. Specifically, we prove the following results:

(1) The space of the states of unsharp particle number is a Hilbert bundle based on the circle S^1. Gauge transformations are implemented on the bundle space via bundle maps.

(2) The Hilbert space of states with sharp particle number (which may also be viewed as Hilbert bundle based on the discrete space of integers) provides a unitary representation of the homotopy group $\pi_1(S^1)$—the fundamental group of the circle. The "topological quantum number" associated with $\pi_1(S^1)$ provides a superselection rule, whose existence accounts for the reducibility of the representation of the algebra generated by the smeared fields.

(3) The Hilbert space of states with sharp particle number is related to the bundle space of states with unsharp number via the standard mathematical procedure of forming direct integrals, in the sense of von Neumann.[10]

(4) The origin of the homotopy group is traced to the mathematical structure of (constant) gauge transformations. Precisely, this structure is that of a *principal fiber bundle* based on the circle S^1 and with a structure group $\pi_1(S^1)$.

At the risk of repetition, we wish to recall some basic notations. The field $\psi_r(x)(\psi_r^*(x))$ destroys (creates) an electron of spin r ($r=1,2$) at the spatial point x. The anticommutation relations

$$\{\psi_r(x),\psi_s(x')\}=0, \tag{1}$$

$$\{\psi_r(x),\psi_s^*(x')\}=\delta_{rs}\delta(x-x'), \tag{1a}$$

together with the form

$$H_i(V)=\frac{1}{V}\int \psi_1^*(x)\psi_2^*(x+z)\psi_2(x'+z') \times\psi_1(x')v(z,z')\,dx\,dx'\,dz\,dz' \tag{2}$$

for the interaction Hamiltonian defines the model. The function $v(z,z')$ characterizes the interaction. All quantities are defined at one instant of time, taken as $t=0$. The limit $V\to\infty$ of infinite volume is taken at the very beginning. Let $\psi(f)$, $\psi^*(f)$ denote the weighted average of $\psi(x)$, $\psi^*(x)$ with respect to square integrable functions $f(x)$ of position. Let S denote the algebra generated by $\psi(f)$, $\psi^*(f)$ [for all such $f(x)$] and let R be the Von Neumann algbera generated by S. So much for notation.

The presence of a superconducting phase is heralded by the two-point correlation function

$$\phi_\alpha(z)=\langle\alpha|\psi_2(x)\psi_1(0)|\alpha\rangle = \exp[i\alpha]\phi_0(z), \tag{3}$$

where $\phi_0(z)$ is a real function not identically zero, and $|\alpha\rangle$ the ground state. The structure of the corresponding state space, according to Haag,[4] is this: for each value of α, there is a separable Hilbert space H_α; all H_α's are exact copies of each

 0022-2488/80/122839-03$1.00

other and thus of some H; linear combinations and inner products are *not* defined between states that sit over distinct values of α. It is evident that these properties define a Hilbert bundle based on the circle ($0 \leqslant \alpha < 2\pi$) and with fiber H (H_α is the fiber over α). In fact, it is a product bundle (equivalent, *in the group of the bundle*, to a product bundle). Let B denote the bundle space. An element $b \in B$ can be written as $b = (\alpha, \phi)$, where $\phi \in H$. Under a gauge transformation, the field behave as

$$\psi \to \psi \exp[i\beta]; \quad \psi^* \to \psi^* \exp[-i\beta] \tag{4}$$

and Eq. (1) is unchanged. Thus (4) is an automorphism of algebra R. The pair (β, a), $a \in R$, is an example of the general object (G, A)—A is some algebra and G a group of automorphism of A—whose bundle representations have been constructed elsewhere,[11] using a cocycle technique. For the present case, the general bundle representation formula of Ref. 11 reads

$$(\beta, a)(\alpha, \phi) = (\alpha + 2\beta, [D\tau_{-\beta}(a)]\phi), \tag{5}$$

where $\tau_\beta(a)$ is the gauge transform of a and D is a symmetric representation of R (the concrete c^*-algebra). Equation (5) shows that gauge transformation is implemented on B as a *bundle map*. The bundle map acts as a left translation on the base space and acts on the fiber through the cocycle. It now follows rapidly that the (above) bundle maps falls into homotopy classes; the associated group being the fundamental group $\pi_1(S^1)$ of the circle.[12] The significance of the homotopy group becomes more transparent on the states of sharp particle number.

The states of sharp particle number are those for which the expectation value of every gauge-variant quantity vanishes and the transformations (4) reduces to the identity map (on every physical observable). Thus the gauge transformation (4) is implemented unitarily (although trivially!) on these states. Let K denote the space of these states. We thus expect the passage from the bundle space B to the space K to mimic the standard procedure of forming direct integrals, which is used in the theory of induced representations[13] of locally compact groups since the inducing construction does just that i.e., provides a means of constructing unitary representations from bundle representations. Of course, here we are not representing a group but a more general object (G, A), but that does not matter. Let $b \in B$, $b = (\alpha, \phi)$ as before. Then $(\alpha, \phi) \to \phi_\alpha$ defines a cross-section of the Hilbert bundle. Let $(\phi_\alpha, \phi_\alpha) = (\phi, \phi)_\alpha$ denote the norm in H_α. Define now a new norm $[\ ]$ by the rule

$$[\phi, \phi] = \int_0^{2\pi} (\phi, \phi)_\alpha \frac{d\alpha}{2\pi}, \tag{6}$$

where $d\alpha/2\pi$ is the Haar measure on the circle. It is now a well-known fact[12] that with respect to the above $[\ ]$, the ϕ's constitute a linear space equipped with the polarization identity, and hence with an inner product, and is complete in the norm; thus it is a Hilbert space. Let $\bar{K}$ denote this space. Let $|\Omega\rangle$ be a cyclic state (ground state) of $\bar{K}$. From (6) we compute

$$\langle \Omega | \Omega \rangle = \int_0^{2\pi} \langle \alpha | \alpha \rangle \frac{d\alpha}{2\pi} = 1. \tag{7}$$

Thus $|\Omega\rangle$ is a normalized state. An inner product of two vectors in $\bar{K}$ is related to that in H_α again by an integral of the form (6). This implies that

$$\langle \Omega | \psi_2(z)\psi_1(0) | \Omega \rangle = \int_0^{2\pi} \langle \alpha | \psi_2(z)\psi_1(0) | \alpha \rangle \frac{d\alpha}{2\pi} = \phi_0(z) \int_0^{2\pi} \exp[i\alpha] = 0. \tag{8}$$

Note that Eq. (3) has been used at the second step of the above derivation. It is now a simple matter to prove that

$$\langle \Omega | \psi^*(x_1) \cdots \psi^*(x_n) \psi(y_1) \cdots \psi(y_m) | \Omega \rangle = 0, \tag{9}$$

whenever $m \neq n$, since any such expectation value as above can be decomposed[4] into sums of products of the basic two point functions and $\langle \Omega | \psi_2(z)\psi_1(0) | \Omega \rangle$ will appear as a factor in every term in the summation. As for quantities for which $m = n$, Eq. (6) gives

$$\langle \Omega | \psi^*(x_1) \cdots \psi^*(x_n) \psi(y_1) \cdots \psi(y_n) | \Omega \rangle = \langle \alpha | \psi^*(x_1) \cdots \psi^*(x_n) \psi(y_1) \cdots \psi(y_n) | \alpha \rangle \tag{10}$$

since the quantity is independent of α (gauge invariant). Equations (9) and (10) are the *defining relations*[4] for the space K. Thus we have proved that $\bar{K} = K$ and hence K is obtained from B via the direct integral (6).

The homotopy group appears in a different avatar in the space K. Let U be the generator of $\pi_1(S^1)$. Then the following statements are entirely obvious: I) U commutes with S and thus with every element of R, II) U does not annihilate $|\Omega\rangle$, III) the group $\pi_1(S^1)$ is unitarily implemented on K. We thus have the string of ground states

$$|\Omega_{2n}\rangle = U^n |\Omega\rangle, \quad n = 0, \pm 1, \pm 2 \cdots. \tag{11}$$

obtained by applying the elements of the homotopy group to $|\Omega\rangle$. All these are degenerate since U commutes with the Hamiltonian (2). Moreover, $|\Omega_{2n}\rangle$ must be orthogonal to $|\Omega_{2m}\rangle$ if $m \neq n$, since this is the statement of the superselection rule associated with the conservation of the "topological quantum number" arising from the homotopy group. The space K contains the string of Hilbert spaces H_{2n}, with H_{2n} arising from $|\Omega_{2n}\rangle$. Thus K consists of the collection of all *coherent sectors* associated with our superselection rule. In other words, the representation of the algebra R on K *must be reducible*. To conclude the task of showing the connection with Haag's treatment, we finally write down the explicit form of U, which is

$$U = \phi_0^{-1}(z) \lim_{V \to \infty} \frac{1}{V} \int \psi_1^*(x)\psi_2^*(x+z)\, dx. \tag{12}$$

The physical meaning of U and thus of the topological quantum number n is now quite clear. One remark: we are free to look upon the Hilbert space K as a Hilbert bundle based on the discrete space Z of integers (Z is equipped with the discrete topology). We note, in the passing, that the degeneracy structure of the ground states in K is very similar to that of the vacua in Yang–Mills theories.[14] In fact, they are *mathematically* identical [the groups $\pi_1(S^1)$ and $\pi_3(S^3)$ are isomorphic]. This fact seems to have gone unappreciated in the literature.[14]

We ask: what general features of gauge transformation

emerge from the foregoing analysis of the BCS model? The answer is contained in the following two remarks.

A) Gauge transformation is a very special kind of a "symmetry", in that, it is a *nontrivial* automorphism only of the algebra of field operators [e.g., of Eq. (1)] but *not* of the algebra of observables (identity map for the latter). This means that we have at our disposal two distinct but equivalent ways of describing the situation. If our states are constructed via the expectation values of physical observables, then the symmetry is there but is trivial. We can begin to talk about the symmetry nontrivially, only when our states are so constructed that they correspond to nonvanishing expectation values of not only physical observables but also of unphysical, gauge-variant quantities. In the latter event, gauge symmetry, of necessity, is broken. Thus the breakdown of gauge symmetry is a concept which is dependent on the choice of language. However, all is not lost, since the memory of symmetry breakdown persists in the form of the homotopy group. This should be a general feature of all theories, including nonabelian (constant) gauge theories.

B) The gauge transformation (4) is a *covering map* p: $E^1 \to S^1$ from the real line (E^1) to the circle, given explicitly as

$$C = \exp[i\alpha], \tag{13}$$

where α ranges over the reals and C is on the unit circle. Now it is a standard mathematical result that a covering map admits a bundle structure with a discrete fiber.[15] In fact, the bundle structure corresponding to (13) is a principal fiber bundle whose base space is S^1 and with fiber and the group $\pi_1(S^1)$; E^1 is the bundle space. Thus the bundle is $[E^1, p, S^1, \pi_1(S^1), \pi_1(S^1)]$. If we forget about the bundle structure and look only at the base space [as a $U(1)$ group], we are bound to lose information. In problems where the homotopy group plays a role, this loss of information is not desirable.

After explaining the principal fiber bundle characterization of gauge transformations of the second kind. Wu and Yang[16] remarked, "all gauge fields are thus based on geometry". Our analysis shows that the same is true for a gauge transformation of the first kind.

ACKNOWLEDGMENT

This work began during a visit to the department of mathematics, Ben Gurion University of the Negev, Beersheva, Israel. The author takes pleasure in thanking his colleagues at Beersheva for warm hospitality and stimulating company, in particular, Professor R. N. Sen also for many helpful discussions.

[1] J. Bardeen, L. N. Cooper and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).
[2] N. N. Bogoliubov, Physica, **26**, S1 (1960).
[3] Throughout this paper, by gauge transformation we will mean a transformation of the first kind (constant gauge transformations).
[4] R. Haag, Nuovo Cimento **25**, 287 (1962).
[5] H. J. Borchers and R. N. Sen, Commun. Math. Phys. **42**, 101 (1975).
[6] R. N. Sen, Physica **94** A, 39 (1978); **94** A, 55 (1978).
[7] S. K. Bose, Lett. Nuovo Cimento **28**, 146 (1980).
[8] See, for instance, J. A. Swieca, *Cargése Lectures in Physics* (Gordon and Beach, New York, 1970); H. Reeh, Fortschr. Phys. **16**, 687 (1968).
[9] By broken symmetry, we will hereafter mean a spontaneously broken symmetry and never one that is broken by explicitly unsymmetric forces.
[10] J. Von Neumann, Ann. Math. **50**, 401 (1949).
[11] R. N. Sen, talk at the conference on Differential Geometrical Methods in Mathematical Physics, Salamanca, Spain, Sept., 1979. S. K. Bose and R. N. Sen (to be published).
[12] That is, transformations $\exp[i\alpha n]$, $n \in Z$, fall into homotopy classes characterized by n. Here Z is the additive group of integers.
[13] G. Mackey, *Unitary Group Representation in Physics, Probability and Number Theory* (Benjamin, Reading, Mass., 1978).
[14] See, for instance, S. Coleman, "The Uses of Instantons", lectures at 1977 International School of Subnuclear Physics, *Ettore Majorana*.
[15] N. Steenrod, *The Topology of Fibre Bundles* (Princeton U. P., Princeton, N. J., 1951), pp. 67–71.
[16] T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).

On the proper vectors of real third order matrices

Vijay K. Stokes
General Electric Company, Corporate Research and Development, Schenectady, New York 12301

(Received 15 February 1980; accepted for publication 15 July 1980)

The result that a pseudovector can be associated with a real third order skew-symmetric matrix has been used for establishing some properties of the proper vectors of real third order matrices. It turns out that the pseudovector associated with the skew-symmetric part of such matrices characterizes some interesting properties of proper vectors, such as the question of their orthogonality.

INTRODUCTION

The purpose of this paper is to establish some properties of proper vectors of real square matrices of order three, for the case when all the three proper values are real. This work is motivated by the problem of the analysis of stress, wherein the real proper values and proper vectors of the transpose of the stress matrix are, respectively, the principal stresses and the principal directions of stress.[1] In the general polar case in which one or more of internal spin, couple stresses or body moments exist, the stress matrix is not symmetric.

It is well known that all the proper values of a real symmetric matrix are real and that there exists at least one set of mutually orthogonal proper vectors. When the proper values are distinct, the corresponding proper vectors are mutually orthogonal. However, in general, the proper values and proper vectors of a real matrix are complex. In this paper only real proper values and real proper vectors are considered.

Let $\mathbf{M}$ be a real square matrix of order three. The symmetric part $\mathbf{M}^S$, the skew-symmetric part $\mathbf{M}^A$, and the deviatoric part $\mathbf{D}$ of $\mathbf{M}$, are then defined by $\mathbf{M}^S = \frac{1}{2}(\mathbf{M} + \mathbf{M}^T)$, $\mathbf{M}^A = \frac{1}{2}(\mathbf{M} - \mathbf{M}^T)$ and $\mathbf{D} = \mathbf{M} - \frac{1}{3}(\mathrm{tr}\mathbf{M})\mathbf{I}$. The pseudo vector, $\mathbf{m}_A$, associated with the skew-symmetric matrix, $\mathbf{M}^A$, is defined by

$$\mathbf{M}^A = \begin{bmatrix} 0 & m_3 & -m_2 \\ -m_3 & 0 & m_1 \\ m_2 & -m_1 & 0 \end{bmatrix}, \quad \mathbf{m}_A = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}$$

and, for all vectors, $\mathbf{n}$ has the property that

$$\mathbf{M}^A\mathbf{n} = \mathbf{n}\times\mathbf{m}_A, \tag{1}$$

where $\mathbf{n}\times\mathbf{m}_A$ is the vector product of $\mathbf{n}$ and $\mathbf{m}_A$. It is this result, which only holds for real third order matrices, that makes it possible to establish the properties of proper vectors discussed in this paper.

$\mathbf{M}$ and $\mathbf{D}$ have the same proper vectors. If λ and ν are, respectively, the proper values of $\mathbf{M}$ and $\mathbf{D}$, corresponding to the same proper vector, then $\lambda = \nu + \frac{1}{3}\mathrm{tr}\mathbf{M}$. Thus λ is real if and only if ν is real. The condition that $\mathbf{D}$ have three real proper values is that all the three roots of the characteristic equation $\det(\mathbf{D} - \nu\mathbf{I}) = 0$ be real. The conditions for the proper values to be real are then characterized by the following well-known result.

Proposition 1: For a real third order square matrix, $\mathbf{M}$, only one proper value is real when $\phi = \mathrm{tr}^3\mathbf{D}^2 - 54\det^2\mathbf{D} < 0$. When $\phi > 0$, all the three proper values are real and distinct. When $\phi = 0$, all the proper values are real and are given by 2ν, $-\nu$, $-\nu$, where $\nu = (\frac{1}{2}\det\mathbf{D})^{1/3}$.

BASIC RESULTS

It is assumed that all the three proper numbers of $\mathbf{M}$ are real and that $\mathbf{m}_A \neq 0$. Let $\mathbf{n}_1$ and $\mathbf{n}_2$ be two distinct proper vectors corresponding, respectively, to the proper numbers λ_1 and λ_2, so that $\mathbf{M}\mathbf{n}_1 = \lambda_1\mathbf{n}_1$ and $\mathbf{M}\mathbf{n}_2 = \lambda_2\mathbf{n}_2$. It then follows from these two equations and Eq. (1) that

$$(\lambda_2 - \lambda_1)\mathbf{n}_1^T\mathbf{n}_2 = 2\mathbf{n}_1^T\mathbf{M}^A\mathbf{n}_2$$

$$= 2\mathbf{n}_1^T(\mathbf{n}_2\times\mathbf{m}_A) = 2[\mathbf{n}_1,\mathbf{n}_2,\mathbf{m}_A].$$

Now the scalar triple product $[\mathbf{n}_1,\mathbf{n}_2,\mathbf{m}_A]$ vanishes if and only if $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{m}_A$ are coplanar. Hence,

Proposition 2: If $\mathbf{n}_1$ and $\mathbf{n}_2$ are two distinct proper vectors of $\mathbf{M}$ corresponding, respectively, to the proper values λ_1 and λ_2, then $(\lambda_2 - \lambda_1)\mathbf{n}_1^T\mathbf{n}_2 = 0$ if and only if $\mathbf{m}_A$ is a linear combination of $\mathbf{n}_1$ and $\mathbf{n}_2$.

Also, if $\lambda_1 = \lambda_2 = \lambda_0$ and $\mathbf{n}_1$ and $\mathbf{n}_2$ are two different proper vectors corresponding to the repeated proper value λ_0, then it follows that $[\mathbf{n}_1,\mathbf{n}_2,\mathbf{m}_A] = 0$. Hence,

Proposition 3: If $\mathbf{M}$ has two distinct proper vectors corresponding to the same (repeated) proper value, so that every vector in their plane is a proper vector, then $\mathbf{m}_A$ is also a proper vector and lies in this plane.

DISTINCT PROPER VALUES

When $\phi > 0$, the three proper values of $\mathbf{M}$ are real and distinct.

Corollary 4: Two proper vectors of $\mathbf{M}$ corresponding to distinct proper values are orthogonal if and only if the pseudovector, $\mathbf{m}_A$, lies in the plane of the proper vectors.

Corollary 5: If all the three proper values are distinct then, at most, one of the proper vectors is orthogonal to the other two. A proper vector $\mathbf{n}_1$ is orthogonal to the other two proper vectors, $\mathbf{n}_2$ and $\mathbf{n}_3$, if and only if $\mathbf{n}_1$ is parallel to $\mathbf{m}_A$, that is, if and only if $\mathbf{m}_A$ is that proper vector.

 0022-2488/80/122842-02$1.00

Remarks: If follows that if all the proper values are distinct, then the three proper vectors are mutually orthogonal if and only if $\mathbf{m}_A = 0$, that is, when $\mathbf{M}$ is symmetric.

The results of this section are illustrated by the matrix

$$\mathbf{M} = \begin{bmatrix} \lambda_1 & a_3 & a_2 \\ 0 & \lambda_2 & a_1 \\ 0 & 0 & \lambda_3 \end{bmatrix} \tag{2}$$

which has the three distinct proper values, λ_1, λ_2, and λ_3, The corresponding proper vectors are, respectively.

$$\mathbf{n}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{n}_2 = \begin{bmatrix} -a_3 \\ \lambda_1 - \lambda_2 \\ 0 \end{bmatrix},$$

and

$$\mathbf{n}_3 = \begin{bmatrix} a_1 a_3 - a_2(\lambda_2 - \lambda_3) \\ -a_1(\lambda_1 - \lambda_3) \\ (\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3) \end{bmatrix}.$$

None of these proper vectors are orthogonal when a_1, a_2 and a_3 are nonzero, and $\mathbf{m}_A$ does not lie in any one of the planes determined by $(\mathbf{n}_1,\mathbf{n}_2)$, $(\mathbf{n}_2,\mathbf{n}_3)$ and $(\mathbf{n}_3,\mathbf{n}_1)$. If $a_3 = 0$, then $\mathbf{n}_1$ is orthogonal to $\mathbf{n}_2$ but neither of them is orthogonal to $\mathbf{n}_3$. In this case, $\mathbf{m}_A$ lies in the plane of $\mathbf{n}_1$ and $\mathbf{n}_2$ but is not along either one of them. Finally, if $a_2 = 0$ and $a_3 = 0$, then $\mathbf{n}_1$ is orthogonal to both $\mathbf{n}_2$ and $\mathbf{n}_3$, but $\mathbf{n}_2$ is not orthogonal to $\mathbf{n}_3$, and $\mathbf{m}_A$ is along $\mathbf{n}_1$.

TWO PROPER VALUES EQUAL

When $\phi = 0$ but $\det\mathbf{D} \neq 0$, two of the proper values are equal and different from the third one. Let $\lambda_1 \neq \lambda_2 = \lambda_3 = \lambda_0$. Then there are two possibilities: (i) either there is only one proper vector, $\mathbf{n}_0$, corresponding to the repeated proper value λ_0, or (ii) all the proper vectors in a plane are proper vectors corresponding to λ_0. It then follows from Propositions 2 and 3 that the following results hold.

Corollary 6: If corresponding to a repeated root $\lambda_0 = \lambda_2 = \lambda_3 \neq \lambda_1$ there exists only one proper vector, $\mathbf{n}_0$ of $\mathbf{M}$, then $\mathbf{M}$ has only two proper vectors, $\mathbf{n}_0$ and $\mathbf{n}_1$, where $\mathbf{n}_1$ is the proper vector associated with the proper value λ_1. Further, $\mathbf{n}_0$ and $\mathbf{n}_1$ are orthogonal if and only if $\mathbf{m}_A$ lies in their plane.

Corollary 7: If $\mathbf{M}$ has two distinct proper vectors, $\mathbf{n}_2$ and $\mathbf{n}_3$, which correspond to the same repeated proper values, $\lambda_0 = \lambda_2 = \lambda_3 \neq \lambda_1$, so that every vector in their plane is also a proper vector corresponding to λ_0, then $\mathbf{m}_A$ is also a proper vector which lies in this plane. Further, $\mathbf{m}_A$ is orthogonal to $\mathbf{n}_1$, the proper vector associated with λ_1.

Remarks: As an example, consider the matrix in Eq. (2) with $\lambda_2 = \lambda_3 = \lambda_0 \neq \lambda_1$. Then, as long as $a_1 \neq 0$, it only has the two proper vectors $\mathbf{n}_1 = (1\ 0\ 0)$ and $\mathbf{n}_2 = (-a_3\lambda_1 - \lambda_0\ 0)$ which correspond, respectively, to the proper values λ_1 and λ_0. These two vectors are orthogonal only when $a_3 = 0$, in which case, $\mathbf{m}_A$ lies in the plane of $\mathbf{n}_1$ and $\mathbf{n}_2$. However, when $a_1 = 0$, $\mathbf{n}_1 = (1\ 0\ 0)$ is the proper vector corresponding to λ_1 and *every* vector which is orthogonal to $\mathbf{n} = (\lambda_1 - \lambda_0\ a_3\ a_2)$ is a proper vector corresponding to the repeated proper value λ_0. Furthermore, $\mathbf{m}_A$ is orthogonal to $\mathbf{n}_1$.

ALL THREE PROPER VALUES EQUAL

When $\phi = 0$ and $\det\mathbf{D} = 0$, all the three proper values are equal, and the only possibilities are then given in Corollary 8.

Corollary 8: If all the three proper values of $\mathbf{M}$ are equal, then either (i) there is only one proper vector, or (ii) all the vectors in a plane are proper vectors and $\mathbf{m}_A$ is also a proper vector which lies in this plane.

This result is illustrated by the matrix in Eq. (2) with $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_0$. If $a_1 \neq 0$ and $a_3 \neq 0$ then the only proper vector corresponding to the repeated proper value λ_0 is $\mathbf{n}_1$ $(1\ 0\ 0)$. However, if $a_1 = 0$, then all the vectors in the plane normal to $\mathbf{n} = (0\ a_3\ a_2)$ are proper vectors, and $\mathbf{m}_A$ also lies in this plane.

CONCLUDING REMARKS

When the three proper values are real and distinct, $\mathbf{M}$ has only three proper vectors which, in general, are not orthogonal. Two of the proper vectors are orthogonal when $\mathbf{m}_A$ lies in their plane. If $\mathbf{m}_A$ is itself a proper vector, then it is orthogonal to the other two proper vectors, which are not orthogonal.

When two of the proper values are equal and different from the third proper value, then there are two cases. (i) If there is only one proper vector associated with the repeated proper value, then $\mathbf{M}$ has only two proper vectors, the second one corresponding to the distinct proper value. These two proper vectors are orthogonal only when $\mathbf{m}_A$ lies in their plane. (ii) If $\mathbf{M}$ has more than one proper vector associated with the repeated proper value, then all the vectors in a plane are proper vectors and this plane also contains the vector $\mathbf{m}_A$. In addition to the proper vectors in this plane, there is an additional proper vector, corresponding to the distinct proper value, which does not lie in this plane and which is orthogonal to $\mathbf{m}_A$.

Finally, when all the three proper values are equal, then there are two cases: (i) either $\mathbf{M}$ has only one proper vector or (ii) all the vectors in a plane are proper vectors and $\mathbf{m}_A$ also lies in this plane.

Given a real third order matrix $\mathbf{M}$, only one real proper vector exists when only one of the proper values is real, the other two proper vectors being complex. However, only one proper vector can exist even when all the proper values are real and equal. $\mathbf{M}$ can have just two proper vectors only when two of its proper values are equal and different from the third one. It has only three proper vectors when the three proper values are distinct. All the proper vectors may comprise the vectors in a plane, this being possible only when all the proper values of $\mathbf{M}$ are equal. In the last possibility, all the vectors in a plane are proper vectors and a vector, not in this plane, is also a proper vector. This is possible only when two proper values are equal and different from the third.

[1]V. K. Stokes, "On the Analysis of Asymmetric Stress," J. Appl. Mech. **39**, 1133–6 (1972).

Bloch electrons in a magnetic field—reduction to one dimension [a)]

G. H. Wannier
University of Regensburg, Faculty of Physics, D-8400 Regensburg, West Germany [b)]

(Received 10 June 1980; accepted for publication 18 August 1980)

A reduction to one dimension of the above problem, found by Schellnhuber and Obermair for a special model lattice, is shown to be valid for all lattices without restriction. As was the case in their problem, the field must be rational. If the rational number is the reciprocal of an integer a single equation results. This condition is well adapted to the study of fields of practically attainable magnitude. If the rational number is of the form q/p a system of q coupled equations is obtained.

1. INTRODUCTION

Schellnhuber,[1] and Schellnhuber and Obermair[2] have recently adapted to low fields the method introduced by Rauh, Wannier, and Obermair (RWO)[3] to solve the quantum problem of a crystalline electron in a magnetic field. They did this for a set of particular cases where the periodic potential has the simplest possible nontrivial form and the rational number attached to the field is the reciprocal of an integer p. The essentially new step was the reduction to a Schrödinger-like equation in one variable only. It is the purpose of this paper to show that the restriction to a simple potential is not necessary: *all* periodic potentials down to triclinic symmetry allow this type of reduction for *all* rational fields. The reduction arises fairly directly from the structure of the Landau functions in Cartesian coordinates.

2. SPECIAL CASE: THE RECIPROCAL OF THE RATIONAL NUMBER IS AN INTEGER

It was shown earlier[3] that rationality requires the magnetic field to be parallel to a lattice vector $\mathbf{c}$ which we take along the z direction of a Cartesian system of axes. We are then compelled to lay the x axis parallel to one of the reciprocal lattice vectors $\mathbf{a}^*$ which are perpendicular to $\mathbf{c}$; this is needed to get a minimal representation of the magnetic translation group.[3] We therefore write the three basis vectors of the crystal lattice in the form

$$\mathbf{a} = \mathbf{i}a_x + \mathbf{j}a_y + \mathbf{k}a_z, \tag{1a}$$

$$\mathbf{b} = \mathbf{j}b_y + \mathbf{k}b_z, \tag{1b}$$

$$\mathbf{c} = \mathbf{k}c\,. \tag{1c}$$

Equation (1) imposes no restriction on the symmetry or lack of symmetry of the crystal. a_x, b_y, and c must be different from zero; their product is the volume of the unit cell. The only bounding parallelogram of the cell traversed by magnetic flux is the one generated by $\mathbf{a}$ and $\mathbf{b}$. The reciprocal vectors of (1) are

$$\mathbf{a}^* = \mathbf{i}\frac{1}{a_x}, \tag{2a}$$

$$\mathbf{b}^* = -\mathbf{i}\frac{a_y}{a_x b_y} + \mathbf{j}\frac{1}{b_y}, \tag{2b}$$

$$\mathbf{c}^* = \mathbf{i}\frac{a_y b_z - a_z b_y}{a_x b_y c} - \mathbf{j}\frac{b_z}{b_y c} + \mathbf{k}\frac{1}{c}. \tag{2c}$$

The potential is triply periodic

$$V(x + a_x, y + a_y, z + a_z) = V(x,y,z), \tag{3a}$$

$$V(x, \quad y + b_y, z + b_z) = V(x,y,z), \tag{3b}$$

$$V(x, \quad y, \quad z + c) \; = V(x,y,z). \tag{3c}$$

It has a Fourier expansion involving the vectors (2):

$$V(x,y,z) = \sum_{l,n,m} v_{l,n,m} \exp\Bigg[2\pi\mathrm{i}\Bigg\{\Bigg(l\frac{1}{a_x} - n\frac{a_y}{a_x b_y} + m\frac{a_y b_z - b_y a_z}{a_x b_y c}\Bigg)x + \Bigg(n\frac{1}{b_y} - m\frac{b_z}{b_y c}\Bigg)y + m\frac{1}{c}z\Bigg\}\Bigg]. \tag{4}$$

It will come out as usual that the potential has just a band splitting function along the z direction where the large energy term $\hbar^2 k_z^2/2m$ is controlling. We need, therefore, the potential (4) only for $m = 0$ and write it in the form

$$[V(x,y,z)]_{m=0} = \frac{\hbar^2}{2m} V(x,y), \tag{5a}$$

$$V(x,y) = \sum_n w_n(x) \times \exp\Bigg[2\pi i\Bigg(-n\frac{a_y}{a_x b_y}x + n\frac{1}{b_y}y\Bigg)\Bigg], \tag{5b}$$

$$w_n(x) = \frac{2m}{\hbar^2}\sum_l v_{l,n,0} \exp\Bigg[2\pi i l\frac{1}{a_x}x\Bigg], \tag{5c}$$

$$w_n(x + a_x) = w_n(x). \tag{5d}$$

For the reason given above, we write the Schrödinger equation immediately without its z-dependence. It then reads in the Landau gauge

$$\mathscr{H}\psi = \partial^2\psi/\partial x^2 + \partial^2\psi/\partial y^2 - 2i\alpha x\partial\psi/\partial y - \alpha^2 x^2\psi - V(x,y)\psi. \tag{6}$$

Here $V(x,y)$ is defined in (5b) and α is equal to

$$\alpha = eH/\hbar c. \tag{7}$$

Without the potential term a solution would be

$$\psi = h(x - x_0)\, e^{i\alpha x_0 y}, \tag{8}$$

[a)]This work was supported by the National Science Foundation.

[b)]Permanent address: Physics Department, University of Oregon, Eugene, OR 97403.

 0022-2488/80/122844-03$1.00

where h is a Hermite function and x_0 an arbitrary displacement. The introduction of the periodic potential will force us to discard the Hermite functions, and to pay attention to rationality. On the other hand, the coupling of the y exponential and the x displacement can be retained. It is the structurally decisive element for the contemplated simplification.

To bring in the integer p we introduce the flux ϕ through a unit cell, expressed in units of the flux quantum

$$\phi = (e/hc)Ha_x b_y = 1/p. \tag{9}$$

This yields with (7)

$$\alpha = 2\pi/pa_x b_y. \tag{10}$$

We take ψ as a Fourier series in y, as follows:

$$\psi = \sum_{m=-\infty}^{+\infty} f_m(x - p(\mu+m)a_x) \times \exp\left[2\pi i\left\{\nu m - \frac{pa_y}{2b_y}(\mu+m)^2 + (\mu+m)\frac{y}{b_y}\right\}\right]. \tag{11}$$

The exact form written down here needs, strictly speaking, no justification; it will justify itself during the derivation. The rationale for all these terms may be found in my paper on quantum numbers.[4] Two quantum numbers, μ and ν, appear in (11); they are fractions. The quadratic exponential takes care of nonrectangular lattices. The displacement in the argument of f_m respects (8). It will be one of the results of the derivation that there should be no index m on f. No attention has been paid to the normalization of ψ as we do not intend to take matrix elements.

Substitution of (11) and (10) into (6) yields

$$\mathscr{H}\psi = \sum_{m=-\infty}^{+\infty} \exp\left[2\pi i\left\{\nu m - \frac{pa_y}{2b_y}(\mu+m)^2 + (\mu+m)\frac{y}{b_y}\right\}\right] \times\left[\frac{d^2}{dx^2}f_m(x-p(\mu+m)a_x) - \frac{4\pi^2}{p^2 a_x^2 b_y^2} \times (x-p(\mu+m)a_x)^2 f_m(x-p(\mu+m)a_x) - V(x,y)f_m(x-p(\mu+m)a_x)\right]. \tag{12}$$

Special attention must be paid to the term containing the potential. From (5b) and (12)

$$\text{potential term} = -\sum_m \exp\left[2\pi i\left\{\nu m - \frac{pa_y}{2b_y}(\mu+m)^2 + (\mu+m)\frac{y}{b_y}\right\}\right] \times \sum_n \exp\left[2\pi i\left\{-n\frac{a_y}{a_x b_y}x + n\frac{y}{b_y}\right\}\right] \times w_n(x) f_m(x - p(\mu+m)a_x).$$

To restore to this expression the character of a Fourier series in y we substitute

$$m \to m - n,$$

$$(\mu+m)^2 \to (\mu+m)^2 - 2n(\mu+m) + n^2,$$

which yields

$$\text{potential term} = -\sum_m \exp\left[2\pi i\left\{\nu m - \frac{pa_y}{2b_y}(\mu+m)^2 + (\mu+m)\frac{y}{b_y}\right\}\right] \times \sum_n \exp\left[2\pi i\left\{-\nu n - n\frac{a_y}{a_x b_y} \times (x - p(\mu+m)a_x) - \frac{pa_y}{2b_y}n^2\right\}\right] \times w_n(x) f_{m-n}(x - p(\mu+m-n)a_x).$$

Writing (12) as an equation

$$(\mathscr{H} - \epsilon)\psi = 0, \tag{13}$$

and returning to it the potential term in the form just obtained we get

$$\sum_m \exp\left[2\pi i\left\{\nu m - \frac{pa_y}{2b_y}(\mu+m)^2 + (\mu+m)\frac{y}{b_y}\right\}\right] \times\left[\frac{d^2}{dx^2}f_m(x-p(\mu+m)a_x) - \frac{4\pi^2}{p^2 a_x^2 b_y^2} \times (x-p(\mu+m)a_x)^2 f_m(x-p(\mu+m)a_x) - \sum_n \exp\left[-2\pi i\left\{\nu n + n\frac{a_y}{a_x b_y} \times (x-p(\mu+m)a_x) + \frac{pa_y}{2b_y}n^2\right\}\right] \times w_n(x) f_{m-n}(x-p(\mu+m-n)a_x) - \epsilon f_m(x-p(\mu+m)a_x)\right] = 0. \tag{14}$$

Equation (14) is a Fourier series in y which vanishes. It can only do so if every Fourier coefficient vanishes. If we consider the mth Fourier coefficient we see that, except for $w_n(x)$, x always occurs in the combination $x - p(\mu+m)a_x$. So if we set

$$x - p(\mu+m)a_x \to x$$

we get the simpler form

$$\frac{d^2 f_m(x)}{dx^2} - \frac{4\pi^2}{p^2 a_x^2 b_y^2}x^2 f_m(x) - \sum_n w_n(x + p\mu a_x) f_{m-n}(x + pna_x) \times \exp\left[-2\pi i\left\{\nu n + n\frac{a_y}{a_x b_y}x + \frac{pa_y}{2b_y}n^2\right\}\right] = \epsilon f_m(x). \tag{15}$$

Here (5b) has produced the final simplification: the disappearance of m from the equation.

The equations are now all alike except for the appearance of m in the index of f. This would still permit an exponential dependence of f_m on m; however we have anticipated this in (11) by introducing the exponential $\exp[2\pi i\nu m]$. f_m

does, therefore, not depend on its index and we end up with

$$\frac{d^2f(x)}{dx^2} - \frac{4\pi^2}{p^2a_xb_y}x^2f(x) - \sum_n w_n(x + p\mu a_x)f(x + pna_x) \times \exp\left[-2\pi i\left\{\nu n + n\frac{a_y}{a_xb_y}x + \frac{pa_y}{2b_y}n^2\right\}\right] = \epsilon f(x), \quad (16)$$

which is the desired equation. It becomes Schellnhuber's[1] working equation (B58) if we set $a_y = 0$, $a_x = b_y$, and

$$w_0(x) = 2V_0 \cos 2\pi x/a_x,$$

$$w_1(x) = w_{-1}(x) = V_0,$$

with all other w's equal to zero. His variable x equals $(\sqrt{2\pi})/(\sqrt{p})\,a$ times the variable x used here.

Schellnhuber,[1] and Schellnhuber and Obermair[2] have shown that the Ritz method is the proper way to solve (16). To reach practical situations, p must be made very large, between 100 and 1000. In connection with such a situation, one wishes to ask what boundary conditions are to be associated with Eq. (16). This question has no answer in principle because $f(x)$ is not a wave function, but an auxiliary function permitting us to construct the wave function, using (11). Professor Obermair[5] pointed out to me, however, that it is very likely that the solution of (16) converges like a Gaussian. The reason is that the two first terms will dominate in the equation for large x. This leads to the alternative of a square exponential increase or decrease. Since an increase is out of the question only the possibility of a decrease remains.

3. EXTENSION TO ALL RATIONAL FIELDS

The extension of the preceding derivation to a general rational field departs from the preceding text at Eq. (9) which is to be replaced by

$$\phi = q/p;\ q,\ p \text{ integers prime to each other.} \quad (17)$$

Thereupon p is to be replaced by p/q in all subsequent formulas up to Eq. (15). In Eq. (15), the argument of w_n reads now

$$x + (p/q)(\mu + m)a_x. \quad (18)$$

When $q = 1$, Eq. (5b) applies and m can be dropped. With $q \neq 1$, this reasoning does not work: m remains in the equation and the text subsequent to (15) is not correct. However, it is not totally invalid: the equation $m + 1$ is different from the equation m, but the equation $m + q$ is not. The Floquet argument used thus remains valid for an advance by q steps. In Eq. (16) $f_m(x)$ has to retain its index, but only modulo q. The single Eq. (16) becomes therefore a system of q coupled equations. They look essentially like (15), with the argument of w_n modified as discussed above and the index of f_m taken modulo q.

We have thus shown that for all rational fields the Schrödinger equation is reducible to a one-dimensional problem. The problem is a single difference-differential equation if the flux ϕ is the reciprocal of an integer. For $\phi = q/p$ it is a system of a q coupled equations.

The work of Schellnhuber and Obermair[1,2] deals entirely with the single equation arising for $q = 1$. There are several reasons why this should remain so for some time. First of all, ϕ is very small for practically attainable fields and therefore the choice $\phi = 1/p$ leaves us plenty of options. Secondly, the complications of the energy spectrum occurring between $\phi = 1/p$ and $\phi = 1/(p + 1)$ were carefully analyzed by Hofstadter.[6,7] He called such an interval a *cell* and the contents of such a cell are covered by the nesting theorem he suggested. Thirdly, even Eq. (16) is fairly hard to solve. The method available is the Ritz method; with this method progress is made by gradually guessing at structural features of the wave function; this is a cumbersome process. Extension of the Ritz method to equation systems is probably possible, but the guessing would become painfully slow.

ACKNOWLEDGMENT

I wish to express my thanks to Professor G. M. Obermair for his fruitful interest in this topic, and to the University of Regensburg for its hospitality. I also want to thank the National Science Foundation for its continued support of this investigation.

[1] H. J. Schellnhuber, thesis, University of Regensburg, 1980.
[2] H. J. Schellnhuber and G. M. Obermair, Phys. Rev. Lett. **45**, 276 (1980).
[3] A. Rauh, G. H. Wannier, and G. Obermair, Phys. Status Solidi B **63**, 215 (1974).
[4] G. H. Wannier, Phys. Status Solidi B **100**, 163 (1980).
[5] G. M. Obermair (private communication).
[6] D. R. Hofstadter, thesis, University of Oregon, 1975.
[7] D. R. Hofstadter, Phys. Rev. B **14**, 2239 (1976).

Quality factor constrained, antenna pattern synthesis using a restricted class of aperture functions

John M. Jarem
University of Petroleum and Minerals, Dhahran, Saudi Arabia

(Received 8 February 1980; accepted for publication 27 May 1980)

The quality factor [Q] for the E-plane strip source antenna is minimized with respect to a class of aperture functions for which the [Q] converges. A complete basis for the above class of functions is constructed for the first time and this basis is used to minimize the above mentioned [Q]. The functions which minimize the above [Q] turn out to be doubly orthogonal and are used to implement a constrained aperture synthesis procedure. New results concerning the maximum bandwidth (or minimum [Q]) of the E-plane strip source antenna are given.

I. INTRODUCTION

Many investigations[1-7] have dealt with the problem of defining a suitable quality factor Q for antennas. Later, many of these quality factors have been used as a constraint parameter in an antenna source synthesis procedure.[4,8] Collin and Rothschild[4] and also Rhodes[8] have defined the Q of an antenna system operated at resonance as the ratio $2\omega W_{e,m}/P$, where $W_{e,m}$ is the greater of the electric and magnetic energy in the reactive field of the antenna and P is the radiated power of the system. The Q is an important parameter because it is a measure of the energy stored in the near field of the antenna and because it is inversely proportional to the bandwidth of the system.

In the definition of the Q a great difficulty which arises is due to the fact that the wavenumber integrands which describe the electric and magnetic energy densities in the evanescent or invisible radiation regions turn out to exhibit a strong singularity at the wavenumbers where $k = (k_x^2 + k_y^2)^{1/2} = \omega/C = k_0$. (This circle in the k_x, k_y wavenumber plane defines the boundary between the visible and invisible radiation ranges.) The strong singularity in turn causes the integrals describing the electric and magnetic energies $W_{e,m}$ to diverge to infinity.

Rhodes has attempted to resolve this difficulty[8] by removing the singular portions of the divergent integrals and defining a new set of what he terms observable electric and magnetic energies $\langle W_e \rangle$ and $\langle W_m \rangle$ based on the convergent terms of the original energy integrals. Rhodes provided a physical basis for this redefinition by noting that: (1) the removal of the singular terms in the above energy integrals would not cause a corresponding change in the bandwidth of the system ($BW\alpha\ 1/Q$), and that (2) the electric and magnetic field components which made up the singular terms did not contribute to the complex Poynting power at the aperture and for this reason the singular terms were not physically meaningful. A synthesis procedure for the E-plane strip source antenna was based on these observable stored energies. These observable stored energies however have been criticized in Refs. 5 and 6 as not being unique and therefore not related to the bandwidth of the system.

Collin and Rothschild[4] have further observed that the energy integrals are, however, not divergent for all aperture distributions. They have shown that for those aperture distributions whose pattern space factor $F(k)$ (or Fourier transform of the aperture distribution) vanishes at the point $k = k_0$, not only is the Q convergent but it is also proportional to the bandwidth of the antenna system when the Q is large. However, due to the fact that $W_{e,m}$ does diverge for those aperture functions for which $F(k_0) \neq 0$, Collin,[6] has discounted the above Q as being physically unmeaningful.

This investigation will be directed at the second problem mentioned at the beginning of the Introduction, namely the development of a Q constrained synthesis procedure. At the present time no Q constrained synthesis procedure has been developed for the above mentioned $Q = 2\omega W_{e,m}/P$,[4,8] due mainly to the divergence of this Q for certain aperture distributions. A synthesis procedure based on this Q may, however, be constructed if the class of allowable aperture distributions for the synthesis procedure is restricted to that class for which $F(k_0) = 0$. If a synthesis procedure is based on this restricted class of functions then the Q will be convergent and also inversely proportional to the physically measurable parameter, namely the bandwidth. In this case it will not be necessary to discount the Q as physically unmeaningful as did Collin[6] since the Q converges, nor will be necessary to redefine the Q by removing certain terms as did Rhodes.[8]

This investigation will construct synthesis procedure for the one-dimensional E-plane strip source antenna using the above mentioned Q [see Eq. (1) of this paper]. Previously Rhodes has presented a synthesis procedure[8] for this antenna with the $(u^2 - 1)^{-3/2}$ term removed.

The investigation will be divided into three parts. The first part will construct a complete set of aperture functions which will: (1) satisfy the proper physical boundary conditions in the aperture, and will (2) span the space of all aperture functions whose pattern space factor $F(k)$ vanishes at the point $k = k_0$. The second part will be concerned with minimizing the quality factor of Eq. (1) with respect to the above mentioned functions. The purpose of this will be to construct a set of doubly orthogonal functions as in Ref. 9, which may be used to implement an antenna source synthesis procedure, which is the third part. Also the minimum value of the above quality factor will be given which will give for the first time the true maximum bandwidth of the E-plane strip source antenna.

 0022-2488/80/122847-05$1.00

TABLE I. Minimum quality factor for the E-plane strip source antenna.

c \ Q	Q_0	Q_1	Q_2	Q_3
0.1π	0.211710 + 04	0.212460 + 07	...	...
$\frac{3\pi}{2}$	0.419030-02	0.162736 + 00	0.351848 + 01	0.708036 + 02
$\frac{3.01\pi}{2}$	0.417781-02	0.162279 + 01	0.350823 + 01	0.705678 + 02
6.0	0.363281-03	0.167662-01	0.371560 + 00	0.606426 + 01
$\frac{5\pi}{2}$	0.105174-04	0.630705-03	0.173314-01	0.296167 + 00
3π	0.509426-06	0.369508-04	0.123749-02	0.250867-01
6.5π	0.392025-15	0.157452-12	...	...

II. ANALYSIS

The E-plane strip source antenna consists of an infinite strip of width a embedded in a conducting screen with the electric field polarized in the direction normal to the aperture edge. Since the electric energy for this antenna is always greater than the magnetic, the Quality factor for this antenna is given by:

$$[Q] = \frac{2\omega W_e}{P} = \frac{\int_{|u|>1}[(u^2-1)^{-1/2} + \frac{1}{2}(u^2-1)^{-3/2}]|F(u)|^2\,du}{\int_{-1}^{1}(1-u^2)^{-1/2}|F(u)|^2\,du} = \frac{G_2(F)}{G_1(F)}, \tag{1}$$

where $F(u) = \int_{-1}^{1} f(t)e^{jcut}\,dt$. (2)

In this expression $F(u)$ represents the far field pattern space factor, $f(t)$ is a function which is proportional to the electric field in the aperture, $t = 2X/a$ is a normalized aperture variable, $u = k_x/k_0$ is a normalized wavenumber, and $c = \pi a/\lambda = k_0 a/2$ electrical length of the aperture. W_e and P in this expression may be found in Ref. 8, Eqs. (3.37b) and (3.5) respectively after setting $k_y = 0$, $F_y = 0$, and after making a change of variables. The region $|u| > 1$ represents the reactive field or invisible region and $|u| \leqslant 1$ the visible radiation region. Rhodes' observable electric energy is obtained if the term proportional to $(u^2-1)^{-3/2}$ is omitted in Eq. (1).

The first part of this paper will deal with the minimization of Eq. (1) with respect to the restricted class of aperture functions in $L^2[-1,1]$ that (i) satisfies the correct physical boundary conditions in the aperture (Ref. 8, pp. 34–50) namely that

$$f(t)|_{t=\pm 1} = 0, \tag{3}$$

and (ii) satisfies the condition that $F(u)$ is zero at $u = 1$ and $u = -1$, namely that

$$F(\pm 1) = 0 = \int_{-1}^{1} f(t)e^{\pm jct}\,dt. \tag{4}$$

The second condition ensures the convergence of Eq. (1) and is equivalent to the statement that every $f(t)$ that meets (i) and (ii) is orthogonal to $\cos ct$ and $\sin ct$. It may be shown that the class of functions which satisfy (i) and (ii) in $L^2[-1,1]$ forms a complete (in the sense of Cauchy) and therefore closed vector subspace which we shall call V.

To carry out the aforementioned minimization we must first find a set of functions which form a total basis for the vector subspace V. To this end we first note that a complete (in the sense of a total basis) set of functions which satisfy the boundary conditions (i) are given by the functions

$$\phi_k = \begin{cases} \cos\underline{k}t, & k = 0,2,4,\cdots \\ \sin\underline{k}t, & k = 1,3,5,\cdots \end{cases}, \tag{5}$$

where $\underline{k} = (k+1)\pi/2$. We also note that the Fourier transform of the above functions which will be useful later is given by

$$\phi_k(u) = \int_{-1}^{1} \phi_k\, e^{jcut}\,dt = j^k \begin{cases} \dfrac{-2\underline{k}\cos cu}{(cu)^2 - \underline{k}^2}, & k = 0,2,4,\cdots \\ \dfrac{-2\underline{k}\sin cu}{(cu)^2 - \underline{k}^2}, & k = 1,3,5,\cdots. \end{cases} \tag{6}$$

The second step in creating a total basis for V is to construct a set of functions $\{h_k\}_{k=0}^{\infty}$ from appropriate linear combinations of the functions ϕ_k in Eq. (5). These linear combinations will be chosen in such a way that each h_k will satisfy condition (ii), each h_k will be orthogonal to $h_{k-1}, h_{k-2}, \cdots h_0$, and each h_k will be normalized to unity. The functions h_k for which $k = 0,2,4,\cdots$ will be even functions constructed from the even ϕ_k functions and the functions h_k for which $k = 1,3,5,\cdots$ will be odd functions constructed from the odd ϕ_k functions. In this construction two different cases occur, namely the case when $c \neq (k+1)\pi/2$, $k = 0,1,2,\cdots$ and the case when $c = (k+1)\pi/2$ for some particular integer k_0.

In the first case when $c \neq (k+1)\pi/2$, $k = 0,1,2,\cdots$ the series that results for each h_k is given by

$$h_k(t) = \sum_{i=0,1}^{k+2}{}' a_{i,k}\phi_i(t) \tag{7}$$

and the series for its associated Fourier transform, call it H_k, is given by

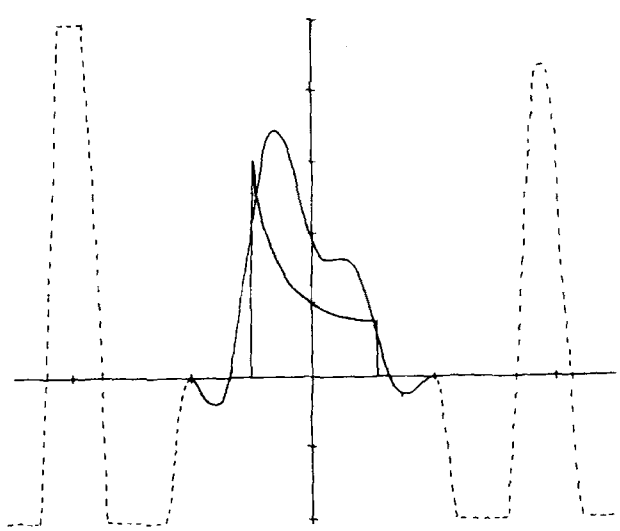


FIG. 1. Radiation approximation of an ideal cosecant radiation pattern with $Q = 10$ and $a = 2.5\lambda$ for the E-plane strip source antenna.

$$H_k(u) = \sum_{i=0,1}^{k+2\prime} a_{i,k}\Phi_i(u)\,. \tag{8}$$

The primed summation in (7) and (8) means summation over the even or odd values of i according to whether k is even or odd. The coefficients $a_{i,k}$ which ensure that h_k satisfies conditions (i) and (ii) are derived in Appendix A.

In the second case if $c = (k_0 + 1)\underline{\pi}/2$ and $k_0 = 0,2,4,\cdots$, then h_k is given by

$$h_k = \phi_k, \quad k = 0,2,\ldots,k_0 - 2\,,$$
$$h_k = \phi_{k+2}, \quad k = k_0, k_0 + 2,\ldots\,, \tag{9}$$

and the h_k for $k = 1,3,5$ are determined from Appendix A. If in the second case $c = (k_0 + 1)\underline{\pi}/2$ and $k_0 = 1,3,5,\cdots$ then h_k is given by

$$h_k = \phi_k, \quad k = 1,3,5,\cdots k_0 - 2,$$
$$h_k = \phi_{k+2}, \quad k = k_0, k_0 + 2,\cdots \tag{10}$$

and the h_k for $k = 0,2,4,\cdots$ are determined from Appendix A. In other words, in the second case a total basis for V is found by simply eliminating the ϕ_{k_0} function from Eq. (5). The functions H_k for the second case found from the Fourier transform of (9) and (10).

III. DOUBLY ORTHOGONAL FUNCTIONS

By using the theory in Ref. 9 the set of functions $\{h_k(t)\}$ may be used to construct a set of doubly orthogonal functions

$$f_n(t) = \sum_{i=0,1}^{\infty\prime} X_{i,n}h_i(t)\,, \tag{11}$$

$$F_n(u) = \sum_{i=0,1}^{\infty\prime} X_{i,n}H_i(u)\,, \tag{12}$$

from the extremal functions of the functional $[Q]$ of Eq. (1). The set of coefficients $X_{i,n}$ are derived from the matrix equation

$$[G_2][X] = Q[G_1][X]\,, \tag{13}$$

where

$$G_{1_{kk'}} = \int_{-1}^{1}(1-u^2)^{-1/2}H^*_{k'}H_k\,du, \tag{14}$$

$$G_{2kk'} = \int_{|u|>1}[(u^2-1)^{-1/2} + \tfrac{1}{2}(u^2-1)^{-3/2}] \times H^*_{k'}H_k\,du\,. \tag{15}$$

The matrices $[G_1]$ and $[G_2]$ are Hermitian and positive definite and the matrix elements are zero whenever k' is even or k is odd or vice versa.

The functions F_n satisfy the double orthogonality relations

$$\int_{|u|>1}[(u^2-1)^{-1/2} + \tfrac{1}{2}(u^2-1)^{-3/2}]F^*_mF_n\,du = Q_nN_n\delta_{m,n}, \tag{16}$$

$$\int_{-1}^{1}(1-u^2)^{-1/2}F^*_mF_n\,du = N_n\delta_{m,n}\,, \tag{17}$$

where N_n is a positive normalization constant.

The eigenvalues $Q_n(c)\alpha 1/BW$ are shown as a function of c, the electrical length of the antenna in Table I. Harwell subroutine EA 12 AD was used to solve the eigenvalue Eq. (13) on an IBM 360–158 computer. The matrices $[G_1]$ and $[G_2]$ were truncated at values of $k = 49$.

It is interesting and also an excellent verification of the numerical procedures used here that the Q for the 1.5005Π case (using the h_k functions of Appendix A) compares so closely with the Q for the $1.5\,\underline{\pi}$ case [(using Eq. (9)] despite the fact that totally different bases were used in each case. This also seems to indicate that the minimum Q is not sensitive to the aperture width when the aperture width assumes integral or half integral values of the wavelength.

IV. SYNTHESIS PROCEDURE

We will now be concerned with using the f_n functions in a constraint synthesis procedure. The synthesis procedure will consist of the minimization of a least-square error functional of the form

$$\epsilon = \int_{-1}^{1}(1-u^2)^{-1/2}|F(u) - \hat{F}(u)|^2\,du + \mu[G_2(F) - Q_RG_1(F)]\,, \tag{18}$$

where $G_1(F)$ and $G_2(F)$ are the functionals of Eq. (1). In this equation $\hat{F}$ is an ideal, desired far-field radiation pattern which is to be approximated, F is the radiation pattern approximation given by Eq. (2), Q_R is a prescribed value of the Q (or BW) to which the approximate pattern $F(u)$ is to correspond, ϵ represents the error between the ideal and approximate radiation patterns, and μ is a Lagrange multiplier. This equation forms the basis of a constraint procedure which was first given by Rhodes[8] and later applied by him to the E-plane strip source antenna based on his observable energies. This equation has also been used in Ref. 10 for the H-plane strip source antenna.

The minimization of ϵ is accomplished by expanding $F(u)$ in a series of the doubly orthogonal functions F_n of Eqs. (16) and (17) and then varying ϵ with respect to each of the coefficients in this series. The double orthogonality properties of the functions F_n reduce Eq. (18) to a simple sum of the squares of the coefficients of the series of F which can then be minimized easily. The details of the minimization are given in Refs. 8 and 10 and are not repeated here.

Figure 1 illustrates the success of this procedure when applied to an ideal cosecant pattern with $c = 2.5\,\underline{\pi}$ and $Q_R = 10$.

V. CONCLUSION

The minimization of the quality factor of the E-plane strip source antenna has been found for the first time with respect to the restricted class of aperture functions V in $L^2[-1,1]$ which (i) vanish at the aperture edges, and (ii) whose pattern space factor vanishes at the boundary of the visible and invisible radiation ranges. The minimization was performed by constructing a total basis $\{h_k\}$ for V, expressing the quality factor functional $[Q]$ in a matrix quadratic form with respect to this basis, and then minimizing the quadratic form. By using the associated matrix eigenvectors, a set of doubly orthogonal functions was contructed and these doubly orthogonal functions were used as a set of basis func-

tions with which a constrained aperture synthesis procedure was implemented.

In conclusion, this author believes that the minimum quality factor which was obtained by minimizing Eq. (1) with respect to the functions in V, is also the minimum quality factor that would be obtained if Eq. (1) was minimized with respect to all functions in $L^2[-1,1]$. This conclusion is reached since any function which is a member of $L^2[-1,1]$ and not a member of V would make the $[Q]$ of Eq. (1) infinite.

APPENDIX A

We will derive the coefficients $a_{i,k}$ for the even h_k functions first and later only state results for the odd h_k functions. Let $h^e_1, h^e_2 \cdots$ and $H^e_1, H^e_2, \cdots$ be the unnormalized series given by

$$h^e_n = \phi_0 + \sum_{i=1}^{n} A_{i,n}\phi_{2i} \tag{A1}$$

$$H^e_n = \Phi_0 + \sum_{i=1}^{n} A_{i,n}\Phi_{2i} \quad n = 1,2,3\cdots . \tag{A2}$$

The coefficient $A_{1,1}$ is chosen so that $H^e_1(1) = 0$ or

$$A_{1,1} = -\Phi_0(1)/\Phi_2(1) . \tag{A3}$$

The coefficient $A_{1,2}$ and $A_{2,2}$ are chosen so that $\int_{-1}^{1} h^e_1 h^e_2\, dt = \langle h^e_1, h^e_2\rangle = 0$ and $H^e_2(1) = 0$, which implies that $A_{1,2} = -1/A_{1,1}$ and

$$A_{2,2} = -\frac{[\Phi_0^2(1) + \Phi_2^2(1)]}{\Phi_0(1)\Phi_4(1)} . \tag{A4}$$

To proceed further we note that for $n \geqslant 3$ the conditions $\langle h^e_n, h^e_i\rangle = 0$, for $i = 1,\ldots, n-2$ leads to the conclusion (after use of the orthogonal properties of the ϕ_k functions) that

$$\begin{aligned} &A_{1,n} = A_{1,n-1} = \cdots = A_{1,2}, \\ &A_{2,n} = A_{2,n-1} = \cdots = A_{2,3}, \\ &\vdots \\ &A_{n-2,n} = A_{n-2,n-1} \quad n \geqslant 3 . \end{aligned} \tag{A5}$$

This then implies that the series for h^e_n may be given by

$$h^e_n = \phi_0 + \sum_{i=1}^{n-1} A_{i,i+1}\phi_{2i} + A_{n,n}\phi_{2n} . \tag{A6}$$

Let us suppose that the coefficients of h^e_{n-1}, $h^e_{n-2}, \ldots$, h_1 have all been found for $n \geqslant 3$. Then to determine the coefficients of h^e_n we need only determine $A_{n-1,n}$ and $A_{n,n}$ as $A_{i,i+1}$ for $i = 1,\ldots, n-2$ are known. By using the condition $\langle h^e_n, h^e_{n-1}\rangle = 0$ and $H^e_n(1) = 0$ and a little algebraic manipulation we find for $n \geqslant 3$.

$$A_{n-1,n} = -\frac{1}{A_{n-1,n-1}}\left[1 + \sum_{i=1}^{n-2} A^2_{i,i+1}\right], \tag{A7}$$

$$A_{n,n} = \frac{1}{A_{n-1,n-1}}\left\{\frac{\Phi_{2n-2}(1)}{\Phi_{2n}(1)}\left[1 + \sum_{i=1}^{n-2} A^2_{i,i+1} + A^2_{n-1,n-1}\right]\right\} . \tag{A8}$$

As can be seen Eqs. (A7) and (A8) provide a simple recursion relation for which the coefficients of $A_{n-1,n}$ and $A_{n,n}$ may be found for any order $n \geqslant 3$. Equation (A5) gives the remaining coefficients for order n. We also note that since $\Phi_k(1) \neq 0$ if $c \neq (k+1)\Pi/2$ [as can be seen from Eq. (6)] that this implies that $A_{1,1}, A_{1,2}, A_{2,2} \neq 0$. This in conjunction with the fact that all of the $\Phi_k(1) \neq 0$, implies that $A_{i,k} \neq 0$ for all i and k.

The coefficients for the odd h_k functions may be found if we let $h^0_1, h^0_2, \cdots$ and $H^0_1, H^0_2, \cdots$ be the unnormalized series given by

$$\begin{aligned} h^0_n &= \phi_1 + \sum_{i=1}^{n} B_{i,n}\phi_{2i+1} , \\ H^0_n &= \Phi_1 + \sum_{i=1}^{n} B_{i,n}\Phi_{2i+1} \quad n = 1,2,3,\cdots , \end{aligned} \tag{A9}$$

and we apply the conditions $H^0_n(1) = 0$ and $\langle h^0_n, h^0_i\rangle = 0$, $i = 1,\ldots, n-1$ to these series. The analysis is similar to the even case and the coefficients are found to be given by the equations

$$\begin{aligned} B_{1,1} &= -\Phi_1(1)/\Phi_3(1) , \\ B_{1,2} &= -1/B_{1,1} , \\ B_{2,2} &= -[\Phi_1(1) + B_{1,2}\Phi_3(1)]/\Phi_5(1) . \end{aligned} \tag{A10}$$

For $n \geqslant 3$

$$\begin{aligned} &B_{1,n} = B_{1,n-1} = \cdots = B_{1,2}, \\ &B_{2,n} = B_{2,n-1} = \cdots = B_{2,3}, \\ &\vdots \\ &B_{n-2,n} = B_{n-2,n-1}, \\ &B_{n-1,n} = -\frac{1}{B_{n-1,n-1}}\left[1 + \sum_{i=1}^{n-2} B^2_{i,i+1}\right], \\ &B_{n,n} = \frac{1}{B_{n-1,n-1}}\left[\frac{\Phi_{2n-1}(1)}{\Phi_{2n+1}(1)}\right] \\ &\quad \times\left[1 + \sum_{i=1}^{n-2} B^2_{i,i+1} + B^2_{n-1,n-1}\right] . \end{aligned}$$

In the above expressions we note that all of the ratios $\Phi_1(1)/\Phi_5(1)$ and $\Phi_{2n-1}(1)/\Phi_{2n+1}(1)$ for $n = 1,2,\cdots$ are real since Φ_n is purely imaginary for $n = 1,3,5,\cdots$.

If we normalize the coefficients $A_{i,n}$ and $B_{i,n}$ and also reorder the indices to conform with Eq. (7) we find for $n = 1,2,3\ldots$

$$\begin{aligned} a_{0,2n-2} &= 1/s_n , \\ a_{2i,2n-2} &= A_{i,n}/s_n \quad i = 1,2,\cdots n , \\ a_{1,2n-1} &= 1/r_n , \\ a_{2i+1,2n-1} &= B_{i,n}/r_n \quad i = 1,2,\cdots n , \end{aligned}$$

where

$$s_n = \left[1 + \sum_{i=1}^{n} A^2_{i,n}\right]^{1/2}$$

and

$$r_n = \left[1 + \sum_{i=1}^{n} B^2_{i,n}\right]^{1/2} .$$

The above scheme has been used to generate numerically all the coefficients up to $h_{49}(t)$ for many values of c. The scheme is very stable and satisfies condition (ii) and the orthogonality requirement very accurately.

[1]R. E. Collin and S. Rothschild, IEEE Trans. Antennas Propag. **AP 12**, 23–7 (1964).
[2]R. M. Kalafus, IEEE Trans. Antennas Propag. **AP 17**, 729–32 (1969).
[3]R. L. Fante, IEEE Trans. Antennas Propag. **AP 17**, 151 (1969).
[4]R. E. Collin, Can. J. Phys. **41**, 1967–79 (1963).
[5]G. V. Borgiotti, IEEE Trans. Antennas Propag. **AP 15**, 565–6 (1967).
[6]R. E. Collin, Trans. Antennas Propag. **AP 15**, 567–9 (1967).
[7]D. R. Rhodes, J. Franklin Inst. **302**, #3 (1976).
[8]D. R. Rhodes, *Synthesis of Planar Aperture Sources* (Oxford U. P., New York, 1974).
[9]J. M. Jarem, J. Math. Phys. **21**, 28 (1980)
[10]J. M. Jarem, IEEE Trans. Antennas Propag. **AP 28**, 36–41 (1980)..

ERRATA

Erratum: On the coupling of self-conjugate systems with $\overline{SL(3,R)}$ symmetry [J. Math. Phys. 20, 1615 (1979)]

J. A. Castilho Alcarás
Instituto de Física Teórica, São Paulo, Brasil

L. C. Biedenharn
Department of Physics, Duke University, Durham, North Carolina 27706

(Received 22 July 1980; accepted for publication 31 July 1980)

(1) In Eq. (2.1) the first commutator is between J_0 and $J_{\pm 1}$.
(2) In the definition of JJT in the second of Eqs. (2.3) there is an overall factor of 6 missing on the right-hand side.
(3) In Eq. (2.5), on the left-hand side the bra SO(2) label is M' and the last ket on the right-hand side must be $|J'M'\rangle$.
(4) In Eq. (2.6) the first factor under the square root should read: $(2e + 1)$.
(5) In Eq. (2.19) under the square root symbol one should have $(2J + 1)$.
(6) In the second of Eqs. (2.21) (the last equation on page 1617) the first factor in the denominator should be $(2J + 1)$ (not $2J + 3$ as printed) while in the third equation (top of page 1618) the complete numerator under the square root should be $2(2J)(2J + 2)$.
(7) In Eq. (2.22), the first ket on the right-hand side should read: $|0,\mu_1; J_1 M_1\rangle$.
(8) In Eq. (2.23): ψ_3 should be replaced by μ_3.
(9) In Eq. (2.24): Delete the first ket on the right-hand side. The sum is over α.
(10) In Eq. (2.31) for $f^4{}_{JJ}$ there is an overall minus sign missing; for $f^4{}_{JJ+4}$ a left parenthesis is missing in the factorial appearing in the numerator.

Erratum: Five-term WKBJ approximation [J. Math. Phys. 21, 90 (1980)]

R. N. Kesarwani
Department of Mathematics, University of Ottawa, Ottawa, Canada K1N 9B4

Y. P. Varshni
Department of Physics, University of Ottawa, Ottawa, Canada K1N 9B4

P. 91, Eq. (10): In the expression for I_5, the second term in the first integrand should read $2065V'^2V''V^{(4)}$ instead of $2065V''^2V^{(4)}$.

 0022-2488/80/122852-01$1.00