JOURNAL OF
MATHEMATICAL
PHYSICS

VOLUME 1, NUMBER 6 NOVEMBER–DECEMBER, 1960

Symmetry-Adapted Functions Belonging to the Symmetric Groups

HAROLD V. MCINTOSH*
RIAS, Baltimore 12, Maryland
(Received June 3, 1960)

Young's factorization of idempotents belonging to the symmetric groups is given a necessary and sufficient characterization, by means of a lemma due to Burrow. The use of these idempotents is contrasted with Yamanouchi's representation, and finally the equivalence of Löwdin's path diagram method to the group-theoretical treatment of the angular momentum states arising from the coupling of an assemblage of spin ½ particles is demonstrated.

INTRODUCTION

SUPPOSE that one is given a representation $\Gamma=\{D(e),D(a),\cdots\}$ of a group $G=\{e,a,\cdots\}$ operating on a vector space V, and that the characters $\chi^1(x)$, $\chi^2(x)$, $\cdots$ of the inequivalent irreducible representations $\Gamma^i=\{D^i(e),D^i(a),\cdots\}$ of G are known. Then, from the theory of group representations,[1] there is known a formula for a set of idempotents $\mathcal{O}^i$

$$\mathcal{O}^i=\frac{l_i}{{}^0G}\sum_{a\in G}\chi^i(a)D(a), \tag{1}$$

where l_i is the dimension of the representation Γ^i and ${}^0G$ is the order of G. These idempotents project onto the subspace V^i of V which are stable under the action of G through the representation Γ. Therefore these stable subspaces reduce the representation Γ according to the inequivalent irreducible representations of G which it contains. If one wishes to obtain a basis for the stable subspaces V^i, he may be sure that among the projections of the basis vectors $\{:1\rangle, :2\rangle, \cdots :n\rangle\}$ of V (and hence among the columns of $\mathcal{O}^i$) there are sufficiently many which are linearly independent that a basis for V^i may be constructed.

However, if the representations Γ^i are known explicitly, a more definite choice of basis may be had by means of the projection operators

$$\mathcal{O}_{jj}{}^i=\frac{l_i}{{}^0G}\sum_{a\in G}d_{jj}{}^i(a)D(a), \tag{2}$$

which are constructed with the aid of the diagonal matrix elements $d_{jj}(a)$ of Γ^i. In fact, with respect to a basis chosen with the aid of these operators, we may even be sure that G will have precisely the representation Γ^i.

The use of these operators encounters two difficulties in practice. The first is that whereas the characters of a group are unique, the diagonal matrix elements of its irreducible representations are not, since any similarity transformation applied to such a representation yields another with the same character. This makes it impractical for one to attempt to compile tables of diagonal matrix elements, in the way that he may form character tables, for he may never be sure in advance which of the many equivalent representations are going to be used. As another aspect of this same difficulty we must mention the fact that it is not always easy to obtain any irreducible representation at all in explicit form, much less to find a desired equivalent representation.

Secondly, many groups which one intends to use in practice are of such a large order that it would be useless even to try to write down all the summands appearing in the expression for the projection operators. For example, one need only recall that the symmetric

* Visiting scientist at Quantum Chemistry Group, Uppsala University, Uppsala, Sweden.

[1] Eugene Wigner, *Gruppentheorie und ihre Anwendung auf die Quantenmechanik der Atmospektren* (Friedrich Vieweg und Sohn, Braunschweig, Germany, 1931), pp. 120–133.

group of degree n has $n!$ elements, which is already a very large number for $n=6$. Therefore one must find some means of simplifying formulas (1) and (2) if he intends to use them for a particular calculation. One scheme of simplification has been discussed by a number of authors: this is the scheme of factorizing the projection operators into a product of simpler operators.

Löwdin[2] has described the factorization of the projection operators belonging to the three-dimensional rotation group in terms of annihilation operators which successively remove unwanted portions of V. Melvin[3] has introduced a factorization which is dependent upon a certain form which the representation Γ^i may perchance exhibit, while Fokker[4] has shown how a projection operator may sometimes be factored in terms of projection operators belonging to subgroups of G. This method was extended[5] to the case in which a group is a semidirect product of two of its subgroups. By the use of projective representations it is even possible to treat the projection operators of any group having a nontrivial normal subgroup in terms of operators belonging to the normal subgroup and factor group. If the normal subgroup is non-Abelian, one does not obtain a factorization into a simple product, but rather a sum of factored operators.

Another factorization of the projection operators belonging to a row of a representation has long been used; it is Young's factorization of a set of primitive idempotents belonging to the symmetric groups S_n.[6] His factorization makes use of the characters belonging to a pair of subgroups, which are, respectively, the identity character $\alpha(x)=1$ of the subgroup of permutations preserving the rows of a certain Young tableau T, and the alternating character $\beta(x)=\pm 1$ of the subgroup of permutations preserving the columns of the same tableau. These characters are themselves idempotent. Quite generally, if we assume that θ is a character of a subgroup H and make it a function belonging to the convolution algebra $C(G)$[7] by extending it to vanish outside H, we have

$$(\theta * \theta)(x) = \sum_{a \epsilon G} \theta(a)\theta(a^{-1}x)$$

$$= {}^0H\theta(x). \tag{3}$$

Indeed, any idempotent in $C(H)$ when so extended, if normalized, remains an idempotent in $C(G)$. Unfortunately, even if an idempotent is minimal in $C(H)$, it need not remain so in $C(G)$. It is a surprising fact that the convolution $\alpha * \beta$, when these are the characters associated with a Young tableau, is nevertheless proportional to a minimal idempotent. This combinatorial result, due to Young,[8] was given an algebraic interpretation by von Neumann,[9] who observed that $\alpha * u * \beta = \lambda(u)\alpha * \beta$, whatever element u was chosen from $C(S_n)$. Here $\lambda(u)$ is a scalar multiplier. This means that α projects onto certain right ideals in $C(S_n)$, β onto certain left ideals, and that the only subspace surviving both projections is the subring hull of $\alpha * \beta$.

Von Neumann's theorem was generalized by Burrow[10] to apply to any group G, although, as stated by him, it simply constituted a sufficient condition whereby one could obtain a minimal idempotent and from it the character of an irreducible representation of G. Supposing that θ and ϕ were two linear characters (characters of one-dimensional representations) of subgroups R and C, respectively, his lemma yields a sufficient condition upon the subgroups R and C and upon the characters θ and ϕ that $\theta * \phi$ will be proportional to a minimal idempotent. Nevertheless, it is possible to show that his requirements are actually the necessary and sufficient conditions for

$$\theta * C(G) * \phi = \Lambda\theta * \phi, \tag{4}$$

where the set of multipliers Λ comprises a field.

This result not only shows how the idempotent $\lambda\theta * \phi$ may be factored into idempotents belonging to two subgroups; it shows that the factors are extraordinarily persistent, in that they perform a projection whenever they appear in the proper order at the extreme ends of an expression. In fact, if δ_a is the characteristic function of the element a, we find

$$\theta * \delta_a * \phi = \begin{cases} 0 & \sim(a \epsilon RC) \\ \beta(c)\theta * \phi & a = rc \epsilon RC, \end{cases} \tag{5}$$

so that it is possible to write an explicit formula for the projection of any function in $C(G)$.

YOUNG TABLEAUX

The hypotheses of Burrow's lemma are

$$\left.\begin{array}{l} [a \epsilon RC \text{ and } c \epsilon (R \cap aCa^{-1})] \rightarrow \phi(a^{-1}ca) = \theta(c) \\ \sim(a \epsilon RC) \rightarrow \exists c \epsilon (R \cap aCa^{-1}) \ni \phi(a^{-1}ca) \neq \theta(c). \end{array}\right\} \tag{6}$$

In applying the lemma to the symmetric groups it is convenient to adopt two restrictive assumptions. One concerns the nature of the subgroups R and C; the other concerns the characters α and β. Regarding the subgroups, we postulate that they must each contain a set of transpositions by which they are generated,

[2] P. O. Löwdin, "Nature of the valence bond functions," Proc. Paris Symposium "Calcul des Fonctions d'Onde Moléculaires," 1957.

[3] M. A. Melvin, Revs. Modern Phys. 28, 18 (1956).

[4] A. D. Fokker, Physica 7, 385 (1940).

[5] H. V. McIntosh, "Symmetry-adapted functions belonging to the crystallographic groups," abstracts, Ohio State University Symposium "Molecular Structure and Spectroscopy," Columbus, Ohio, 1958, p. 22.

[6] See D. F. Rutherford, *Substitutional Analysis* (University Press, Cambridge, England), 1948.

[7] See A. Weil, *L'Integration dans les Groupes Topologiques* (Hermann & Cie, Paris, France, 1951), Chap. III.

[8] See D. F. Rutherford, footnote reference 6, for a bibliography of Young's papers.

[9] See B. L. Van der Waerden, *Moderne Algebra II* (Berlin, 1931), Sec. 129.

[10] M. D. Burrow, Can. J. Math. 6, 498 (1954).

just as S_n itself, for instance, is generated by the set of transpositions $\{(1k)\,|\,k=1, 2, \cdots, n\}$. Concerning the characters, we postulate that they be the restriction of either the identity character $\alpha(x)=1$ or the alternating character $\beta(x)=\pm 1$ to their respective subgroup. Since we assume R and C to be generated by transpositions, it follows that their linear characters must take the value ± 1. Since the transpositions all belong to the same class in S_n, α and β are the only linear characters possible; we make the second postulate to preserve this uniformity with respect to the classes of S_n.

When these two restrictions are accepted, it is possible to prove a number of properties of the subgroups R and C of S_n. These properties are such that we may introduce two equivalence relations into the set of digits $\{1,2,\cdots,n\}$. We call x and y R equivalent if the transposition (x,y) belongs to R, accepting the degenerate transposition (x,x) as the identity. If, on the other hand, $(x,y)\epsilon C$, we call x and y C equivalent. The properties of R and C are such that both these relations are equivalence relations. Moreover, R and C possess the additional properties that x and y are not both R and C equivalent (for $x\neq y$), and that there is always a digit a such that *either* x and a are R equivalent and a and y are C equivalent *or* x and a are C equivalent and a and y are R equivalent.

These two equivalence relations may be diagramed, in order that they may be more easily visualized, by drawing a rectangular array like that shown in Fig. 1, in which the R-equivalence classes are horizontal rows and the C-equivalence classes are vertical columns. The properties of R and C are such that: (1) There is at most one digit in each intersection of a row and a column, since whenever two digits lie in the same row they cannot lie in the same column; (2) If x and y lie neither in the same row nor in the same column, then there is either a digit in the same row as x and the same column as y, or else in the same column as x and the same row as y. That is, if we have two diagonally opposite corners of a rectangle filled, then at least one other corner must be filled as well. This property also is shown in Fig. 1.

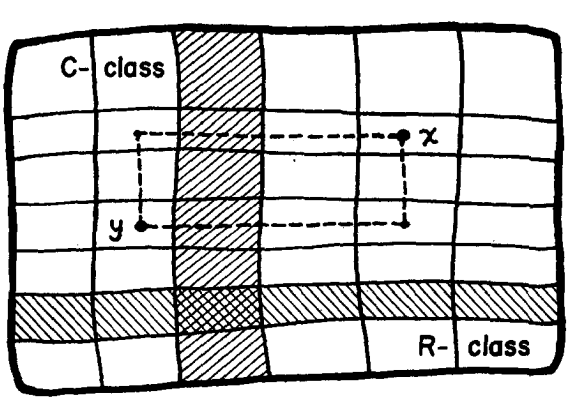

FIG. 1. Equivalence relation diagram.

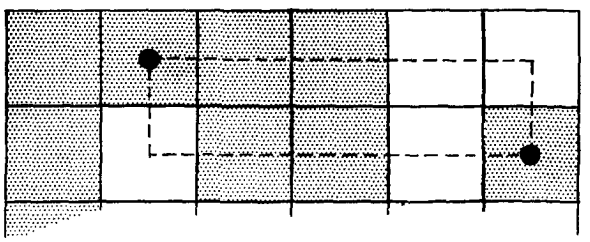

FIG. 2. Construction of the Young tableau.

The diagram may be stylized yet further as shown in Fig. 2, by placing the longest (or one of the longest, if there are several) row at the top of the diagram, so that it is filled solidly, and then the next longest, and so on, always taking care not to exceed the left margin established by the first row. Should there be some digits in the second row which are not C equivalent to any in the first row, they are to be placed to the extreme right in the diagram; however, since there are no more elements in the second row than in the first, this would create a vacancy in the second row below a filled position in the first; this pair of positions taken together with the overjutting element in the second row and the vacancy above it would combine to violate the second property of the diagram. This argument shows that the second row cannot overjut the first. If it is not filled solidly from the left margin, the columns may be rearranged to make it so. A similar argument shows that the third row may not overjut the second, which allows the columns to be rearranged so that the first three rows are filled solidly from the left margin, and so on. Thus the digits $\{1, 2, \cdots, n\}$ may be arranged into a figure which is known in the theory of the symmetric groups as a Young tableau. Since the existence of the pair (x,y) in a common row implies that the transposition (x,y) belongs to R, we see that R may simply be described as the subgroup of permutations preserving the rows of the tableau T. Likewise C is the subgroup preserving the columns of T.

The fact that such a derivation as this is possible shows that if one desires a "split idempotent" in the sense of Burrow's lemma for the symmetric group S_n, he is uniquely led (except for the two assumptions concerning the subgroups and characters) to the Young tableaux. The converse is of course possible—one may show that two subgroups R and C derived from any Young tableau will satisfy the postulates of Burrow's lemma. The only requirement necessary for the characters, other than our hypothesis that they be the restriction of α or β to their respective subgroup, is that they not be both the same. One must be the restriction of α; the other must be the restriction of β.

Thus the Young symmetrizers, as the quantities $\alpha * \beta$ and $\beta * \alpha$ are called, acquire a necessary as well as a sufficient significance insofar as they are the factored idempotents determined by Burrow's lemma.

To be of the greatest utility, as idempotent—even a factored primitive idempotent—should not stand alone.

One would prefer to have a family of commuting idempotents comprising a resolution of the identity. Such a family could be used to form an irreducible representation of S_n, of which they would constitute the diagonal matrix elements. This is a requirement of which the Young symmetrizers fall short. Nevertheless it is interesting to note the way in which they fail of this requirement, and how they are in many instances acceptable as substitutes.

It is first of all apparent that there are far more Young symmetrizers—$n!$ in fact—than dimensions of any representation of S_n, and thus that these idempotents at the very least comprise a highly redundant set. That they are complete, and that tableaux of different shapes belong to different irreducible representations may be shown in the usual fashion.[11] Thanks to formula (5), it may be shown that the redundancy is of a very simple nature, namely, that the idempotents all project *onto* the same linearly independent ideals, but that they each project *along* a different ideal, so that it is a matter of discarding all but one of the idempotents projecting upon a given ideal.

It may be shown that when a number of Young idempotents project upon a given ideal, their tableaux differ in the respect that one can first rearrange the digits within the different rows, and then within the columns after the rows have been rearranged, to get the other. Therefore if one selects that tableau, among the many which are equivalent, for which the digits 1, 2, $\cdots$, n are to be found in increasing order along whatever row one selects, increasing from left to right, and in increasing order down any column, he has systematically selected one tableau from each class of redundant tableaux. These are the "standard tableaux" which play such an important role in the representation theory of the permutation group.

The standard tableaux may be arranged in lexicographic order according to the digits which they contain, and when this is done it will be found that the symmetrizers have the property

$$(\alpha_i * \beta_i) * (\alpha_j * \beta_j) = 0, \quad i > j, \tag{7}$$

where the subscripts distinguish the symmetrizers belonging to different tableaux. With this it is found that one has a set of idempotents such as one obtains in the first stages of the proof of Wedderburn's structure theorem.[12] The introduction of a new set of idempotents defined recursively by the equations

$$\begin{aligned} f_1' &= f_1, \\ f_2' &= (1 - f_1') * f_2, \\ f_3' &= (1 - f_1' - f_2') * f_3, \\ &\cdots \\ f_p' &= (1 - f_1' - f_2' - \cdots - f_{p-1}') * f_p, \end{aligned} \tag{8}$$

where $f_i = \lambda(\alpha_i * \beta_i)$, finally yields a set of idempotents which serve as the diagonal matrix elements of an irreducible representation of S_n. By the use of Eq. (5), this expression may be reduced further, and yields

$$f_s' = (1 - \sum_{0<i<s} b(\sigma_{is})\delta_{\sigma_{is}} + \sum_{0<j<i<s} b(\sigma_{ji})b(\sigma_{is})\delta_{\sigma_{js}} + \cdots) * \alpha_s * \beta_s, \tag{9}$$

where σ_{ij} is the permutation which changes the tableau T_j into the tableau T_i and $b(\sigma_{ij})$ vanishes unless $\sigma_{ij} \epsilon R_j C_j$. In this case we may factor σ_{ij} uniquely into a product $\sigma_{ij} = \sigma_{ik}\sigma_{kj}$ such that $\sigma_{ik} \epsilon C_i$, $\sigma_{kj} \epsilon R_j$; and then we may set $b(\sigma_{ij}) = \beta(\sigma_{ik})$.

If we define

$$\omega_s = (1 - \sum_{0<i<s} b(\sigma_{is})\delta_{\sigma_{is}} + \sum_{0<j<i<s} b(\sigma_{ji})b(\sigma_{is})\delta_{\sigma_{js}} + \cdots), \tag{10}$$

we may set

$$f_s' = \omega_s * \alpha_s * \beta_s, \tag{11}$$

$$= \alpha_s' * \beta_s, \tag{12}$$

thereby defining α_s'. Calculation shows that the general matrix elements of this representation, which is called Young's "natural" representation, are

$$f_{ij} = \lambda \alpha_i' * \delta_{\sigma ij} * \beta_j. \tag{13}$$

IRREDUCIBLE REPRESENTATIONS OF THE SYMMETRIC GROUPS

Once one is in possession of these results, he may proceed to make certain remarks about the irreducible representations of the symmetric groups. First of all, although the idempotents of the symmetric group are given by the formula (2) with the diagonal matrix elements defined by Eq. (9), nevertheless the Young idempotents $\lambda\alpha_i * \beta_i$ project onto the correct subspaces, albeit along the wrong subspaces. In this sense, they are "almost" the projection operators conceived in group representation theory.

Also, since

$$\delta_{\sigma ij} * \alpha_i * \beta_i * \delta_{\sigma ji} = \alpha_j * \beta_j, \tag{14}$$

if we have

$$X = \{\sum_{a \epsilon G} (\alpha_i * \beta_i)(a) D(a)\} X, \tag{15}$$

then

$$D(\sigma_{ji})X = \{\sum_{a \epsilon G} (\delta_{\sigma ji} * \alpha_i * \beta_i * \delta_{\sigma ij})(a) D(a)\} D(\sigma_{ji}) X$$

$$= \{\sum_{a \epsilon G} (\alpha_j * \beta_j)(a) D(a)\} D(\sigma_{ji}) X, \tag{16}$$

and consequently we see that $D(\sigma_{ji})$ plays the role of a transition operator, even though $\delta_{\sigma ji}$ is not an off-diagonal matrix element. Just as the idempotents $\lambda\alpha_i * \beta_i$

[11] See, e.g., H. Boerner, *Darstellungen von Gruppen* (Springer-Verlag, Berlin, Germany, 1955), Chap. IV.

[12] See, e.g., H. Boerner, *Darstellungen von Gruppen* (Springer-Verlag, Berlin, Germany, 1955), Chap. IV, Satz 42, p. 61.

project along the wrong subspaces, so the elements $\delta_{\sigma ji}$ are not nilpotent, but nevertheless cause the correct transitions between certain pairs of states. Other transitions will of course be described incorrectly.

The "transition operators" $\delta_{\sigma ij}$ may be used to simplify formula (9), by selecting a certain one of the standard tableaux, say T_1, and referring all other idempotents to it. Thus we have

$$f_s=\delta_{\sigma s1} * \alpha_1 * \beta_1 * \delta_{\sigma 1s}, \tag{17}$$

or, using the correct diagonal matrix elements,

$$f_s'=(1-\sum_{0<i<s} b(\sigma_{is})\delta_{\sigma i1} + \sum_{0<j<i<s} b(\sigma_{ji})b(\sigma_{is})\delta_{\sigma j1}+\cdots) * \alpha_1 * \beta_1 * \delta_{\sigma 1s}. \tag{18}$$

Formula (17) has been used by Specht[13] and, more recently, by Trainor,[14] the latter to great advantage in certain nuclear problems[15] where formula (17) proves to be a simplification over alternative formulas.

By using Burrow's lemma substantially as a necessary and sufficient condition for the Young idempotents, we have obtained a clearer conception of their use in the theory of the symmetric groups. However, as we have seen, they lead to Young's "natural" representation of these groups, and that is by no means the only possible, nor even useful, representation. For example, it is not necessarily unitary, a requirement imposed by many problems. Yamanouchi,[16] Thrall,[17] and Rutherford[6] have all discussed a direct means of obtaining Young's "orthogonal" representation (since all the classes of S_n are two-sided, all its irreducible representations are equivalent to real representations[18]).

Their discussion is based upon an inductive scheme wherein one assumes that his irreducible representation, when restricted to the subgroup S_{n-1} of permutations leaving the digit n fixed, is reduced according to the irreducible representations of S_{n-1}. This property is of a recursive nature, for it is expected that upon further restriction of the representation to the subgroups S_{n-2}, S_{n-3}, $\cdots$, S_1, which have, respectively, the digits n, $n-1$; n, $n-1$, $n-2$; $\cdots$; n, $n-1$, $\cdots$, 2 fixed, the representation will also be completely reduced according to the irreducible representations of these groups. In other words, one assumes that one deals with a representation of S_n which is completely reduced over the chain of subgroups $S_n \supset S_{n-1} \supset S_{n-2} \supset \cdots \supset S_1$.

When this assumption is made, in Yamanouchi's derivation, one notices that the transposition $(n, n-1)$ commutes with all the elements of the subgroup S_{n-2}, and therefore that Schur's lemma may be invoked to deduce $D(n, n-1)$. This is an inductive scheme, wherein $D(n-1, n-2)$, $D(n-2, n-3)$, $\cdots$ may be determined, and thus the whole representation is determined (in principle) from the knowledge that the set of transpositions $\{(k, k-1)\}$ generates S_n. In practice this set of generators is rather awkward.

APPLICATIONS OF THE SYMMETRICAL GROUPS

In any discussion of the properties of the representations of a group it is worthwhile to consider the applications to which these representations will be put, for they will have a bearing upon the form which it is desirable for the representation to take. In physics there are two principal applications of the general symmetric groups S_n, both quite distinct. One is a consequence of the Pauli exclusion principle, which requires the wave functions of an n-particle system to belong either to the representation α or to the representation β of the group S_n of permutations among the n particles. The other application is to the classification of the resultant angular momentum states arising from the coupling of a number of angular momenta. This latter application is important in the shell model theories of both atomic and nuclear spectroscopy and is no less consequential in other instances.

The importance of the symmetric groups in the theory of angular momentum is founded upon a certain relationship between representations of the full linear groups and the symmetric groups. If one takes the primitive representation of the full linear group of degree n, $GL(n)$ (by which is meant its own faithful representation as the set of all nonsingular $n \times n$ matrices), and forms its kth Kronecker power,

$$D(a) \otimes D(a) \otimes \cdots \otimes D(a),$$

he obtains a reducible representation Γ_k of the full linear group, acting upon a certain vector space $V \times V \times \cdots \times V$. The permutations which exchange the factors in this tensor power leave the Kronecker product unchanged, and thus induce a set of linear transformations in $V \times V \times \cdots \times V$ which commute with the matrices of Γ_k. On the other hand they form a representation Π of the symmetric group S_k. It may be shown[19] that the linear combinations of these matrices, which comprise $((\Pi))$, are the only matrices commuting with Γ_k, and thus with $((\Gamma_k))$. This mutual relation between $((\Pi))$ and $((\Gamma_k))$, that they are one another's commuting algebras, is responsible for the use of the symmetric group in classifying angular momentum states.

The fact that $((\Pi))$ and $((\Gamma_k))$ are one another's commuting algebras means that Γ_k and Π may be brought simultaneously to the form appearing in Fig. 3. We have supposed in reducing Γ_k that the irreducible representations Γ^i of $GL(n)$ may occur with a multi-

[13] W. Specht, Math. Z. **39**, 696 (1935); **42**, 774 (1937).
[14] L. E. H. Trainor, Can. J. Phys. **35**, 555 (1957).
[15] L. E. H. Trainor, Phys. Rev. **85**, 962 (1952); **95**, 801 (1954).
[16] Takehito Yamanouchi, Proc. Phys. Math. Soc. Japan (3) **19**, 436 (1937).
[17] R. M. Thrall, Duke Math. J. **8**, 611 (1941).
[18] G. Frobenius and I. Schur, Sitzber. preuss. Akad. Wiss. Physik. math. Kl. **1906**, 186.
[19] Richard Brauer, Ann. Math. **38**, 857 (1937).

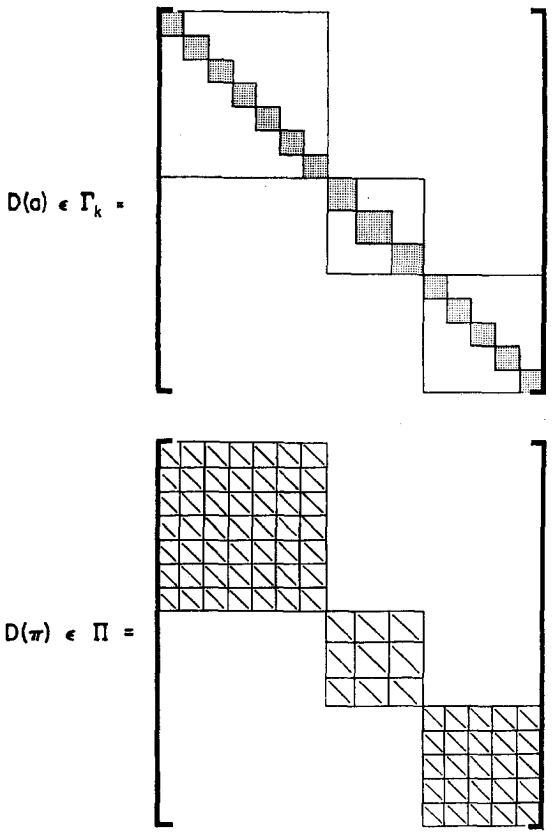

FIG. 3. Γ_k and Π.

plicity $n(i)$ and that Γ has been so reduced that these equivalent representations are equal and occupy contiguous positions along the diagonal. Then, by Schur's lemma, Π consists of matrices having zero matrix elements connecting different representations Γ^i, and multiples of the unit matrix connecting the equal representations. Thus Π is also reduced, and certain irreducible representations of the symmetric group S_n will appear along its diagonal.

A projection operator belonging to a row of Γ^i will have the form of $\gamma_{pp}{}^i$, shown in Fig. 4(a), since this row occurs with a certain multiplicity in Γ_k. A projection operator belonging to a row of Π^j (the jth irreducible representation of S_k), on the other hand, has the form of Fig. 4(b), since the submatrices entering into Π are multiples of the unit matrix. These projection operators commute with one another, and their product, Fig. 4(c), projects onto a single state. Thus we see that for a unique classification of states we require projections belonging both to Π and Γ_k.

This theory remains valid when the full linear group $GL(n)$ is replaced by its unitary subgroup $U(n)$ of the same degree, or even by the unitary unimodular subgroup $SU(n)$. When a representation of the full linear group is restricted to the unitary unimodular subgroup, it remains irreducible, although some distinct representations of $GL(n)$ may become equivalent. An important special case is that of the spin $\frac{1}{2}$ representations of the ordinary three-dimensional rotation group. They comprise the 2×2 unitary unimodular group $SU(2)$, and the theory which we have just outlined is applicable.

Let us consider, as an example, the coupling of six spin $\frac{1}{2}$ particles in such a fashion that their total spin and the z component of the total spin are constants of the motion. Then, by making use of the well known "branching rule" for angular momentum, which states that

$$D^i\otimes D^j=D^{i+j}\oplus F^{i+j-1}\oplus\cdots\oplus D^{|i-j|},$$

where D^j is a representation of dimension $2j+1$ of the rotation group, one can form the diagram[20] shown in Fig. 5, which displays the possible angular momentum states as one particle at a time is added to the system until a total of 6 is reached. Starting with spin $\frac{1}{2}$ for a single particle, we see that we may have $s=\frac{1}{2}+\frac{1}{2}$ or $s=\frac{1}{2}-\frac{1}{2}$ for the combination of two particles, yielding an s state or a p state. At each stage we may add or subtract spin $\frac{1}{2}$ from the previously existing state (except to subtract from the $s=0$ state), and thus we draw as many lines leading up from a point as led to it altogether, and likewise as many leading down, except from $s=0$. For six particles we have a single f state, a

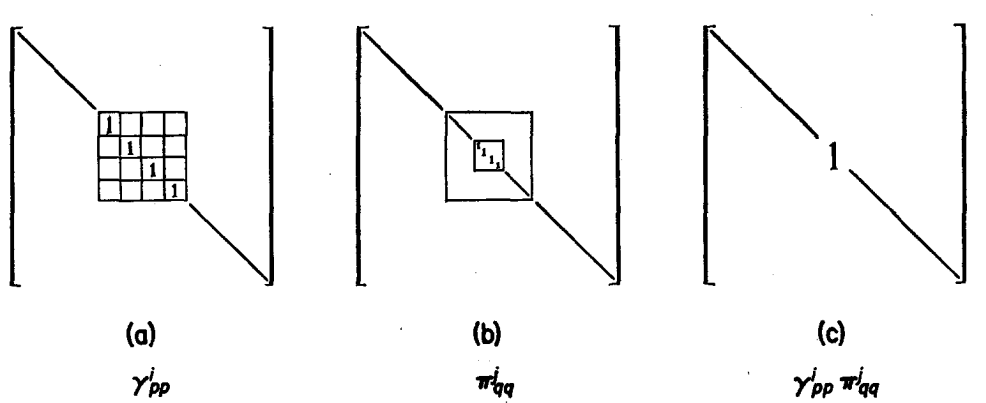

FIG. 4. Projection operators of Γ_k and Π.

[20] See George W. Pratt, Jr., Phys. Rev. 92, 278 (1953).

fivefold degenerate d state, a ninefold degenerate p state, and finally five singlet states.

The reduced form of $\otimes^6 D^{\frac{1}{2}}$ is shown in Fig. 6, together with an indication of the irreducible representation of S_6 which forms the commuting algebra of each degenerate set.

Wigner has shown[21] that not all the irreducible representations of S_k appear in the reduction of Π, and in fact only those occur which belong to Young tableaux having not more than two rows. If the rows have lengths μ and ν, respectively, where $\mu+\nu=k$, then the representation whose tableau is (μ,ν) generates the commuting algebra of the representation $D^{\frac{1}{2}(\mu-\nu)}$ of dimension $(\mu-\nu+1)$ of the rotation group.

Löwdin[2] has obtained a formula which yields projections, onto the states with $m_l=l$, of the product basis functions for a general Kronecker product $\otimes^p D^{\frac{1}{2}}$ of the spin $\frac{1}{2}$ representations. It is a remarkable fact that these projections coincide with the result of pro-

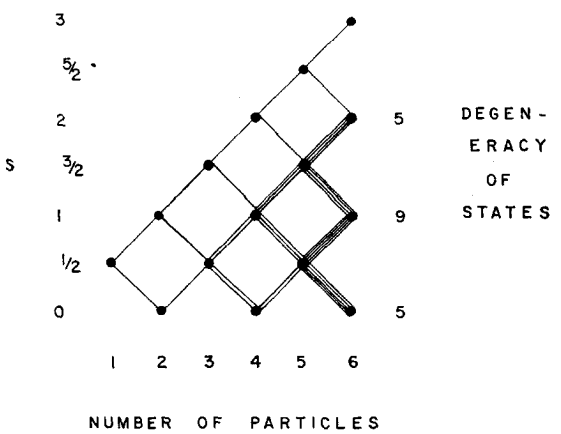

FIG. 5. Possible angular momentum states.

jection by the operator $\alpha * \beta$ belonging to the tableau T_1 of Fig. 7, as is readily seen by comparing his formula

$$\mathcal{O}_{ll}\vartheta_1(\mu,\nu)=\frac{2l+1}{\mu+1}\sum_{p=0}^{p}(-1)^p\binom{\mu}{p}^{-1}[\alpha^{\mu-p}\beta^p][\alpha^p\beta^{\nu-p}] \tag{19}$$

with $\alpha * \beta\vartheta_1(\mu,\nu)$. By $\mathcal{O}_{ll}$ he means the projection operator onto the state of total angular momentum l, with the azimuthal quantum number m_l equal to l. $\vartheta_i(\mu,\nu)$ is one of the product wave functions composed by $\mu\alpha$'s (spin up) and $\nu\beta$'s (spin down) arranged in lexicographic order. The square bracket expression $[\alpha^{\mu-p}\beta^p]$ indicates an average over all the *distinct* permutations of α's and β's in the first μ coordinates, $[\alpha^p\beta^{\nu-p}]$ a similar average over the last ν coordinates.

To calculate $\alpha * \beta\vartheta_1(\mu,\nu)$ we note first that

$$\vartheta_1(\mu,\nu)=(\alpha\alpha\cdots\alpha\beta\beta\cdots\beta),$$

[21] Eugene Wigner, footnote reference 1, Sec. XIII.

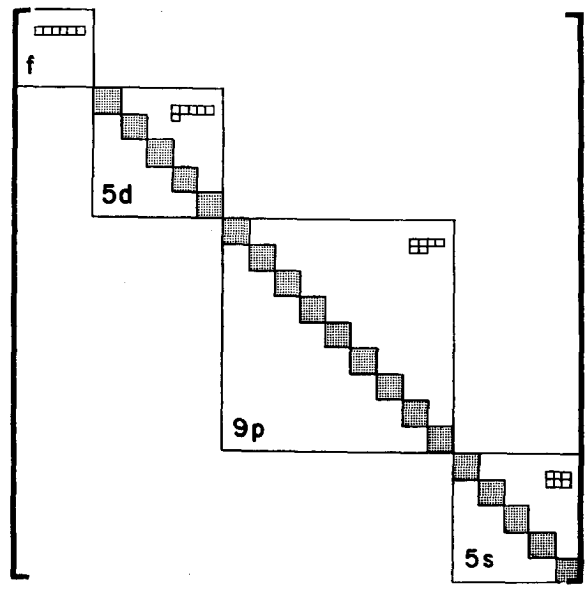

FIG. 6. $\otimes^6 D^{\frac{1}{2}}$ reduced.

with $\mu\alpha$'s and $\nu\beta$'s, and that $\beta\vartheta_1(\mu,\nu)$ is a sum

$$\beta\vartheta_1(\mu,\nu)=(\alpha\alpha\cdots|\beta\beta\cdots)-(\beta\alpha\cdots|\alpha\beta\cdots)$$
$$-(\alpha\beta\cdots|\beta\alpha\cdots)\cdots+(\beta\beta\cdots|\alpha\alpha\cdots)+\cdots+\cdots,$$

writing a vertical bar to separate the first μ from the last ν coordinates. There are $\binom{\nu}{p}$ terms in which $p\alpha$'s have been replaced by β's, and each has the sign $(-1)^p$. When α is applied to this expression, we observe that if one computes the average over all permutations of an expression ϑ containing $\xi\alpha$'s and $\eta\beta$'s, then

$$\alpha\vartheta=\xi!\eta![\vartheta],$$

since on the right-hand side one sums over *distinct* permutations only (or over permutation-values rather than permutations, as it were). Bearing this in mind, and observing that $[\vartheta]$ depends only upon the number of β's that ϑ contains and not upon their order, we find that

$$\alpha * \beta\vartheta_1(\mu,\nu)=\sum_{p=0}^{p}(-1)^p\binom{\nu}{p}(\mu-p)!p!$$
$$\times[\alpha^{\mu-p}\beta^p]p!(\nu-p)![\alpha^p\beta^{\nu-p}]$$
$$=\mu!\nu!\sum_{p=0}^{p}(-1)^p\binom{\mu}{p}^{-1}$$
$$\times[\alpha^{\mu-p}\beta^p][\alpha^p\beta^{\nu-p}], \tag{20}$$

$T_1 =$

1	2	...					...	μ
$\mu+1$	...	...	$\mu+\nu$					

S($\mu+\nu$)

FIG. 7. Standard tableau.

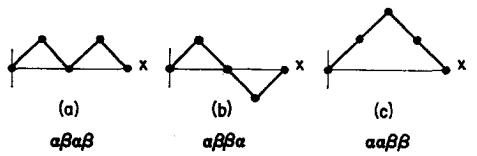

FIG. 8. Path diagrams.

which, except for a numerical factor, is precisely Löwdin's result.

CONCLUSION

This result is of great practical importance, for the angular momentum projection operator is a constant of the motion and therefore commutes with the Hamiltonian. The symmetry projection operators belonging to the symmetric group are in general *not* constants of the motion and do not commute with the Hamiltonian. Thus, by this result, one may use the transition operators of the symmetric group to obtain a complete and nonredundant set of projections, while still being able to write them as angular momentum projections and thus being able to use an operator which commutes with the Hamiltonian.

In fact, the pseudotransition operators $D(\sigma_{ij})$ may be used to generate a complete set of symmetry adapted functions. In the present case we would use σ_{j1}, which is the permutation changing the standard tableau T_1 (Fig. 7) into the standard tableau T_j. Löwdin has described a diagram by means of which one may display the product wave functions $\vartheta_i(\mu,\nu)=(\alpha\alpha\cdots\beta\cdots\alpha\cdots\beta)$. One starts at the origin of a Cartesian plane, and draws a unit segment, inclined at an angle of 45°, from the origin. If the first coordinate ξ_1 is α, the segment inclines upward to the right; if $\xi_1=\beta$, it inclines downward to the right. If $\xi_2=\alpha$, another segment continues upward; downward if $\xi_2=\beta$; and so on. The diagrams in Fig. 8 illustrate $\alpha\beta\alpha\beta$, $\alpha\beta\beta\alpha$, and $\alpha\alpha\beta\beta$.

Now, there is another means of describing a Young tableau; namely, by a Yamanouchi symbol.[22] A Yamanouchi symbol is a sequence of digits $(x_1,x_2\cdots x_n)$, and each digit x_k is simply the number of the row in which k sits in the tableau. Thus the tableau shown in Fig. 9 has the Yamanouchi symbol (2 1 3 2 1 1 1). If a permutation is carried out on the tableau T, the same permutation π is to be carried out on the Yamanouchi symbol Y. That is, x_k becomes $x_{\pi(k)}$.

The Yamanouchi symbols belonging to standard tableaux are characterized by the fact that, reading from left to right, the digit t cannot occur more often than $t-1$, which in turn cannot occur more often then $t-2$, and so on. Several Young tableaux may have the same Yamanouchi symbol, but exactly one standard

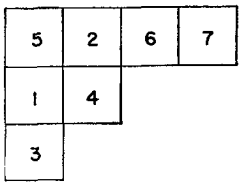

FIG. 9. Young tableau.

tableau corresponds to a given symbol. Specifically, in filling the shape with the digits 1, 2, $\cdots$, n, in their natural order, we always use the leftmost empty box in any row to contain the new digit. This ensures that the digits increase in order across a row, while the fact that the previous rows must be filled at least as far out as the box under consideration means that only smaller digits can occur above any number in a given column.

There is, in the spin $\frac{1}{2}$ case, where the Young tableaux have at most two rows, a one-to-one correspondence between the product wave functions $\vartheta_i(\mu,\nu)$ and the Yamanouchi symbols belonging to the shape (μ,ν), namely, one identifies α with 1, β with 2. If one applies σ_{j1} to $\vartheta_1(\mu,\nu)$, he obtains the new function $\vartheta_j(\mu,\nu)$; meanwhile Y_1 becomes Y_j and T_1 becomes T_j, and these transformations are all consistent, so that $\vartheta_j(\mu,\nu)$ corresponds to Y_j, etc. If T_j is a standard tableau, this means that, in reading $\vartheta_j(\mu,\nu)$ from left to right, more α's must have occurred at any stage than β's, and hence that the diagram for $\vartheta_j(\mu,\nu)$ must lie entirely above the x axis. Conversely, for any line which does lie entirely above the axis there is a standard tableau and hence a permutation σ_{j1} generating it, so that such a path must correspond to a basis function. Thus a class of functions whose projections are linearly independent and nonredundant is composed of just those functions whose path diagrams lie entirely above the x axis.

To summarize the discussion, we have seen how it is possible to factor the projection operators belonging to the symmetric groups, and to give a necessary and sufficient condition for this factorization. We have indicated the fact that alternative forms of the representations are useful in certain problems, and finally we have demonstrated the equivalence of Löwdin's path-diagram method to the formal group-theoretical treatment of the angular momentum states formed from a collection of spin $\frac{1}{2}$ particles.

ACKNOWLEDGMENTS

This paper has been prepared for presentation at the Thanksgiving meeting of the American Physical Society in Chicago, Illinois, while the author has been a guest of the Quantum Chemistry Group at the University of Uppsala, Sweden. It is a pleasure to thank Professor P. O. Löwdin for his hospitality and for stimulating discussions of the subject matter of the paper.

[22] See H. A. Jahn, Proc. Roy. Soc. (London) **A205**, 192 (1951).

On the Calculation of the Inverse of the Overlap Matrix in Cyclic Systems*

P. O. LÖWDIN, R. PAUNCZ,† AND J. DE HEER‡
Quantum Chemistry Group for Research in Atomic, Molecular and Solid-State Theory, Uppsala University, Uppsala, Sweden
(Received June 27, 1960)

The inverse of the overlap matrix of a cyclic system has been computed in three different ways, each of which throws light on a different aspect of the subject and has its own range of applicability to related problems. In all these derivations extensive use has been made of the properties of the Chebyshev polynomials. The results hold for every cyclic symmetric matrix.

INTRODUCTION

IN our work on the quantum mechanics of the infinite linear chain with the Born-von Kármán periodicity condition,[1] we encounter the problem how to obtain the inverse and the inverse half power of a cyclic symmetric matrix. Such a matrix arises if one takes into account the nonorthogonality of the atomic orbitals concerned. The importance of including overlap has already been emphasized in earlier work,[2] thus the solution of this problem is of a more general interest, justifying a separate publication.

In the following we shall give three alternative methods to obtain the inverse, each of which throws light on a different aspect of the subject and has its own range of applicability to related problems. Thus, as we shall show, the last method to be discussed can easily be extended to yield the inverse half-power of the relevant matrix.

I. GENERAL PROBLEM

If we consider n identical atoms arranged symmetrically on the periphery of a circle, the corresponding overlap matrix can be written in the following form:

$$\mathbf{\Delta}=(1,S_1,S_2,\cdots,S_2,S_1)_{\text{cyclic}}, \tag{1}$$

where S_i is the overlap integral between the rth and the $(r+i)$th atomic orbitals ($S_{n+i}=S_i$). The special case $n=6$ (benzene) has already been treated[3] in detail and the following results, equally valid for the general case, were obtained:

(a) The matrix can be transformed to diagonal form by the unitary matrix

$$U_{jl}=n^{-\frac{1}{2}}\exp(2\pi ijl/n) \quad j,l=0,1,\cdots,n-1. \tag{2}$$

(b) The eigenvalues are given by

$$d_l=\sum_{p=0}^{n-1} S_p\exp(2\pi ipl/n). \tag{3}$$

(c) The elements of the matrix corresponding to any function of $\mathbf{\Delta}$ are the following:

$$[F(\mathbf{\Delta})]_{\mu\nu}=\frac{1}{n}\sum_{l=0}^{n-1} F(d_l)\exp[2\pi i(\mu-\nu)l/n]. \tag{4}$$

In the case of the inverse, we wish to avoid the evaluation of these sums, and rather give an explicit expression for its elements. We shall first treat the case when only S_1 is retained and the other S_i are put equal to zero, and afterwards return to the general problem.

II. RESTRICTION TO NEAREST-NEIGHBOR OVERLAP

In this case $\mathbf{\Delta}$ can be written as

$$\mathbf{\Delta}=\mathbf{1}+S_1\mathbf{M}_1, \tag{5}$$

where $\mathbf{M}_1$ is the topological matrix discussed extensively by Ruedenberg.[4]

$$M_{1,pq}=\begin{cases}1 & \text{if } p \text{ and } q \text{ are neighbors}\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

The eigenvalues of $\mathbf{\Delta}$ lie within the range

$$1-2S_1\leqslant d_l\leqslant 1+2S_1 \quad \text{for } n=2\nu, \tag{7a}$$

and

$$1-2S_1\cos\pi/(2\nu+1)\leqslant d_l\leqslant 1+2S_1 \quad \text{for } n=2\nu+1. \tag{7b}$$

The overlap matrix corresponding to any set of linearly independent orbitals can never have a zero or negative eigenvalue. Thus S_1 must be smaller than 0.5[4] (note that for $n=2\nu$ and $S_1=0.5$, the inverse does not exist at all). We shall therefore restrict our treatment to

* Supported in part by the King Gustaf VI Adolf's 70-Years Fund for Swedish Culture, Knut and Alice Wallenberg's Foundation, The Swedish Natural Science Research Council, and in part by the Aeronautical Research Laboratory, Wright Air Development Division of the Air Research and Development Command, U. S. Air Force, through its European Office, under a contract with Uppsala University.

† On leave from the Chemistry Department, Technion-Israel Institute of Technology, Haifa, Israel.

‡ This work was performed under a fellowship grant from the John Simon Guggenheim Foundation. Permanent adress: Chemistry Department, University of Colorado, Boulder, Colorado.

[1] R. Pauncz, J. de Heer, and P. O. Löwdin (to be published).

[2] P. O. Löwdin, J. Chem. Phys. **18**, 365 (1950); **19**, 1579 (1951).

[3] P. O. Löwdin, J. Chem. Phys. **21**, 496 (1953).

[4] K. Ruedenberg, J. Chem. Phys. (to be published); the authors are indebted to Dr. Ruedenberg for making preprints available to them.

this case. Physically, the breakdown for $S_1>0.5$ is understandable since with such high overlap it becomes utterly inconsistent to neglect all but nearest-neighbor interactions.

A. Direct Method

Divide $\mathbf{\Delta}$ by S_1 and introduce the notation

$$\mathbf{N}(a)=(a,1,0,0,\cdots,0,1)_{\text{cyclic}},$$

with $a=1/S_1$. Then

$$\mathbf{\Delta}^{-1}=a\mathbf{N}^{-1}(a),$$

where $a>2$. The determinant and the cofactors of $\mathbf{N}(a)$ can be related to a special type of determinant, "continuant" [5]:

$$P_k(a)=\begin{vmatrix} a & 1 & & & & \\ 1 & a & 1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & 1 & a & 1 \\ & & & & 1 & a \end{vmatrix}$$

They fulfill the recursion relations

$$P_k=aP_{k-1}-P_{k-2}, \tag{9}$$

with $P_1=a$ and $P_2=a^2-1$. These determinants have been frequently used in connection with the simple MO-LCAO treatment of linear polyene chains, and it is well known that the Chebyshev polynomials of the second kind fulfill the same relations (9):

$$\mathcal{S}_k(x)=x\mathcal{S}_{k-1}-\mathcal{S}_{k-2}, \tag{10}$$

with $x=2\cos\vartheta$, $\mathcal{S}_0=1$, and $\mathcal{S}_1=x$. Since in our case $a>2$, we must use an imaginary argument, hence put $\vartheta=i\xi$, or $a=2\cosh\xi$ (see Appendix). The $P_k(a)$ are then given by

$$P_k(a)=\sinh[(k+1)\xi]/\sinh\xi=\mathcal{S}_k(a).$$

By the Cauchy expansion of the determinant $|\mathbf{N}(a)|$ according to its first row and column, we easily obtain

$$|\mathbf{N}_n(a)|\equiv N_n(a)=P_n-P_{n-2}-2(-1)^n =2[\cosh n\xi-(-1)^n]. \tag{11}$$

Since both $\mathbf{N}$ and $\mathbf{N}^{-1}$ are cyclic and symmetric, it is sufficient to evaluate the cofactors pertaining to the elements of the first row of $\mathbf{N}$. On denoting by N_{ik} the determinant with the ith row and the kth column stricken out, we have

$$N_{00}=P_{n-1}(a)=\sinh(n\xi)/\sinh\xi, \tag{12a}$$

$$N_{0\mu}=(-1)^\mu[P_{n-1-\mu}+(-1)^nP_{\mu-1}] =(-1)^\mu[\sinh(n-\mu)\xi+(-1)^n\sinh\mu\xi]/\sinh\xi. \tag{12b}$$

[5] See, e.g., T. H. Goodwin and V. Vand, J. Chem. Soc. **1955**, 1683.

By combining (12a) and (12b) with (11), we obtain the expression for the element of the inverse:

$$[\mathbf{N}^{-1}(a)]_{\mu 0}=(-1)^\mu\frac{\sinh(n-\mu)\xi+(-1)^n\sinh\mu\xi}{2\sinh\xi[\cosh n\xi-(-1)^n]}, \tag{13}$$

where $\mu\leqslant n/2$. The use of the cyclic and symmetric properties of the matrix yields all the remaining elements.

By a simple manipulation we can write the foregoing result in the following form:

$$[\mathbf{N}^{-1}(a)]_{\mu 0}=\frac{(-1)^\mu}{2\sinh\xi}\left\{\cosh\mu\xi-\sinh\mu\xi +\left[\frac{\sinh n\xi}{\cosh n\xi-(-1)^n}-1\right]\cosh\mu\xi\right\}. \tag{14}$$

If n becomes large and $\mu\ll n/2$, the preceding expression yields the following asymptotic formula:

$$[\mathbf{N}^{-1}(a)]_{\mu 0}\simeq(-1)^\mu(e^{-\mu\xi}/2\sinh\xi). \tag{15}$$

Hence

$$\mathbf{\Delta}_{\mu 0}{}^{-1}\simeq(-1)^\mu\coth\xi e^{-\mu\xi}, \tag{16}$$

with $S_1=1/2\cosh\xi$. This is a very simple expression, which shows that the elements of the inverse in the first column (row) decrease very rapidly. The asymptotic formula becomes worse as μ approaches $n/2$; for $\mu=n/2$ the exact expression is twice as great as the asymptotic one. Owing to the rapid decrease of the elements with increasing μ, the absolute error becomes practically negligible.

B. Chebyshev Polynomials of the Topological Matrix

In the second derivation, we shall use the Chebyshev polynomials of the first kind $\mathcal{C}_n$ defined as

$$\mathcal{C}_{n+1}=x\mathcal{C}_n-\mathcal{C}_{n-1},\quad \mathcal{C}_0=2 \text{ and } \mathcal{C}_1=x. \tag{17}$$

Next let us introduce the following matrices which are cyclic of order n:

$$(\mathbf{m})_1=(0,1,0,\cdots 0),\quad (\mathbf{m})_2=(0,0,1,\cdots 0), \tag{18}$$

hence with the general elements:

$$[\mathbf{m}_p]_{kl}=\delta_{k+p,l}. \tag{19}$$

Then we can easily prove that

$$\mathbf{m}_p\mathbf{m}_q=\mathbf{m}_{p+q}. \tag{20}$$

Since $\mathbf{m}_0$ and $\mathbf{m}_n$ are unit matrices, we also have

$$\mathbf{m}_p\mathbf{m}_{n-p}=\mathbf{1}, \tag{21}$$

$$\mathbf{m}_{-p}\equiv\mathbf{m}_{n-p}=\mathbf{m}_p{}^{-1}. \tag{22}$$

On the other hand

$$\mathbf{m}_{n-p}=\mathbf{m}_p^{\dagger},$$

so that the $\mathbf{m}$ are unitary matrices.[6]

Finally we define

$$\mathbf{M}_p=\mathbf{m}_p+\mathbf{m}_{-p}, \quad \mathbf{M}_n=\mathbf{M}_0=2\mathbf{1}. \tag{23}$$

$\mathbf{M}_1$ is the topological matrix, $\mathbf{M}_2$ the analogous matrix for next-nearest interaction, and so on. We now have the following relations between the $\mathbf{M}$ and $\mathbf{m}$:

$$\mathbf{M}_\nu=2\mathbf{m}_\nu, \quad \text{if } n=2\nu, \tag{24}$$

$$\mathbf{M}_p\mathbf{M}_q=\mathbf{M}_{p+q}+\mathbf{M}_{p-q}, \tag{25}$$

$$\mathbf{M}_{n-p}\equiv\mathbf{M}_{-p}=\mathbf{m}_{-p}+\mathbf{m}_p=\mathbf{M}_p. \tag{26}$$

The $\mathbf{M}_p$ are cyclic and symmetric matrices; they all commute and can be brought to diagonal form by the same unitary transformation.

We can easily see that

$$\mathbf{M}_p=\mathfrak{C}_p(\mathbf{M}_1), \tag{27}$$

since this relation holds for $p=0$ and 1 and the $\mathbf{M}_p$ satisfy the same recursion formula as the $\mathfrak{C}_p$:

$$\mathbf{M}_0=2\mathbf{1}=\mathfrak{C}_0(\mathbf{M}_1), \quad \mathbf{M}_1=\mathfrak{C}_1(\mathbf{M}_1),$$
$$\mathbf{M}_{p+1}=\mathbf{M}_1\mathbf{M}_p-\mathbf{M}_{p-1}.$$

By using this result, we can obtain a number of characteristic equations, the first of which reads, for example,

$$\mathbf{M}_{n-1}=\mathfrak{C}_{n-1}(\mathbf{M}_1)=\mathbf{M}_1, \tag{28}$$

or

$$\mathfrak{C}_{n-1}(\mathbf{M}_1)-\mathbf{M}_1=0,$$

which is an equation of degree $(n-1)$. The equation of the lowest degree which is not trivial reads

$$\mathbf{M}_{\nu+1}=\mathbf{M}_{\nu-1} \quad \text{or} \quad \mathfrak{C}_{\nu+1}(\mathbf{M}_1)-\mathfrak{C}_{\nu-1}(\mathbf{M}_1)=0 \quad \text{for } n=2\nu, \tag{29}$$

$$\mathbf{M}_{\nu+1}=\mathbf{M}_\nu \quad \text{or} \quad \mathfrak{C}_{\nu+1}(\mathbf{M}_1)-\mathfrak{C}_\nu(\mathbf{M}_1)=0 \quad \text{for } n=2\nu+1. \tag{30}$$

Introduce $F_r(x)$ defined by

$$F_r(x)=\mathfrak{C}_{r+1}(x)-\mathfrak{C}_{r-1}(x); \tag{31}$$

then we have the identity

$$\frac{F_r(x)-F_r(a)}{x-a}=\sum_{k=0}^{r}\epsilon_k\mathfrak{C}_k(a)\mathfrak{C}_{r-k}(x), \tag{32}$$

with $\epsilon_0=\epsilon_r=\frac{1}{2}$ all other $\epsilon_k=1$. The identity is easily verified if we multiply both sides by $(x-a)$ and use the recursion formulas

$$x\mathfrak{C}_{r-k}(x)=\mathfrak{C}_{r+1-k}(x)+\mathfrak{C}_{r-1-k}(x),$$
$$a\mathfrak{C}_k(a)=\mathfrak{C}_{k+1}(a)+\mathfrak{C}_{k-1}(a).$$

[6] $\mathbf{m}_p^{\dagger}$ denotes the transpose of $\mathbf{m}_p$.

We must now discuss separately the case of even and odd n.

(a) $n=2\nu$

By using the definition (31), the characteristic equation (29) can be written

$$F_\nu(\mathbf{M}_1)=0. \tag{33}$$

By adding and subtracting $F_\nu(-a)$ and dividing by $[\mathbf{M}_1-(-a)]$, we have the following equation for the inverse:

$$[\mathbf{M}_1+a\mathbf{1}]^{-1}=-\frac{1}{F_\nu(-a)}\frac{F_\nu(\mathbf{M}_1)-F_\nu(-a)}{\mathbf{M}_1-(-a)\mathbf{1}}$$
$$=-\frac{1}{F_\nu(-a)}\sum_{k=0}^{\nu}\epsilon_k\mathfrak{C}_k(-a)\mathfrak{C}_{\nu-k}(\mathbf{M}_1).$$

But $\mathfrak{C}_k(-a)=(-1)^k\mathfrak{C}_k(a)$ and $F_\nu(-a)=(-1)^{\nu+1}F_\nu(a)$; hence the final result reads

$$[\mathbf{M}_1+a\mathbf{1}]^{-1}=\frac{1}{F_\nu(a)}\sum_{k=0}^{\nu}\epsilon_k(-1)^{\nu-k}\mathfrak{C}_k(a)\mathbf{M}_{\nu-k}$$
$$=\frac{1}{F_\nu(a)}\sum_{k=0}^{\nu}\epsilon_k(-1)^k\mathfrak{C}_{\nu-k}(a)\mathbf{M}_k. \tag{34}$$

(b) $n=2\nu+1$

Now two characteristic equations will be used:

$$\begin{aligned}&\mathbf{M}_{\nu+1}=\mathbf{M}_\nu, \quad &&\mathfrak{C}_{\nu+1}(\mathbf{M}_1)-\mathfrak{C}_\nu(\mathbf{M}_1)=0,\\ &\mathbf{M}_{\nu+2}=\mathbf{M}_{\nu-1}, \quad &&\mathfrak{C}_{\nu+2}(\mathbf{M}_1)-\mathfrak{C}_{\nu-1}(\mathbf{M}_1)=0.\end{aligned} \tag{35}$$

The addition of the two equations yields, using (31),

$$F_{\nu+1}(\mathbf{M}_1)+F_\nu(\mathbf{M}_1)=0. \tag{36}$$

It is convenient to introduce the shorthand notations

$$G_\nu=F_{\nu+1}+F_\nu, \quad \bar{G}_\nu=F_{\nu+1}-F_\nu, \tag{37}$$

$$\mathfrak{D}_\nu=\mathfrak{C}_{\nu+1}+\mathfrak{C}_\nu, \quad \bar{\mathfrak{D}}_\nu=\mathfrak{C}_{\nu+1}-\mathfrak{C}_\nu, \tag{38}$$

so that (35) and (36) read

$$G_\nu(\mathbf{M}_1)=0 \quad \bar{\mathfrak{D}}_\nu(\mathbf{M}_1)=0. \tag{39}$$

Then, using (32), the following identity can be established:

$$\frac{G_\nu(x)-G_\nu(a)}{x-a}=\tfrac{1}{2}[\mathfrak{C}_0(a)\bar{\mathfrak{D}}_\nu(x)+\mathfrak{D}_\nu(a)\mathfrak{C}_0(x)]+\sum_{k=0}^{\nu-1}\mathfrak{D}_k(a)\mathfrak{C}_{\nu-k}(x). \tag{40}$$

In order to get the expression for the inverse, we proceed in the same way as in the case of an even number of

atoms:

$$[\mathbf{M}_1+a\mathbf{1}]^{-1}=-\frac{1}{G_\nu(-a)}\frac{G_\nu(\mathbf{M}_1)-G_\nu(-a)}{\mathbf{M}_1-(-a)\mathbf{1}}$$

$$=-\frac{1}{G_\nu(-a)}\{\tfrac{1}{2}[\mathfrak{C}_0(-a)\overline{\mathfrak{D}}_\nu(\mathbf{M}_1)+\mathfrak{D}_\nu(-a)\mathfrak{C}_0(\mathbf{M}_1)]+\sum_{k=0}^{\nu-1}\mathfrak{D}_k(-a)\mathfrak{C}_{\nu-k}(\mathbf{M}_1)\}. \quad (41)$$

From (37) and (38) we obtain

$$\begin{aligned}G_\nu(-a)&=(-1)^\nu\bar{G}_\nu(a),\\ \mathfrak{D}_k(-a)&=(-1)^{k+1}\overline{\mathfrak{D}}_k(a);\end{aligned} \quad (42)$$

hence we obtain finally

$$[\mathbf{M}_1+a\mathbf{1}]^{-1}=\frac{1}{\bar{G}_\nu(a)}[\overline{\mathfrak{D}}_\nu(a)\mathbf{1}+\sum_{k=1}^{\nu}(-1)^k\overline{\mathfrak{D}}_{\nu-k}\mathbf{M}_k]. \quad (43)$$

In order to compare the results obtained in Sec. II.A and II.B we write Eq. (13) in the form

$$\mathbf{N}(a)^{-1}=[\mathbf{M}_1+a\mathbf{1}]^{-1}=\sum_{k=0}^{\nu}(-1)^k\eta_k\frac{\mathcal{S}_{n-1-k}(a)+(-1)^n\mathcal{S}_{k-1}(a)}{\mathcal{S}_n(a)-\mathcal{S}_{n-2}(a)-2(-1)^n}\mathbf{M}_k, \quad (44)$$

where for $n=2\nu$ we must take $\eta_0=\eta_\nu=\frac{1}{2}$, i.e., the η_k are identical with the ϵ_k introduced in Eq. (32); however, for $n=2\nu+1$, it is implied that $\eta_0=\frac{1}{2}$ and all other $\eta_k=1$. We can easily see that Eq. (44) is indeed obtained if we multiply numerator and denominator in (34) by $\mathcal{S}_{\nu-1}(a)$ (for $n=2\nu$), and those in (43) by $\mathcal{S}_\nu(a)-\mathcal{S}_{\nu-1}(a)$ $(n=2\nu+1)$. In both cases we also have to use the relation

$$\mathcal{S}_k\mathcal{C}_l=\mathcal{S}_{k+l}+\mathcal{S}_{k-l}. \quad (45)$$

C. Power Series Method

The third derivation gives at once the asymptotic formula in a very useful form. This method can easily be modified to yield other powers of $\mathbf{\Delta}$, for example $\mathbf{\Delta}^{-\frac{1}{2}}$.

Let us take a power series expansion of an analytic function

$$f(z)=\sum_{k=0}^{\infty}\alpha_k z^k;$$

the series is convergent if $|z|<\rho$.

Put $z=re^{i\vartheta}$ and $x=2\cos\vartheta$,

$$z=r\{(x/2)+i[1-(x^2/4)]^{\frac{1}{2}}\};$$

then

$$z^k=r^ke^{ik\vartheta}=r^k\{\tfrac{1}{2}\mathcal{C}_k(x)+i[1-(x^2/4)]^{\frac{1}{2}}\mathcal{S}_{k-1}(x)\}, \quad (46)$$

and

$$f(z)=\tfrac{1}{2}\sum_{k=0}^{\infty}\alpha_k r^k\mathcal{C}_k(x)+i[1-(x^2/4)]^{\frac{1}{2}}\sum_{k=1}^{\infty}\alpha_k r^k\mathcal{S}_{k-1}(x). \quad (47)$$

Hence we obtain two very useful relationships:

$$\sum_{k=0}^{\infty}\alpha_k r^k\mathcal{C}_k(x)=2\,\mathrm{Re}\,f\left\{r\left[\frac{x}{2}+i\left(1-\frac{x^2}{4}\right)^{\frac{1}{2}}\right]\right\}, \quad (48a)$$

$$\sum_{k=0}^{\infty}\alpha_{k+1}r^k\mathcal{S}_k(x)=\{1/r[1-(x^2/4)]^{\frac{1}{2}}\}\times\mathrm{Im}f\left\{r\left[\frac{x}{2}+i\left(1-\frac{x^2}{4}\right)^{\frac{1}{2}}\right]\right\}, \quad (48b)$$

where $r<\rho$, $|x|<2$. We shall use the following power series expansion:

$$f(z)=1/(1+z)=1-z+z^2-\cdots, \quad \alpha_k=(-1)^k, \quad \rho=1.$$

The real and imaginary parts are easily found:

$$\frac{1}{1+z}=\frac{1+r(x/2)}{1+rx+r^2}-i\frac{r[1-(x^2/4)]^{\frac{1}{2}}}{1+rx+r^2}.$$

Hence

$$\sum_{k=0}^{\infty}(-1)^k r^k\mathcal{C}_k(x)=\frac{2+rx}{1+rx+r^2}=1+\frac{(1-r^2)/r}{[(1+r^2)/r]+x}, \quad (49a)$$

and

$$\sum_{k=0}^{\infty}(-1)^k r^k\mathcal{S}_k(x)=-1/(1+rx+r^2), \quad |x|<2. \quad (49b)$$

In order to relate Eq. (49a) to the present problem, put

$$\begin{aligned}(1+r^2)/r&=a=1/S_1;\\ r=(a/2)\pm[(a^2/4)-1]^{\frac{1}{2}}&=[1\pm(1-4S_1^2)^{\frac{1}{2}}]/2S_1.\end{aligned} \quad (50)$$

The conditions $S_1<\frac{1}{2}$ and $|r|\leqslant 1$ allow only the minus sign. By a simple manipulation we can see that

$$(1-r^2)/r=(1-4S_1^2)^{\frac{1}{2}}/S_1.$$

Inserting (50) into Eq. (49a) we have

$$\frac{1}{a+x}=\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}[1-r\mathcal{C}_1(x)+r^2\mathcal{C}_2(x)-\cdots]. \quad (51)$$

Hence

$$[\mathbf{M}_1+a\mathbf{1}]^{-1}=\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}(\mathbf{1}-r\mathbf{M}_1+r^2\mathbf{M}_2-\cdots)=\sum_{k=0}^{\infty}c_k\mathbf{M}_k, \quad (52)$$

where

$$r=[1-(1-4S_1^2)^{\frac{1}{2}}]/2S_1=2S_1/[1+(1-4S_1^2)^{\frac{1}{2}}], \quad |r|<1, \quad S_1<\tfrac{1}{2}. \tag{53}$$

For large n it is sufficient to consider only the first $n/2$ terms in the expansion, which yields immediately the asymptotic form for the coefficient of $\mathbf{M}_k$ as

$$c_k=\eta_k(-1)^k\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}\{2S_1/[1+(1-4S_1^2)^{\frac{1}{2}}]\}^k, \tag{54}$$

with $\eta_0=\frac{1}{2}$ all other $\eta_k=1$. We must note, however, that this formula becomes worse as k approaches $n/2$.

Next we use the present method in order to derive the exact coefficients. By taking into account the property of the $\mathbf{M}_k$ that

$$\mathbf{M}_k=\mathbf{M}_{ln-k}=\mathbf{M}_{ln+k}, \tag{55}$$

where n is the order of the overlap matrix, we can convert the infinite sum (52) into a finite one:

$$\sum_{k=0}^{\infty}c_k\mathbf{M}_k=\sum_{k=0}^{\nu}b_k\mathbf{M}_k. \tag{56}$$

On the basis of (55) we obtain the following relations between the c_k and b_k:

$$b_k=c_k+c_{n-k}+c_{n+k}+c_{2n-k}+\cdots, \quad k\neq 0, \tag{57}$$

and

$$b_0=c_0+c_n+c_{2n}+\cdots, \tag{58a}$$

while for even $n(=2\nu)$ only

$$b_\nu=c_\nu+c_{3\nu}+c_{5\nu}+\cdots. \tag{58b}$$

Again we have to discuss the cases of even and odd n separately.

(a) $n=2\nu$

From Eqs. (52), (57), and (58), we obtain

$$b_0=\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}(\tfrac{1}{2}+r^n+r^{2n}+\cdots)=\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}\left(\frac{1}{1-r^n}-\frac{1}{2}\right), \tag{59}$$

$$b_\nu=\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}\frac{(-r)^\nu}{1-r^n}, \tag{60}$$

$$b_k=(-1)^k\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}[r^k+r^{n+k}(1+r^n+\cdots)+r^{n-k}(1+r^n+\cdots)]=(-1)^k\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}\frac{r^k+r^{n-k}}{1-r^n}. \tag{61}$$

(b) $n=2\nu+1$

On the basis of (57) and (58), we have

$$b_0=\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}(\tfrac{1}{2}-r^n+r^{2n}-\cdots)=\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}\left(\frac{1}{1+r^n}-\frac{1}{2}\right), \tag{62}$$

$$b_k=(-1)^k\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}[r^k-(1-r^n+r^{2n}-\cdots)\times(r^{n+k}+r^{n-k})]=(-1)^k\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}\frac{r^k-r^{n-k}}{1+r^n}. \tag{63}$$

In order to compare the results of the present section with those obtained by the direct method, we put $r=e^{-\xi}$. Then by virtue of Eq. (53) we readily establish that

$$2\cosh\xi=r+r^{-1}=\frac{2S_1}{1+(1-4S_1^2)^{\frac{1}{2}}}+\frac{2S_1}{1-(1-4S_1^2)^{\frac{1}{2}}}=\frac{1}{S_1}=a \tag{64}$$

in accordance with the earlier definition. Obviously

$$\frac{S_1}{(1-4S_1^2)^{\frac{1}{2}}}=\frac{1}{(a^2-4)^{\frac{1}{2}}}=\frac{1}{2}\sinh\xi. \tag{65}$$

Turning our attention first to the asymptotic formula (54), it can now be transformed to the following form:

$$c_k=\eta_k(-1)^k(e^{-k\xi}/2\sinh\xi), \tag{66}$$

which clearly shows its equivalence with (15). Next it is easy to show that the results expressed in Eqs. (59)–(63) are equivalent to that contained in Eq. (13) if we multiply numerator and denominator of the former by $[1-(-1)^n r^{-n}]$.

III. INVERSE OF $\mathbf{\Delta}=1+S_1\mathbf{M}_1+S_2\mathbf{M}_2$

On returning to the general $\mathbf{\Delta}$ of Eq. (1), we can obtain an explicit expression for its inverse using the results obtained for the restricted case of nearest neighbors only. For Eq. (27) shows that any $\mathbf{M}_k$ can always be written as a polynomial in $\mathbf{M}_1$ of degree k. Hence we can easily carry out the following factorization:

$$\mathbf{\Delta}=\sum_{k=0}^{\nu}\epsilon_k S_k\mathbf{M}_k=\text{const}\times\prod_{k=1}^{\nu}(a_k 1+\mathbf{M}_1), \tag{67}$$

where in general the a_k (no longer equal to S_k^{-1}) are complex numbers. We can easily generalize the derivations given in Sec. II.A to the case when a is complex, and the form of the elements of the inverse remains unaltered. The only difference is that now the argument is a complex number (see Appendix). The factors $(a_k\mathbf{1}+\mathbf{M}_1)$ commute, so that the inverse of $\mathbf{\Delta}$ is found by multiplying together the inverses of the individual factors.

We shall illustrate the method in the case of

$$\mathbf{\Delta}=\mathbf{1}+S_1\mathbf{M}_1+S_2\mathbf{M}_2, \tag{68}$$

for a system with an even number of atoms, $n(=2\nu)$. This matrix can be written in the factorized form

$$\mathbf{\Delta}=\mathbf{1}+S_1\mathbf{M}_1+S_2(\mathbf{M}_1^2-2\mathbf{1}) =S_2(a_1\mathbf{1}+\mathbf{M}_1)(a_2\mathbf{1}+\mathbf{M}_2), \tag{69}$$

with

$$a_1a_2=(1-2S_2)/S_2 \quad \text{and} \quad a_1+a_2=S_1/S_2. \tag{70}$$

As mentioned previously, a_1 and a_2 are usually conjugate complex numbers:

$$a_1=a_2^*=2\cos\eta=2(\cos u\cosh v-i\sin u\sinh v), \tag{71}$$

where u and v are real. They can be determined on the basis of Eqs. (70) and (71):

$$\cos u=\frac{1}{2(S_2)^{\frac{1}{2}}}[(1+2S_1+2S_2)^{\frac{1}{2}}-(1-2S_1+2S_2)^{\frac{1}{2}}],$$
$$\cosh v=\frac{1}{2(S_2)^{\frac{1}{2}}}[(1+2S_1+2S_2)^{\frac{1}{2}}+(1-2S_1+2S_2)^{\frac{1}{2}}]. \tag{72}$$

On multiplying together the inverses of the individual factors and performing the relevant summations, we obtain the following result:

$$\begin{aligned}[\mathbf{1}+S_1\mathbf{M}_1+S_2\mathbf{M}_2]_{0k}^{-1} &=\{(-1)^k/8S_2(\cosh^2 v-\cos^2 u)(\cosh^2\nu v-\cos^2\nu u)\}\\ &\times\{\cot u[\sin ku\cosh(2\nu-k)v\\ &+\sin(2\nu-k)u\cosh kv]+\coth v[\sinh kv\\ &\times\cos(2\nu-k)u+\sinh(2\nu-k)v\cos ku]\}.\end{aligned} \tag{73}$$

The same method can be used if we wish to take into account third, fourth, and so on, neighbor interactions, but in view of the exponential decrease of the overlap integral with distance, it is hardly worthwhile to include many more terms.

IV. INVERSE HALF-POWER OF $\mathbf{\Delta}=\mathbf{1}+S_1\mathbf{M}_1$

In the discussion of a number of physical problems the inverse half-power of the overlap matrix plays an important part.[2]

Since $\mathbf{\Delta}$ is a cyclic and symmetric matrix, $\mathbf{\Delta}^{-\frac{1}{2}}$ must also have those properties. Thus it can be written in the following form:

$$\mathbf{\Delta}^{-\frac{1}{2}}=\sum_{k=0}^{\nu}\gamma_k\mathbf{M}_k. \tag{74}$$

The jth eigenvalue can be given in two alternative ways:

$$d_j^{-\frac{1}{2}}=1/(1+2S_1\cos j\alpha)^{\frac{1}{2}}=\sum_{k=0}^{\nu}\gamma_k 2\cos kj\alpha, \tag{75}$$

with $\alpha=(2\pi)/n$. We determine the γ_k in expanding the lhs in terms of $\cos lj\alpha$ and comparing the coefficients with those occuring in the rhs. For this expansion we use the following relation[7]:

$$\frac{1}{\sqrt{2}}(\cosh\xi+\cos j\alpha)^{-\frac{1}{2}}=A_0+\sum_{1}^{\infty}(-1)^l A_l 2\cos lj\alpha \tag{76}$$

with

$$A_l=\frac{e^{-(l+\frac{1}{2})\xi}}{l!}\frac{\Gamma(l+\frac{1}{2})}{\Gamma(\frac{1}{2})}F(\tfrac{1}{2},l+\tfrac{1}{2};l+1,e^{-2\xi}), \tag{77}$$

where F is the hypergeometric series and $1/S_1=a=2\cosh\xi$ as before.

The result can be rewritten in the form

$$(1+2S_1\cos j\alpha)^{-\frac{1}{2}}=\sum_{l=0}^{\infty}2\delta_l\cos lj\alpha=\sum_{k=0}^{\nu}\gamma_k 2\cos kj\alpha, \tag{78}$$

with $\delta_0=(S_1)^{\frac{1}{2}}A_0/2$ and $\delta_l=(-1)^l A_l(S_1)^{\frac{1}{2}}$.

The relation between the δ_l and the γ_k can be established in the same way as in Sec. II.C, taking into account the fact that

$$\cos(rn\pm l)j\alpha=\cos lj\alpha. \tag{79}$$

In spite of the fact that we cannot give such a simple expression as in the case of the inverse, the use of (74)–(78) is rather more advantageous than that of the formulas given in the earlier derivation.[3] The coefficients given in the present series expansion can be computed easily since it converges very fast for $S<\frac{1}{2}$, and the so computed A_l themselves decrease very rapidly with increasing l. So for greater n there is even no need to use the periodicity relations (79); the corrections from the higher periods are completely negligible. For the sake of illustration we give the first few A_l's for $S=\frac{1}{4}$:

l:	0	1	2	3	4	5
A_l:	1.05465	0.28522	0.02848	0.01286	0.00151	0.00073
	6	7	8			
A_l:	0.00009	0.00004	0.00000_5.			

[7] W. Magnus and F. Oberhättinger, *Formeln und Sätze für die speziellen Funktionen der matematischen Physik* (Springer Verlag, Berlin, 1948), p. 214.

APPENDIX

In this Appendix we shall show that Chebyshev polynomials defined in terms of a complex argument satisfy the same basic equations as those commonly defined in terms of a real argument.

Let

$$z=2\cos\eta \quad \text{with} \quad \eta=u+iv \tag{A1}$$

and put

$$\mathcal{C}_n(z)=2\cos n\eta, \tag{A2}$$

$$\mathcal{S}_n(z)=\sin[(n+1)\eta]/\sin\eta. \tag{A3}$$

As the relations

$$\sin(n+1)\eta+\sin(n-1)\eta=2\cos\eta\sin n\eta, \tag{A4}$$

$$\cos(n+1)\eta+\cos(n-1)\eta=2\cos\eta\cos n\eta, \tag{A5}$$

still hold for the complex argument, it is seen that $\mathcal{C}_n(z)$ and $\mathcal{S}_n(z)$ fulfill the same recurrence relations as given in Eqs. (10) and (17) for real argument.

In particular, if $\eta=i\xi$, $x=2\cosh\xi$ and

$$\mathcal{C}_n(x)=2\cosh n\xi, \tag{A6}$$

$$\mathcal{S}_n(x)=\sinh[(n+1)\xi]/\sinh\xi. \tag{A7}$$

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 1, NUMBER 6 NOVEMBER-DECEMBER, 1960

Quantization of Nonlinear Systems

I. E. SEGAL*
Department of Mathematics, University of Chicago, Chicago, Illinois†
(Received April 25, 1960)

A direct method of quantization, applicable to a given nonlinear hyperbolic partial differential equation, is indicated. From such classical equations alone, without a given Lagrangian or Hamiltonian, or *a priori* linear reference system such as a bare or incoming field, a quantized field is constructed, satisfying the conventional commutation relations. While mathematically quite heuristic in part, local products of quantized fields do not intervene, and there are grounds for the belief that the formulation is free from nontrivial divergences.

1. INTRODUCTION

THERE has been interest recently in the development of purely nonlinear quantum field theories, i.e., theories which are formulated without the use of such physically somewhat dubious and mathematically linear notions as those of a "free" field or of an "elementary" particle.[1] Despite the promise of this work, the difficulties are such that there has sometimes been lacking a satisfying demonstration of the internal consistency of the formal structure on which the theory is based, or, on occasion, a reasonably clear-cut physical interpretation of the formalism. The purpose of the present work is to indicate a method of quantization that seems on the whole somewhat less subject to these defects. The main result is a new framework for covariant quantum field theory, which appears to be convergent, although mathematically quite heuristic. From any of a fairly wide class of given manifolds of "classical" wave functions, there is constructed an associated quantum field, as well as a possible means of determining theoretically vacuum expectation values of functions of field operators, and aspects of a formal elementary particle interpretation. In particular, the work provides some basis for a renewal of the traditional intuitive belief—which has been strongly tempered by the persistence of divergences during the past 30 years—that for any simple covariant coupling of the conventional elementary particles of relativistic quantum field theory, there should be a corresponding quantum field theory of their interaction; but at the same time casts further doubt on the rigorous relevance to such theories of the notion of elementary and/or physical (dressed) particle, as well as the possibility of expressing such a theory in terms of an *a priori* type of incoming field.

Besides the perturbation-theoretic divergences of quantum field theory, and its use of an *a priori* linear reference space, there is another feature that is rather unsatisfactory from a foundational viewpoint. This is the dependence of the theory on a notion, the product of local fields [e.g., $\phi(x)\psi(x)\psi(x)^*$ in conventional notation] which seems inevitably remote from any physical measurement. As is clear from a line of work originating with the well-known classical paper of Bohr and Rosenfeld, a suitably smoothed average $\int\phi(x)f(x)d_4x$ ($f=a$ "test" function, corresponding to a probe into the field) is the most that one can hope to measure even in principle. However, no way has been found to express such a product as $\phi(x)\psi(x)\psi(x)^*$, or averages of it, in terms of such smoothed averages of individual fields; and quite apart from the divergences which such products directly lead to, it is odd that a notion so lacking in direct physical meaning (as well as in rigorous mathematical significance, so that it rests purely on traditional formalism and metaphysics) should play the essential role in the construction of the field dynamics. The attempt to bypass this kind of difficulty by a purely axiomatic approach as in the work of Haag, Källén and Wightman, Lehmann *et al.*, and some others, has clarified the logical situation, but on the whole the results are still rather inconclusive, and certain of the axioms are rather strong from a physical standpoint. A more constructive (in the technical sense) line of attack is given by Segal,[2] the essential presently relevant idea being the use not of the $\phi(x)\psi(x)\psi(x)^*$ themselves, but only of integrals of the type $H=\int_{t=t'}\phi(x)\psi(x)\psi(x)^*d_3x$, involving only commuting (and so more amenable) fields; and the use of these not as operators, but as generators of motions of the dynamical variables of the field. That is, roughly speaking, only the $[H,X]$ need be finite for any field observable X, and not H itself, which leads to a mathematically quite well-defined category of objects H materially broader than the class of self-adjoint operators in a Hilbert space. Although some definite results in quantum electrodynamics, of a rigorous character, can be obtained in

* Research supported in part by the Air Force OSR, and conducted in part at the University of Copenhagen while an NSF fellow.

† Present address: Massachusetts Institute of Technology, Cambridge, Massachusetts.

[1] See notably W. Heisenberg, Revs. Modern Phys. **29**, 269 (1957), and S. Deser, *ibid.* **29**, 417 (1950) (and other articles, in addition to the last-named, reporting the Chapel Hill Conference on Gravitation); and especially articles by Heisenberg and Yukawa, Proc. Internatl. Conf. High-Energy Nuclear Phys., Geneva, 1958.

[2] I. E. Segal, Kgl. Danske Videnskab. Selskab Mat.-fys. Medd. **31**, No. 12, 1–39 (1959).

this way (cf. footnote reference 3), this work has the important limitation that the physical vacuum is not constructed, and there are very substantial, if relatively well-understood difficulties in showing that $H[=H(t)]$ has the required properties for the rigorous existence of the time-ordered (product) integral $\exp[i\int H(t)dt]$ that defines the transformation taking the in- into the out-field observables. It is hard to believe that a definitive foundational treatment of a system whose dynamics are conceptually as simple as those of quantum electrodynamics, say, must depend on the resolution of the intricate and special problems that arise here.

At any rate, it seems reassuring to have a general scheme for setting up quantum field interactions in which the singular products of the type $\phi(x)\psi(x)\psi(x)^*$ play no role whatever. The formalism involves only quantities which are in principle capable of being related to direct physical measurement. It rests mathematically on the combination of certain simple if abstract ideas from the theory of differentiable manifolds with the intrinsic (representation-independent) theory of operator systems applied to field theory in,[2] and the development of mathematical analysis in function space. On the other hand, many of the relevant mathematical developments are presently available only in highly rudimentary form (e.g., it is not yet proved that the relevant *classical* partial differential equations have any nontrivial solutions—even of a generalized character—in the large), and we merely assume their eventual existence on the basis of plausibility considerations. Also, the particle interpretation of the present scheme refers in the first instance essentially to the "primary" particles, and it seems doubtful whether a precise "empirical" particle interpretation will exist with any generality, in view of the applicability of the scheme to both renormalizable and nonrenormalizable field equations, and to fields involving bound states and unstable particles. In particular, when a theory of the present sort is specialized, say, to quantum electrodynamics, it gives no *a priori* labeling of the states of the incoming field in terms of finite aggregates of "free physical" electrons and photons. Whether or not such labels can be rigorously established—as is a well-defined mathematical question according to the present framework, along with the question of the existence and character of bound states and unstable particles—it is difficult to make specific computations of real empirical effects without them or some approximate equivalent. Without these various developments there is no assurance either that the theory can be made mathematically irreproachable or can be accurately correlated with the crucial experimental results pertinent to field theory. It is only from a theoretical physical point of view and relative to the present state of quantum field theory that the present work appears to represent a contribution of possible significance. There have after all been extremely few truly unambiguous theoretical developments in the subject since it was set up by Dirac, Heisenberg, and Pauli, despite the large number of fragmentary contributions that have been made. It seems that for foundational purposes only a quite comprehensive attack employing conservative but global methods has much hope of ultimate success. As this has never really precisely been undertaken, there is no reason for undue pessimism, but the scope of such a development is necessarily such that it is unrealistic to begin highly explicit analytical computations until the fundamental design is well established. It is to the settlement of this design question—of what is actually a quantum field theory—that this article is intended to contribute.

The present theory is related to linear quantum field theory (or the theory of noninteracting fields) in roughly the same way that the theory of differentiable manifolds is related to the theory of linear vector spaces—the interaction has its source in the nonlinear structure of the manifold representing the classical states of the system being quantized. The conventional type of theory of interacting fields (which may be called quasi-linear) is related to the present theory in somewhat the way that the theory of a Riemannian manifold as described by normal coordinates at a distinguished point is related to the intrinsic theory of the manifold. The vectors in the tangent space to the manifold at that point represent the bare particles of the theory, which would make the extrinsic theory convenient for giving a particle interpretation, if the apparent need for infinite mass and charge renormalization did not make it impossible then to give *ab initio* in the theory the precise relation between the manifold and the tangent space. The extrinsic theory is also disadvantageous from a theoretical point of view in its use of *ad hoc* assumptions as to the structure of the incoming field, which make the role of bound states and unstable particles in the theory highly elusive. For example, in the case of quantum electrodynamics it is conventionally *assumed* implicitly that the incoming field is describable by the Fock representation, with a renormalized tangent space as single-particle space; in general, such an assumption overdetermines the theoretical structure of a quantum field, and may well lead to internal inconsistencies.

In its simplest form the nonintrinsic character of conventional theory is exemplified by the *ad hoc* separation of the total Hamiltonian into "free-field" and "interaction" parts, a separation that is required for the usual analytical treatment of scattering. The kinematics of the interacting field is derived from the free-field part and is linear, while the dynamics is superimposed on the kinematics through the statement of the interaction Hamiltonian or Lagrangian (or more operationally, of the S operator). In the present work, no Hamiltonian or Lagrangian (or S operator) needs

[3] I. E. Segal, Ann. Math. (to be published).

to be specified; the theory is built up entirely from the classical equations of motion. The distinction between the free-field and total Hamiltonians is seen to be essentially that between the linear motion in the tangent plane at a fixed point in a manifold under a group of transformations that is induced in the natural manner by the group action, and the nonlinear motion that is obtained by transferring, through the use, e.g., of normal coordinates, the given group to the tangent plane. In general relativity there is no distinguished invariant point of the manifold of solutions of the equations that is physically analogous to the point defined by vanishing fields in the case of elementary particle theory, and hence no covariant separation of the motion into two parts, but the theory may still be quantized by the intrinsic approach.

From the viewpoint of general analytical dynamics, a theory of the present type is determined primarily by the specification of a differentiable manifold B representing the classical phase space of the system under consideration, together with a second-order Hermitian differential form Ω on B, and a corresponding notion of multiplication by complex scalars in the tangent spaces of B. In classical mechanics the fundamental bilinear covariant is quite analogous to Ω, but complex scalars in the tangent spaces of phase space have apparently not been used. In the case of a field, where B is infinite-dimensional, Ω is better known in the form of the singular functions $D(x,x')$ that arise as field commutators in the quantization of a linear equation. The tangent space at any point ϕ of B is parametrized by functions $f,g,\cdots$ on space-time satisfying the first-order variations of the coupled field equations in the infinitesimal vicinity of ϕ (taking the case of a scalar field for simplicity), while Ω is determined by its imaginary part Ω_i, which is by definition a rule that assigns to a point ϕ a bilinear form in the tangent vectors at ϕ, and is given by the equation

$$\Omega_i(f,g;\phi)=\int\int f(x)g(x')D_\phi(x,x')d_4xd_4x'.$$

In conventional theory only the singular functions of covariant free fields seem to have been used, and in this case $D(x,x')$ depends only on $x-x'$, but here the singular functions defined by similar Cauchy data for all first-order variations of the coupled field equations are relevant, and $D(x,x')$ will have the usual type of dependence only in the special case $\phi=0$ (or other constant solution, if any, of the equation defining the manifold). This canonical construction for Ω_i in the case of a manifold in function space defined by a nonlinear hyperbolic partial differential equation has been explored in certain cases and in a rather different form by Peierls.[4] The fundamental symmetry group G of the theory may be any group of transformations on B that leaves Ω invariant, as does, e.g., the Lorentz group in the case of a manifold defined by a Lorentz-invariant equation of the foregoing type. A complete set of primary quantum numbers of the usual type—of group-theoretic origin—will exist if, and only if, the induced action of G in the tangent space at $\phi=0$ is such that the linear operators in the tangent space that are left invariant commute with one another (or equivalently, the irreducible constituents of this representation are all distinct). When the generators of G are suitably labeled as "energy," "angular momentum," etc. (the Lorentz group is by no means the only one for which this is possible; cf., e.g., Segal[5]) the resulting physical theory, in particular the formal S operator, is in essence completely determined.

The idea of constructing a purely nonlinear quantum field theory has been developed in recent years most extensively by Heisenberg (see footnote reference 1 where further references are given), with whose standpoint the theory described in the foregoing is in general harmony. While it thereby lends some support to Heisenberg's idea that a purely nonlinear theory should be convergent, its specific form deviates in some important respects from that suggested by Heisenberg's program, notably in the significant role played in it by singular functions associated with linear partial differential equations. It has been a cardinal principle of Heisenberg to avoid the use of such functions, with the aim of eliminating the divergences of conventional theory, which arise from the *a priori* meaningless character of their products. The latter are involved in computations based on perturbation theory, as well as, in essence, in the formulation of conventional dynamics. As indicated in the foregoing, in the present work no such products arise, so that the use of these singular functions introduces no divergences. But this does not in itself indicate that a fully satisfactory theory may be based on the Lorentz group and conventional space-time, for divergences may well be introduced by the use of *ad hoc* labels for the states of the incoming field. Such desiderata as the observation of stable single-particle states of sharply-defined mass may well ultimately require the introduction of a fundamental length into the structure of B and/or lead to the replacement of the Lorentz group by another to which it is a partial approximation in the sense considered in footnote reference 5. We may also note a rather obvious rough analogy between the role of the infinite-dimensional tangent spaces to nonlinear function manifolds in the present work and those of the finite-dimensional tangent spaces of the space-time manifold in general relativity. Partial parallels with important ideas of Feynman concerning the use of functional integration, of Dirac dealing with covariance questions, and of Wiener concerning nonlinear analysis, will also be evident to the knowledgeable reader.

[4] R. E. Peierls, Proc. Roy. Soc. (London) **A214**, 143 (1952).

[5] I. E. Segal, *Report of Lille Conference on Quantum Fields* (C.N.R.S., Paris, 1959), pp. 57–103.

Briefly, we show how a generalized canonical variable $R(X)$ can be associated with an infinitesimal generator X of the group of classical contact transformations on the given classical system. The construction of these variables depends inessentially on the choice of a first-order differential form whose covariant differential gives Ω_i, and which is analogous to the form $\sum_k p_k dq_k - Hdt$ employed in classical mechanics. In the case of a linear manifold, the $R(X)$ associated with infinitesimal translations X in phase space give the conventional (Heisenberg) commutation relations, while the present generalized ones satisfy the rule

$$[R(X),R(Y)] = -iR([X,Y]) + \Omega(X,Y),$$

where $[X,Y]$ denotes the usual bracket of two vector fields. When Ω vanishes (and only then), R gives a representation of the infinitesimal contact group, which is in fact in the finite-dimensional case the well-known one introduced by Koopman and studied by him and von Neumann, in connection with classical mechanics. It may also be noted that in not being concerned with an actual representation of the contact group, as well as in a number of other respects, the specialization of the present approach to the case of a finite-dimensional linear manifold differs from the noteworthy investigations of van Hove[6] directed toward a basis for a rational correspondence between classical and quantum mechanical Hamiltonians.

The avoidance of convergence difficulties depends in part on the elimination of any *ad hoc* Hilbert space in the foundation of the theory for the representation of the states of the incoming field, such a space being, however, convenient for correlation with experiment and also used, implicitly or explicitly, in most of the recent literature on quantum fields in a rigorous direction (cf. the authors already cited). Rather, the incoming field becomes one of the objects whose structure the theory is to determine. The method is roughly to work with the system of all bounded functions of *finite* sets of the canonical variables, together with their limits in the sense of *uniform* convergence, as in footnote reference 2; this gives a covariant class of observables that is representation-independent, in contrast to the set of bounded observables obtained by using functions of infinite sets of canonical variables and/or limits in the sense of so-called strong or weak convergence. States are defined through their expectation value functionals on the foregoing system, which is both more physical, and mathematically more effective than their *a priori* representation by vectors in a Hilbert space. Yet ultimately a Hilbert space can be constructed which represents the states of the incoming field, by the use of the physical vacuum expectation values, which are in turn connected with a process resembling integration over the classical manifold B (such integration is made fully rigorous in the case of infinite-dimensional linear manifolds by Segal,[7a] and a formal adaptation of this work to the relevant nonlinear manifolds will be indicated later).

In more analytical terms, the main ideas of the present work may be indicated in their simplest form as follows. The manifold B of all real solutions of a given Lorentz-invariant hyperbolic nonlinear partial differential equation in four-dimensionsl space-time carries a distinguished Hermitian structure. Quantization involves in essence analysis over this manifold (in contrast to classical mechanics, which is concerned with the construction of the manifold and the action of various groups on it), i.e., the study of certain operators (in particular the values of the "quantum field") in spaces of functionals over the manifold. Canonical variables may be attached to the vector fields on the manifold through the use of a differential form of first order related to the given Hermitian structure. The field itself arises from the projection of the variational derivative $\partial/\partial f$ in function space onto the (sub-) manifold B; taking f as a delta function at a point yields formally the field at the point. The quantum-theoretical physical vacuum is represented by the unit function on B, the vacuum state being characterized by invariance under the group of isometries of B leaving invariant the vanishing field $\phi=0$. The primary elementary particle species of the theory are given by the irreducibly invariant subspaces of the tangent space to B at the point $\phi=0$ (or other Lorentz-invariant point of B, if any), and formally the theory may be expressed entirely in terms of this tangent space, which corresponds essentially to the most conventional procedure. The conventional free fields (those given by the quantizations of the Klein-Gordon, Maxwell, etc. equations in empty space) correspond precisely to the special case in which the manifold B is a complex Hilbert space and the Lorentz group action is unitary, the Hermitian structure being that given by the fundamental inner product, and the physical vacuum as characterized before being unique and the familiar one associated essentially with an isotropic normal distribution in Hilbert space.

In considerable part, the foregoing description is valid only for Bose-Einstein fields. While it appears that the Fermi-Dirac fields can probably be treated in a rather analogous way, it will presumably be necessary to replace vector by spinor fields (over function manifolds), and the notion of integration by that treated in the linear case in footnote reference 7[b], etc. In view of the substantial character of such modifications, the present paper is confined to the Bose-Einstein case.

In Sec. 2, the nonrelativistic quantum mechanics of a finite number of degrees of freedom is extended to the

[6] L. Van Hove, Acad. roy. Belg. Classe sci. Mém. Collection in 8° **29**, No. 6, 1–102 (1951).

[7] (a) I. E. Segal, Trans. Am. Math. Soc. **81**, 106 (1956); (b) I. E. Segal, Ann. Math. **63**, 160 (1956).

case when space is not necessarily flat; this involves in particular the reformulation of the conventional quantum conditions so as to be covariant under general point transformations in physical space. Section 3 completes the preliminaries by showing how canonical variables and commutation rules may be set up in terms of phase, rather than physical, space, in a form covariant under contact transformations. The hydrogen atom and harmonic oscillator problems in the presence of nonvanishing curvature are briefly discussed in these sections. In Sec. 4, the earlier developments are combined with methods previously developed in connection with certain aspects of infinite systems to obtain a quantization scheme for a class of infinite nonlinear systems, represented by the case of a classical system defined by a nonlinear hyperbolic partial differential equation. The concluding Sec. 5 discusses the present results in relation to some of the existing literature and possible further developments.

2. FINITE SYSTEMS AND POINT TRANSFORMATIONS

Consider a quantum-mechanical system whose position is described by a point of an n-dimensional manifold M (and so is a system of $2n$ degrees of freedom). If as in conventional theory M is a linear manifold, one proceeds by introducing Hermitian operators $p_1,p_2,\cdots,p_n$ and $q_1,q_2,\cdots,q_n$ satisfying the Heisenberg commutation relations. If however, M is nonlinear, it is unclear *a priori* to what extent it is possible to proceed in a suitably parallel way. In the case of a sphere or torus, results can be obtained by making use of the simple natural parametrizations available for these manifolds. However, physically the availability of a suitable special parametrization appears as a rather technical restriction on M; intuitively it would appear possible to quantize a classical system whose position is represented by a point of a relatively arbitrary manifold.

To develop an appropriate quantization method, we note that the canonical P's are naturally associated with vector fields on M, and the canonical Q's with position coordinates; the Schrödinger representation for the linear case associates P_j with $\partial/i\partial x_j$, and Q_j with the coordinate x_j. To handle the nonlinear case we merely allow the P's to be associated with arbitrary vector fields—i.e., linear forms in the $\partial/i\partial x_j$, with variable rather than constant coefficients (which are undefined in the absence of distinguished coordinates or related special features of M)—and the Q's with arbitrary functions on M, and not merely linear functions (which are likewise undefined on a general manifold). The commutation relations are virtually automatically generalized thereby; any commutator of canonical variables is required simply to be that associated with the commutator of the corresponding transformations on the functions over M.

To make this approach mathematically effective, it is necessary to formulate the P's and Q's as well-defined operators in a Hilbert space. To set up an appropriate Hilbert space, take a measure m on the given manifold M that has a continuous nonvanishing density at every point[8]; in general there will be no distinguished measure analogous to the Euclidean volume element used in conventional theory, but we proceed, tentatively, with an arbitrary measure of the foregoing type; it will develope that actually the theory is independent of the choice of measure.

The Hilbert space $\mathfrak{H}$ is then defined as consisting of all square-integrable functions f on M (the values of f being complex numbers) with the inner product

$$(f,g)=\int f(x)g(x)^*dm(x).$$

Now if T is a general vector field on M, the associated canonical momentum $P(T)$ might be provisionally defined as the operator in $\mathfrak{H}$ taking f into $(\hbar/i)Tf$; this is appropriate from a formal algebraic viewpoint, but it gives rise to difficulties originating in the non-Hermitian character of T as an operator in $\mathfrak{H}$. With the modified definition

$$P(T)=(\hbar/2i)(T-T^+),$$

the fundamental commutation relations are unchanged, and $P(T)$ is now manifestly Hermitian. A simple computation shows that the foregoing definition works out concretely as

$$P(T)=(\hbar/i)(T+K_T),$$

where K_T denotes the operation of multiplication by the function k_T, which is defined by the equation

$$2k_T(x)=Tw+l(T),$$

where m has the element $w\prod dx_j$ (locally), and $l(T)$ is defined as $\sum_j(\partial a_j/\partial x_j)$ for T of the form $\sum_j a_j(\partial/\partial x_j)$.

A straightforward computation that is here omitted yields the commutation relation

$$[P(S),P(T)]=(\hbar/i)P([S,T]). \tag{1}$$

The symbol $[S,T]$ denotes the commutator of the two vector fields S and T in the usual sense of the theory of manifolds. In the case of a linear manifold this vanishes for two infinitesimal translations, and (1) specializes merely to the commutativity of the conventional linear momenta. For an infinitesimal trans-

[8] For convenience, it is assumed, as seems no essential loss of generality from a physical standpoint, that the manifold M is infinitely differentiable, i.e., that it is possible near each point to choose local coordinates in such a manner that whenever a point is assigned two sets of coordinates, then near the point the one set may be expressed as infinitely differentiable functions of the other set. It is known (virtually as a matter of definition) that the existence of a measure with a nowhere vanishing continuous density function is mathematically equivalent to the orientability of M, which will be assumed in the present section.

lation and an infinitesimal rotation, however, (1) gives the conventional commutation relations between linear and angular momenta.

The canonical Q's are defined more simply: if f is a general function on M, $Q(f)$ is defined as the operator in H taking h into fh (i.e., the operation of multiplication by f). For real f, $Q(f)$ is Hermitian, and there is no difficulty in verifying the additional commutation relations

$$[P(T),Q(f)]=(\hbar/i)Q(Tf), \tag{2}$$

$$[Q(f),Q(g)]=0. \tag{3}$$

The first of these relations includes the conventional commutation relations between an angular momentum and a coordinate, as well as the basic relations between a linear momentum and a coordinate given explicitly in the Heisenberg form. The second merely asserts the commutativity of all the Q's.

The foregoing construction is essentially merely an adaptation of the Schrödinger representation to an arbitrary manifold, together with a reformulation that makes manifest the invariance of the scheme under arbitrary coordinate transformations. When M is three-dimensional Euclidean space, as in the conventional theory of a single particle, the basic canonical variables are taken as the $P(T)$ with T restricted to be a first-order linear differential operator with constant coefficients, and as the correspondingly restricted $Q(f)$ (i.e., f, a linear function on the space). Since such $P(T)$ and $Q(f)$ however already suffice to give an irreducible set of operators on $L_2(M)$, the additional $P(T)$ and $Q(f)$ defined earlier are already observables in the conventional scheme, so that the phenomenological structure of the theory—the observables, states, and notions defined in terms of these—is unaltered by the present reformulation. The broadened definitions of the $P(T)$ and $Q(f)$ merely amount to a labelling of certain of the observables, which facilitates a general treatment of kinematics, in which transformations that do not have constant coefficients are treated on the same footing as those that do.

Thus, as far as phenomenology and kinematics are concerned, the present formalism is quite equivalent to the conventional one of nonrelativistic quantum mechanics, for the case of a system of finitely many particles in three-dimensional Euclidean space. Since our ultimate aim is to treat systems whose dynamics is implicit in their kinematics, that is all that is primarily relevant. Nevertheless, it is of interest to consider how the application of the correspondence principle to the determination of the quantum dynamics is affected. This may also serve to clarify and make more concrete the development just described.

Conventionally, the quantum-theoretic Hamiltonian is derived from the classical one by a familiar, although generally somewhat ambiguous, process of substituting variables satisfying the Heisenberg relations for the commuting classical canonical variables. From the present standpoint, this means that a special frame (or class of frames) of reference is used in the manifold that will have no analog on a general manifold. The substitution method thus appears as less applicable in the case of a nonlinear manifold, but there is another effective method of implementing the correspondence principle, notably that of matching the invariance and other formal features of the classical Hamiltonian.

Consider for example the problem of the hydrogen atom in an arbitrary Riemannian manifold. The relevant classical Hamiltonian is (or, strictly speaking, is defined as) the sum of the kinetic energy with the Coulomb potential (the latter being defined in general as proportional to the elementary solution for the Laplace equation for the manifold). There is no need to describe the use of normal coordinates, etc., in obtaining a precise analog for the conventional classical kinetic energy, for the Laplacian gives immediately an operator that satisfies the key desiderata of generalizing the kinetic energy in conventional nonrelativistic quantum mechanics and of being intrinsically defined in terms of the Riemannian geometry. It is clear that any finite number of particles with Coulomb interactions may be similarly treated.

This example may be not without some realistic relevance. The validity of three-dimensional Euclidean space as a model for macroscopic space at the nonrelativistic level is open to direct verification, but that the same model is valid in dealing with microscopic space (i.e., that in which it is theoretically appropriate to consider an electron as imbedded, if indeed such exists) is quite a different postulate, which can only be verified experimentally by indirect means, such as through its implications for atomic spectra (cf., e.g., Schrödinger[9]). In particular, in the event that with increasing precision of measurement discrepancies from present theory are found in the spectrum of hydrogen, it might well be of interest to compare them with the first-order perturbations in the spectrum arising from a nonvanishing constant curvature, a problem which seems technically quite accessible.

The correspondence principle as just applied does not have rigorous mathematical character, but is based partly on the exercise of judgement as to what is physically appropriate and mathematically natural. In involving possible ambiguity, the present form of the correspondence principle does not, however, differ from the conventional process, in which the assignment of the order of factors in a product of canonical operators is generally quite essentially nonunique. There have been many efforts toward the solution of this uniqueness problem (see notably footnote reference 6, which is definitive in certain respects), but no completely satisfactory mathematical process has yet been presented. Thus the application of the correspondence

[9] E. Schrödinger, Naturwissenschaften **22**, 518 (1934).

principle within the present formalism appears to be fundamentally not more difficult than its application by means of the conventional formalism, in the case of a linear manifold. Actually, in the following a unique method is given for passing from a *covariant* classical motion to a quantum-mechanical one in line with the present approach, but nontrivial applications are limited to systems of infinitely many degrees of freedom.

It remains to consider the dependence of the foregoing quantization scheme on the choice of a measure m on M. In case another measure m' is used, operators $P'(T)$ and $Q'(f)$ in another Hilbert space $\mathfrak{H}'=L_2(M,m')$ are obtained. But the transformation

$$V:f \to f(dm/dm')^{\frac{1}{2}}$$

is unitary from $\mathfrak{H}$ onto $\mathfrak{H}'$, and it is straightforward to verify that

$$VP(T)V^{-1}=P(T), \quad VQ(f)V^{-1}=Q'(f).$$

Thus the two systems of canonical variables are unitarily equivalent, and in fact the equivalence is implemented by the relatively trivial transformation V.

The following paragraphs of this section concern questions of rigor, and some readers may prefer to omit them.

In the foregoing work a certain loophole for irrelevant mathematical pathology has been left open through the use of the unbounded P's and Q's, which operate not on all of $\mathfrak{H}$, but on certain dense domains in $\mathfrak{H}$ (this domain varying from operator to operator), and what is more serious, cannot be unambiguously multiplied and added together freely. The well-known device of Weyl for eliminating pathological canonical systems and making in a natural fashion the P's and Q's mathematically more clear-cut in the case of a linear manifold can however be adapted to general manifolds. It consists in the replacement of the consideration of the P's and Q's in the foundations of the theory by the consideration of the one-parameter unitary groups they generate. Actually it is convenient to modify this device and consider in place of the one-parameter groups generated by the Q's the smooth bounded functions of them, for this merely amounts to using only those $Q(f)$ for which f is such a function. In this way one is led to make the following definition reminiscent of that for a representation of a group of transformations given by G. W. Mackey.

Definition 1. A generalized Heisenberg canonical system over a finite-dimensional infinitely differentiable manifold M is a pair of maps $[U,Q]$, which are respectively from the group G of all nonsingular infinitely differentiable transformations in M and the class $\mathfrak{R}$ of all real bounded infinitely differentiable functions on M that vanish at infinity, to the bounded operators in a Hilbert space $\mathfrak{K}$, such that:

(1) U is a unitary representation of G: $U(gg')=U(g)U(g')$, $U(e)=I$ (e=unit of G, I=identity operator on $\mathfrak{K}$), $U(g)^{-1}=U(g)^*$; and is continuous on finite-dimensional subgroups of G.

(2) Q is an isomorphism: $Q(f+f')=Q(f)+Q(f')$, $Q(lf)=lQ(f)$, $Q(ff')=Q(f)Q(f')$, and $Q(f)\neq 0$ if $f\neq 0$.

(3) $U(g)Q(f)U(g)^{-1}=Q(f_g)$, where $f_g(x)=f(g^{-1}(x))$ (this essentially gives in finite form the commutation relations between a P and a Q).

(4) The $Q(f)$ generate a maximal commuting subsystem of the total system of operators generated by the $U(g)$ and $Q(f)$.

In the case when M is a finite-dimensional Euclidean space, the only such system, within physical equivalence (or observables and states) is that in which $\mathfrak{H}=L_2(M)$, $U(g)h(x)=h[g^{-1}(x)]$, and $Q(f)h=fh$. But if M is not a simply connected manifold, there will be unitarily inequivalent Heisenberg systems.[10] Nevertheless there is always a fully covariant way to specify the representation that is relevant here, i.e., to make Definition 2.

Definition 2. A generalized Schrödinger canonical system over a finite-dimensional infinitely differentiable orientable manifold M is the pair of maps $[U,Q]$ from G and $\mathfrak{R}$ described earlier, to operators on $L_2(M,m)$, where m is an arbitrary measure on M with infinitely differentiable nonvanishing density function, given by the equations

$$U(a)h(x)=h(a^{-1}(x))(dm_a/dm)^{\frac{1}{2}},$$
$$Q(f)h=fh.$$

Here a is an arbitrary element in G, and m_a denotes the transform of m under the transformation of measures induced by the transformation a on M.

As noted earlier, all the Schrödinger systems are unitarily equivalent, and no essential ambiguity will

[10] The number of inequivalent such is an invariant of M closely related to its one-dimensional cohomology in the following way: if ω is any closed first-order differential form on M, then the equations $P'(X)=P(X)+\omega(X)$, $Q'(f)=Q(f)$, define a Heisenberg system $[P',Q']$ (in infinitesimal terms) which will be equivalent to the system $[P,Q]$ if ω is exact, but not generally otherwise. Specifically, there is equivalence if, and only if, ω is logarithmically exact, in the sense that $\omega=dF/F$ for some function F on M. It follows from a study of the logarithmically exact forms (cf. a forthcoming paper by R. S. Palais; similar but less complete and unpublished results are due to E. Dyer and R. Swan) that on a manifold with first Betti number r, there is an r-parameter family of inequivalent Heisenberg systems.

Mathematically it is interesting to weaken statement (4) by requiring only (4'), ergodicity: no nontrivial function of the P's and Q's commutes with all the P's and Q's. The analog of the Schrödinger representation with square-integrable functions replaced by square-integrable tensor fields is an example of a system satisfying (4') but not (4). The foregoing connection with closed differential forms and cohomology can be extended, but some of the quantum-mechanical invariants of M obtained in the indicated fashion may be new, depending in part on the extent to which the tensor field examples exhaust the possibilities, within unitary equivalence and the intervention of a closed form. This is a point having a certain differential-geometric interest, and conceivably there is a physical role for the tensor, etc. representations in other physical connections, but in the present paper only the "scalar" Heisenberg representations given by Definition 1 are used.

result if any one of these systems is referred to as *the* Schrödinger system. Thus we may summarize the foregoing section as:

Principle I. There exists a unique and mathematically precise scheme for setting up quantization conditions on an arbitrary orientable finite-dimensional manifold M; this extends the conventional scheme for the case of three-dimensional Euclidean space, and is covariant under arbitrary transformations on the manifold. In essence, the generalized canonical variables are represented by Hermitian first-order linear differential operators on M, relative to an arbitrary measure on M.

3. FINITE SYSTEMS AND CONTACT TRANSFORMATIONS

Let us now consider the method of the preceding section in relation to the problem of the quantization of an infinite nonlinear system. At a nonrelativistic level the problem is that of developing a parallel to Dirac's extension to a radiation field of Heisenberg's original quantization procedure. A field whose state at a particular time is represented classically by a solution of a certain nonlinear partial differential equation, rather than by a linear equation as in the case treated by Dirac, has its nonlinear canonical Q's associated with functionals on the manifold M of all classical such solutions of the equation, while the canonical nonlinear P's are associated with vector fields on M, as indicated in Sec. 2. If the field equation is first order and hyperbolic in the weak sense that the values of the solutions at a particular time $t=t_0$ determine the solutions throughout space-time, and if the initial values form a linear vector space (assumptions which in essence are frequently made), this set of initial values may be taken as the manifold M, and the nonlinearity enters primarily in the nonlinear action of displacement in time on M. The adaptation of Sec. 2 to this case requires a notion of integral in M, and the development of its transformation properties under nonlinear transformations of M, of the type presented by Gross,[11] as well as, for a rigorous treatment of certain divergent cases, an as yet unavailable combination of the transformation theory of footnote reference 11 with the representation-free approach of footnote reference 2. Basically however—in particular as regards the formulation of the quantized field itself—the development of this nonrelativistic theory is closely related to that of the covariant theory that is our central concern, and to which we shall therefore restrict our further consideration.

The quantization of a nonlinear covariant system involves new formal elements roughly analogous to those involved in the Heisenberg-Pauli extension of the Dirac theory to the relativistic case. The circumstance that there is no separation between the P's and Q's that is invariant under the entire Lorentz group in the case of a conventional field shows that there is no fully Lorentz-invariant manifold of classical wave functions in the covariant case that plays the same role as the manifold M in Sec. 2. Rather, the manifold of classical wave functions that is usually given in the field-theoretic case by a partial differential equation is analogous to the phase space in the case of a classical system of finitely many degrees of freedom. An element of such a manifold (e.g., a particular solution of Maxwell's equations, as an element of the manifold of all solutions) completely describes the "classical" state of the system. A point of the manifold M in Sec. 2, however, merely determined the location in physical space of the classical system; to specify its state completely requires in addition the momentum vector at the point. The collection of all such complete specifications forms a manifold B of twice the dimension of M.

Thus in the relativistic field-theoretic case, one is given an analog to the classical phase space B, but is not given any analog for the space M describing the spatial location of the system, nor is there any explicitly relativistic way to define such an analog. Therefore, in passing from the treatment of Sec. 2 to the case of an infinite covariant physical system it is natural to attempt to interpolate a treatment of a finite system directly in terms of its phase space, in such a manner that the P's and Q's are dealt with on an equal footing. The point of this interpolation is primarily theoretical; there are in fact no nontrivial and realistic Lorentz-invariant systems of finitely many degrees of freedom. But it is useful to be able to develop the formalism free from the analytical complications that are present in the case of infinite systems, and in fact the results for the finite case will be needed in dealing with the infinite case.

A classical phase space such as B is not at all an arbitrary space, but has a special structure. In the case of a conventional classical system of n degrees of freedom, a point of B is often specified by a vector $(q_1,\cdots,q_n, p_1,\cdots,p_n)$ whose first n components give the spatial location of the system, and whose last n give its momenta. When the spatial location is described by a point of a nonlinear manifold M, such a coordination is generally only locally valid. In intrinsic terms, a point of the phase space B is a pair consisting of a point of M together with a vector in M at the point, the components of the latter being the various momenta. (Cf., e.g., Veblen and Whitehead.[12]) The conventional $(q_1,\cdots,q_n)$ give a nonintrinsic way of specifying the point of M, while the $(p_1,\cdots,p_n)$ give a similar specification for the vector. The key property of B from the standpoint of dynamical theory is its covariant association with a distinguished differential form of second degree, say Ω, which is defined by the

[11] L. Gross, Trans. Am. Math. Soc. **94**, 404 (1960).

[12] O. Veblen and J. H. C. Whitehead, *Foundations of Differential Geometry* (Cambridge University Press, New York, 1932).

equation

$$\Omega=\sum_{i=1}^{n} dp_i dq_i,$$

in the vicinity of a point $(\mathbf{p},\mathbf{q})$ of B, where $(q_1,\cdots,q_n)$ are local coordinates in M near $\mathbf{q}$, and $(p_1,\cdots,p_n)$ are the corresponding coordinates for the vectors at $\mathbf{q}$. This form is nondegenerate and determines a nowhere vanishing positive element of measure $dm=\Omega^n=\prod_i dp_i dq_i$. There is no difficulty in verifying that Ω, and hence also m, are independent of the local coordinates, and are globally defined on B. A dynamical or "contact" transformation is then defined as a point transformation on B that leaves invariant the form Ω.

To quantize the system, starting from the phase space B, observe first that for any vector field X on M there is, as is well known (cf. Whittaker[13]) a corresponding contact transformation $p(X)$ on B. It suffices to defined $p(X)$ locally, in terms of the local description of X in a particular coordinate system. Writing $X=\sum_i a(\partial/\partial q_i)$, then

$$p(X)=X-\sum_{i,j} p_i[(\partial a_i/\partial q_j)(\partial/\partial p_j)]$$

(this is the contact transformation corresponding to the Hamiltonian $H=\sum_i p_i a_i$). Next, for any function f on M, there is a corresponding infinitesimal contact transformation $q(f)$ on B: $q(f)=-\sum_j[(\partial f/\partial q_j)(\partial/\partial p_j)]$ (this corresponds to the Hamiltonian f).

Next observe that the $p(X)$ and $q(f)$ satisfy almost the same algebraic relations as the $P(X)$ and $Q(f)$. Specifically, it is straightforward to compute

$$[p(X),p(X')]=p([X,X']),$$
$$[p(X),q(f)]=q(Xf),$$
$$[q(f),q(f')]=0.$$

There is, however, a certain difference, which is quite fundamental, namely, that $q(f)=0$ if f is constant; in particular $p(X)$ and $q(f)$ commute in the linear case when X is an infinitesimal translation and f a linear function. Thus the $p(X)$ and $q(f)$ do not directly give quite a canonical system; but there is an invariant construction employing them that gives such a system.

If T is an infinitesimal contact transformation, it defines a Hermitian operator in $L_2(B,m)$, where the measure m is determined by the fundamental form $dm=\prod_i dp_i dq_i$, by its direct action: $h\rightarrow -iTh$, for any function h that is square-integrable over B. Now the form Ω is an exact differential: $\Omega=d\omega$, where ω is the differential form of first order $\sum_i p_i dq_i$ which is invariant on M. Associated invariantly with T and ω is the function on B, $\omega(T)$, and the definition

$$R(T)=-iT+\omega(T)$$

then gives a Hermitian operator in $L_2(B,m)$. It follows from the formula in the theory of differentiable manifolds for the derivative of a one-form (or alternatively by direct computation) that the $R(T)$ satisfy the commutation relations

$$[R(T),R(T')]=-iR([T,T'])+\Omega(T,T').$$

Now when T and T' are taken as the $p(X)$ and $q(f)$, one has, by direct computation

$$\Omega[p(X),q(f)]=Xf,$$
$$\Omega[p(X),p(X')]=\omega([X,X']),$$
$$\Omega[q(f),q(f'')]=0.$$

In particular, substituting in the foregoing commutation relations and defining $\tilde{P}(X)=R[p(X)]$ and $\tilde{Q}(f)=R[q(f)]$, there results

$$[\tilde{P}(X),\tilde{P}(X')]=\tilde{P}([X,X']),$$
$$[\tilde{P}(X),\tilde{Q}(f)]=\tilde{Q}(Xf),$$
$$[\tilde{Q}(f),\tilde{Q}(f')]=0.$$

Here $\tilde{P}(X)$ and $\tilde{Q}(f)$ vanish if X vanishes or f is constant, respectively, but they have the proper commutation relations in the case of an infinitesimal coordinate and a linear function. The last set of equations are in fact identical with the commutation relations given at the beginning of the preceding section.

Now the foregoing commutation relations not only extend the conventional ones of the nonrelativistic quantum mechanics of finite systems, but are closely analogous to those used in footnote reference 2 for the quantization of general Bose-Einstein fields. In view of this, and since we seek a formulation in which the p's and q's are treated symmetrically, we make definition 3.

Definition 3. A Schrödinger canonical system over a phase space B with exact fundamental differential form Ω is a mapping $X\rightarrow R(X)$ from the infinitesimal contact transformations on B to the self-adjoint operators in the space $L_2(B,\Omega^n)$ of square-integrable functions over B with respect to the canonical measure on B, of the form $R(X)=X+\omega(X)$, where ω is a first-order differential form such that $d\omega=\Omega$.

In case B is simply connected, any two ω's differ by the differential of a function, multiplication by the complex exponential of which gives a unitary transformation taking the one Schrödinger system into the other. Assuming now, that B is simply connected, a rather slight restriction as far as our purposes go, we may speak of *the* Schrödinger system on B with no essential ambiguity, as in Sec. 2.

Any contact transformation on B, say T, gives rise to a unique transformation of the canonical variables defined by the property of taking $R(X)$ into $R(X^T)$, for an arbitrary vector field X, where X^T denotes the vector field into which X is transformed by T. In this way it is possible to pass uniquely from a given classical kinematics (or even dynamics) to corresponding quantum-mechanical ones. The foregoing would appear

[13] E. T. Whittaker, *Analytical Dynamics* (Cambridge University Press, New York, 1959).

to be the simplest quantization scheme that is invariant under all classical contact transformations, although, as discussed below, it is open to serious question whether *all* the $R(X)$ are truly observable, or equivalently whether some additional selection principle does not operate, as well as to what extent the dynamics just defined agrees with the conventional substitution rule.

To see the connection with conventional theory, consider the case when B is the phase space for three-dimensional Euclidean space M. At first glance it would appear that the Hilbert space $L_2(B)$ is far too large, and that the present theory must be materially different from the conventional one. The point is however that the elements of $L_2(B)$ serve only to set up our observable algebra, and have primarily analytical rather than physical significance; our states are linear forms on our observable algebra, and only coincidentally expressible in terms of vectors in specific Hilbert spaces. The $R(X)$ with X restricted to be the extension to B of a Euclidean motion in M, or the infinitesimal contact transformation whose Hamiltonian is a linear function on M, or a sum of two such vector fields, satisfy the very same commutation relations as the conventional linear and angular momenta, and position observables. It follows therefore from the Stone-von Neumann theorem on the uniqueness of the Schrödinger operators,[14] or actually by a fairly simple direct reduction in this case, that these $R(X)$ are identical with the conventional Schrödinger operators, not within unitary equivalence, but what is physically just as effective, within unitary equivalence and multiplicity. There is no difficulty in verifying that the kinematics defined above for the $R(X)$ is in corresponding identity with the conventional kinematics. The dynamics is also in agreement in the two formulations, for the case when the Hamiltonian is at most quadratic in the canonical variables; but for a general Hamiltonian the two formulations are incomparable *a priori* because the class of $R(X)$ singled out in connection with conventional theory is not invariant under a general contact transformation. Thus for a free particle or harmonic oscillator the two theories are in precise agreement (cf. Segal[15]); but for, say, the hydrogen atom problem, the relationship is obscure. This is not of special concern to us because our primary interest is in the covariant case, and we could hardly expect to solve in an incidental way the much considered problem of formulating a unique way of passing from a classical nonrelativistic Hamiltonian to a quantum-mechanical one, which is invariant under contact transformations, etc. It would nevertheless be of significant independent interest to determine in the hydrogen atom case the precise connection between the theories, which may possibly be in agreement within terms of order h^2. [An eigenstate of the present motion in $L_2(B)$ gives rise to a linear form on the subsystem generated by the special class of $R(X)$ designated before, which in turn gives a linear form on the conventional system of operators on $L_2(M)$; this should be a pure state within $O(h^2)$ which has a wave function agreeing with a conventional hydrogen atom wave function within $O(h^2)$.]

A natural and general way to pick out the relevant special class of $R(X)$ seems to be to make use of a Riemannian structure in B, which it will inherit from that of M, in case B originates from an M. When M is Riemannian and $q_1,\cdots,q_n$ are normal coordinates at a point, while $p_1,\cdots,p_n$ are corresponding vector coordinates, the symmetric quadratic form $(\frac{1}{2})\sum_k (dp_k^2 + dq_k^2)$ defines a Riemannian structure in B. An infinitesimal complex structure can be introduced in B by defining multiplication by i to act in each tangent space of B by taking the dp's into the corresponding dq's and the dq's into the corresponding $-dp$'s; this structure is evidently intrinsic, and in combination with the form Ω, gives a positive definite Hermitian structure to each tangent space of B. When B arises from an M the infinitesimal complex structure will be integrable only when M has vanishing curvature, according to a result obtained by K. Kodaira (written communication via N. Steenrod) and also by A. Frölich and A. Nijenhuis (oral communication). The case of a given Hermitian manifold B not necessarily originating from an M, is, however, more relevant to relativistic field-theoretic situations. In any event a transformation on a Hermitian manifold B may be called isometric in case it preserves the Hermitian inner product in each tangent space to the manifold; and the observables $R(X)$ may, in the case when B is endowed with a Hermitian structure having Ω as the imaginary part of the inner product, be restricted to those for which X is infinitesimally isometric. This is natural from a mathematical viewpoint, and it will be seen later that it gives the conventional theory in the case of covariant free fields, as well as, as noted earlier, in the case of elementary quantum mechanics. It should perhaps be emphasized that, in any case, the presence of the additional $R(X)$ does not in any way alter the physical conclusions concerning the subsystem generated by some restricted class of $R(X)$—the stationary states and expectation values, transformation properties, etc., of the subsystem are unaffected by treating it as a subsystem rather than as a full system in itself.

A theoretically less severe limitation on the $R(X)$ to be used in forming the subsystem of interest, although for many manifolds apparently an equivalent limitation, is the use only of those for which X is holomorphic, i.e., commutes with the operation defining multiplication by i in each tangent space. In a formal way one may in fact describe the relevant states explicitly, as represented by the holomorphic functions on B.

[14] J. von Neumann, Math. Ann. **104**, 570 (1931).
[15] I. E. Segal, Can. J. Math. (to be published).

Example. Let B be a complex n-dimensional space, with coordinates $z_1, z_2, \cdots, z_n$, and fundamental Hermitian form $\sum_k dz_k dz_k^*$. The isometry group is generated by the translations together with the homogeneous unitary transformations. Writing $z_k = p_k + iq_k$, where p_k and q_k are real, and setting $P_k = R(\partial/\partial p_k)$ and $Q_k = R(\partial/\partial q_k)$, gives the conventional commutation relations. Choosing $\omega = (\frac{1}{2}) \sum_k (z_k dz_k^* - z_k^* dz_k)$ gives specifically $P_k = [(1/i)(\partial/\partial p_k)] + q_k$ and $Q_k = [(1/i)(\partial/\partial q_k)] - p_k$. These look rather different from the conventional quantum-mechanical variables, but by the uniqueness result cited, must be the same, apart from a unitary transformation, and the introduction of a multiplicity (visible in the circumstance that B has twice the dimensionality of the real manifold on which the Schrödinger representation is based). It is actually not difficult in this case to exhibit specifically, without the use of the uniqueness result, the decomposition of the present operators in the form

$$P_k = P_k^0 \times I; \quad Q_k = Q_k^0 \times I,$$

where (P_k^0, Q_k^0) are the conventional Schrödinger operators and I is the identity operator in a certain Hilbert space.

For any unitary transformation U on B, there will be a corresponding unitary transformation $\Gamma(U)$ on $L_2(B)$, transforming the canonical P's and Q's into corresponding linear combinations of themselves, by virtue of the fact that transformation of a translation by a unitary is another translation. More generally this is true of any linear contact transformation (i.e., so-called "symplectic" transformation). Of particular interest is the case when $U(=U_t)$ is multiplication by e^{it}, $z_k \to e^{it} z_k$; $\Gamma(U_t)$ is then a one-parameter group of operators whose generator is the conventional harmonic oscillator (isotropic) Hamiltonian. Its ground state, as an expectation value linear functional on the algebra generated by the $R(X)$ with X an infinitesimal isometry is invariant under the $\Gamma(U)$ with U unitary.

Now let B be any simple connected Hermitian manifold with associated closed form Ω, for short phase manifold. There is then a corresponding theory. Because there is in general no distinction analogous to that between the translations and homogeneous transformations, all the various momenta, and even the position coordinates, are treated on the same footing. Transformation properties of these canonical variables under the subgroup G_0 of isometries leaving fixed a point ϕ_0 of B are quite analogous to those in the linear case, where ϕ_0 is the origin and G_0 the unitary group.

To formulate the relevant intrinsic nonlinear analog to the ground state of the harmonic oscillator, fix a point ϕ_0, and consider the connection between the tangent plane T_{ϕ_0} at ϕ_0 and the manifold B. The "exponential" map introduced in differential geometry by Whitehead, taking a vector l of T_{ϕ_0} into a corresponding point $\exp(l)$ in B, lying an appropriate distance from ϕ_0 on the geodesic from ϕ_0 in the direction of l, gives a local linear parametrization of B, which will have, in general, certain singularities in the large. These singularities will, however, form only sets of measure zero in T_{ϕ_0} and in B, in the case of many manifolds, particularly those whose deviation from linearity arises from the nontriviality of the fundamental Hermitian form, rather than from the nontriviality of the connectivity properties of the manifold B, as is formally the case of basic interest here. (The manifold of solutions of a nonlinear hyperbolic equation is from the quite heuristic standpoint usually employed in theoretical physics topologically flat, as it is generally implicitly assumed that the admissible Cauchy data at a particular time do not need to satisfy any special nonlinear conditions, and determine the solution throughout space-time.) At any rate, for a fairly extensive and interesting class of manifolds M, the map $l \to \exp l$ will give rise to a well-defined mapping of sets into sets, if sets of measure zero are neglected, and thereby to a linear and multiplicative correspondence between the measurable functions on T_{ϕ_0} and those on B.

Any unitary transformation U on T_{ϕ_0} will give a corresponding transformation $\Gamma_0(U)$ on $L_2(T_{\phi_0})$, and by virtue of the foregoing correspondence, a transformation $\Gamma(U)$ of $L_2(B, \Omega^n)$. Choosing U to be multiplication by e^{it} (t real) gives then a one-parameter group on $L_2(B, \Omega^n)$, whose generator may be designated as the Hamiltonian for the *generalized harmonic oscillator on B at ϕ_0*. This will not necessarily be self-adjoint relative to the given inner product, but it will have real, and in fact integral eigenvalues. It may also reasonably be conjectured that in the cases of interest, and in particular when B is obtainable by continuous deformation of a linear manifold, the spectrum will be bounded from below, and the ground state will be unique, as an expectation value functional on the functions of the $R(X)$ for isometric X.

The point of this construction is that it picks out in a natural and well-defined way a particular state that is invariant under the group G_0 of isometries leaving invariant the point ϕ_0. This will be useful in getting at the physical vacuum in the case of fields, where such invariance presumably characterizes the physical vacuum, although in the finite-dimensional case there will generally be other invariant states under G_0.

Now when B is a complex unitary space, the corresponding physical situation is considered to be free of interaction, and in a certain sense this is evidently true of the situation for a general Hermitian manifold B. But from the standpoint of an observer who utilizes as a reference system the tangent plane to B at a particular point ϕ_0—i.e., the reference system appropriate for the examination of small displacements from a particular classical state—interaction is present. In

particular the ground state of B and the ground state of the linear system associated with the tangent plane are not simple transforms of one another, nor does the isometry group leaving ϕ_0 invariant transform B in the same way as does the transform via the exponential map of the linear action of the isometry group on the tangent plane at ϕ_0.

The preceding line of development may be summarized as follows.

Principle II. Let there be given in a simply connected manifold B of states of a physical system a distinguished locally exact Hermitian differential form. There is then a unique and mathematically precise scheme for setting up quantization conditions, which extends elementary quantum mechanics as well as the conventional quantization theory for relativistic free fields. The relevant covariance group is that of all transformations on the manifold leaving invariant the fundamental form (isometries, that is) as well as a distinguished point of B. The canonical variables $R(X)$ are associated with infinitesimal isometries, and satisfy the commutation relations (4); *they are Hermitian first-order linear differential operators on B, relative to the canonical measure determined by* Ω. *There exists with significant generality a ground state on B analogous to the lowest eigenstate of a harmonic oscillator in elementary quantum mechanics.*

The situation as regards field quantization needs to be elaborated somewhat; this will be done in the next section, where the foregoing principle will be applied to the quantization of a nonlinear hyperbolic partial differential equation.

4. INFINITE SYSTEMS

This section extends the preceding one to the case of infinite systems, and indicates how this extension may be used to quantize a given nonlinear hyperbolic partial differential equation.

A. Formal Differential Geometry of Nonlinear Hyperbolic Equations

For simplicity and concreteness we treat here primarily the classical (unquantized) system M defined by the equation

$$\Box\phi = m^2\phi + p(\phi),\ p \text{ a polynomial vanishing at } 0; \quad (5)$$

the extension to rather more general cases appears to involve no great difficulties. M may be regarded as an infinite-dimensional manifold that is imbedded in the manifold (say S) of all scalar functions on space-time. Consequently, at any point ϕ of M there will be a tangent plane T_ϕ defined by the equation

$$\Box\lambda = m^2\lambda + p'(\phi)\lambda. \quad (6)$$

For arbitary given ϕ in S, this equation defines a linear manifold in S, and in fact M may be considered as an integral manifold of this distribution of linear manifolds, that passing through the point $\phi=0$. (It is worth noting that the satisfaction of the relevant integrability conditions by this distribution of linear manifolds is purely a matter of linear analysis, and so on a much more accessible level than the questions of classical nonlinear analysis involved in the structure of M as first defined.) At this generally substantially unique Lorentz-invariant point of M, the tangent plane is defined by the so-called "free-field" equation

$$\Box\lambda = m^2\lambda. \quad (7)$$

Since Eq. (6) is linear and hyperbolic, there is for any fixed function ϕ a unique function $D_\phi(x,x')$ of ordered pairs of points of M, which satisfies (6) as a function of the first point x, and also the following initial conditions [employing the notation $x=(\mathbf{x},x_0)$]:

$$\left.\begin{aligned} D_\phi(x,x') &= 0 \\ (\partial/\partial x_0)D_\phi(x,x') &= \delta(\mathbf{x}-\mathbf{x}') \end{aligned}\right\} \text{ when } x_0 = x_0'. \quad (8)$$

Now this function also satisfies the differential equation as a function of x', or more exactly:

Heuristic Proposition 1. $D_\phi(x,x') = -D_\phi(x',x)$ *for arbitrary* x *and* x'.

Argument: It suffices to show that $-D(x',x)$ (suppressing the dependence on ϕ, which is here irrelevant) satisfies the defining conditions for $D(x,x')$. The first condition of Eq. (8) is obvious, and for the second condition, it may be noted that

$$\begin{aligned} \left.\frac{\partial[-D(x',x)]}{\partial x_0}\right|_{x_0=x_0'} &= -\lim_{\epsilon\to 0}\epsilon^{-1}D(\mathbf{x}',x_0',\mathbf{x},x_0'+\epsilon) \\ &= \lim_{\epsilon\to 0}\epsilon^{-1}[D(\mathbf{x}',x_0'+\epsilon,\mathbf{x},x_0'+\epsilon) \\ &\quad - D(\mathbf{x}',x_0',\mathbf{x},x_0'+\epsilon)] \\ &= \left.\frac{\partial}{\partial x_0'}D(\mathbf{x}',x_0',\mathbf{x},x_0)\right|_{x_0'=x_0} \\ &= \delta(\mathbf{x}-\mathbf{x}'). \end{aligned}$$

It remains only to show that $M(x,x')$ vanishes identically, where $M(x,x') = [\Box_x - V(x)]D(x',x)$, writing $V = m^2 + p'(\phi)$. To this end it suffices to show that $M(x,x')$ is the solution to a Cauchy problem with vanishing initial data. We shall regard it as a function of x' with initial values given on the hyperplane $x_0=x_0'$. Since $[\Box_{x'} - V(x')]D(x',x) = 0$ by the definition of $D(x,x')$, and since $\Box_{x'} - V(x')$ as an operator commutes with $\Box_x - V(x)$, we have

$$[\Box_{x'} - V(x')]M(x,x') = 0.$$

Now let us evaluate $M(x,x')$ for $x_0=x_0'$. The only contribution whose vanishing is not apparent is

$$(\partial^2/\partial x_0)[2D(x',x)]|_{x_0=x_0'}.$$

This may be written as

$$\lim_{\epsilon\to 0}(2\epsilon)^{-2}[D(x_0, x_0+2\epsilon)-2D(x_0, x_0+\epsilon)+D(x_0,x_0)],$$

where for the moment we suppress the dependence on $\mathbf{x}$ and $\mathbf{x}'$. Now

$$D(x_0, x_0+2\epsilon)=-[D(x_0+2\epsilon, x_0+2\epsilon)-D(x_0, x_0+2\epsilon)],$$

which by Taylor's expansion and the definition of $D(x,x')$ is

$$-2\epsilon\delta(\mathbf{x}-\mathbf{x}')+2\epsilon^2\left[\frac{\partial^2}{\partial x_0'}2D(x_0', x_0+2\epsilon)\right]_{x_0'=x_0+2\epsilon}+0(\epsilon^3).$$

From the fact that $D(x',x)$ satisfies the differential equation as a function of x', it follows that $(\partial^2/\partial x_0)\times[2D(x,x')]$, evaluated for $x_0=x_0'$, is the same as $[\Delta_x-V(x)]D(x,x')$, likewise evaluated for $x_0=x_0'$, where Δ denotes the Laplacian; and hence the middle terms in the preceding expression vanishes. A similar evaluation applies to $D(x_0, x_0+\epsilon)$, from which it results that $(\partial^2/\partial x_0)[2D(x',x)]\,|_{x_0=x_0'}=\lim_{\epsilon\to 0}(2\epsilon)^{-2}0(\epsilon^3)=0$.

It remains only to show the vanishing of $(\partial/\partial x_0')\times[M(x,x')]$ for $x_0=x_0'$. Writing $L_x=\Delta_x-V(x)$, we need to examine

$$\left[L_x\frac{\partial D(x',x)}{\partial x_0'}\right|_{x_0=x_0'}-\frac{\partial^2}{\partial x_0^2}\frac{\partial}{\partial x_0'}D(x',x)\Big|_{x_0=x_0'}.$$

Since L_x involves no differentiation with respect to time, the first term is the same as

$$L_x\left[\frac{\partial D(x',x)}{\partial x_0'}\right|_{x_0=x_0'}=L_x\delta(\mathbf{x}-\mathbf{x}').$$

In evaluating the second term, we write $x_0=t$, $x_0'=t'$, and note that

$$\frac{\partial^2}{\partial t^2}\frac{\partial}{\partial t'}=\frac{\partial}{\partial t}+\frac{\partial}{\partial t'}\frac{\partial^2}{\partial t\partial t'}-\frac{\partial}{\partial t'}\frac{\partial^2}{\partial t\partial t'}.$$

The term $[(\partial/\partial t')(\partial^2/\partial t^2)]$ develops as follows:

$$\frac{\partial^2}{\partial t'^2}\frac{\partial}{\partial t}D(x',x)\Big|_{t'=t}=\frac{\partial}{\partial t}L_{x'}D'x',x)\Big|_{t'=t}$$

but $L_{x'}$ and $\partial/\partial t$ commute, and $L_{x'}$ involves no differentiation with respect to time, so that the expression reduces to

$$L_{x'}\left[\frac{\partial}{\partial t}D(x',x)\right]_{t=t'}=-L_{x'}\delta(\mathbf{x}-\mathbf{x}')$$

by an earlier result. This precisely cancels the first term, so to conclude the argument it suffices to show that

$$\left[\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial t'}\right)\frac{\partial^2}{\partial t\partial t'}D(x',x)\right]_{t=t'}=0.$$

Now

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial t'}\right)F(t,t')\Big|_{t=t'}$$

vanishes identically if $F(t,t)$ is a constant, so it suffices to show that $\{[(\partial^2/\partial t\partial t')]D(x',x)\}_{t=t'}$ is a constant, as a function of t. Actually it vanishes, for it may be written as

$$\lim_{\epsilon\to 0}\epsilon^{-2}[D(t+\epsilon, t+\epsilon)-D(t, t+\epsilon)-D(t+\epsilon, t)+D(t,t)]$$
$$=\lim\epsilon^{-2}[(\epsilon^2/2)L_{x'}D(x,x')-(\epsilon^2/2)L_{x'}D(x,x')+0(\epsilon^3)]_{t=t'}=0,$$

since $L_{x'}D(x,x')$ vanishes for $t=t'$.

The function $D_\phi(x,x')$ thus determines a skew-symmetric bilinear form $B_\phi(l,l')$ in the solutions of (6):

$$B_\phi(l,l')=\int\int D_\phi(x,x')l(x)l(x')d_4xd_4x'.$$

Thus for each pair of tangent vectors at ϕ there is a skew-symmetric bilinear functional of them; this is by definition a second-order differential form on M. This form will be denoted as Ω, and called the fundamental form on M. To see the connection between this form and the similarly designated form in classical mechanics, it is useful to observe that

in the case of the Klein-Gordon equation ($p=0$ identically),

$$\Omega=\sum_{k=1}^{\infty}dp_kdq_k,$$

where $p_1, p_2, \cdots, q_1, q_2, \cdots$ *are "natural" coordinates on* M.

Specifically, the p_k and q_k are obtained by choosing any complete orthonormal set of Klein-Gordon wave functions invariant under time reversal, say $f_1, f_2, \cdots$, and writing a general real Klein-Gordon wave function f as

$$f=\textstyle\sum_k(p_kf_k+q_k\tilde{f}_k),$$

where the tilde denotes the action of forming the Hilbert transform with respect to time. The convergence of the infinite sum presents no essential difficulty, as is clear from the following argument, which also serves to make clear how such sums are to be interpreted.

In a linear space, any differential form may be expanded into a product of differentials of linear coordinates, showing that two differential forms are the same if they agree on all generators of infinitesimal translations. Hence it suffices to show that if f and g are arbitrary normalizable real Klein-Gordon wave functions, then

$$\Omega\left(\frac{\partial}{\partial f},\frac{\partial}{\partial g}\right)=\sum_{k=1}^{\infty}(dp_kdq_k)\left(\frac{\partial}{\partial f},\frac{\partial}{\partial g}\right);$$

here $\partial/\partial f$ stands for the vector field on M generating the transformations $\phi \to \phi + sf (-\infty < s < \infty)$. The left side reduces to

$$\int\int f(x)g(x')D(x-x')d_4x d_4x',$$

where D denotes the familiar scalar particle commutation function. The right side is

$$\tfrac{1}{2}\sum_k \left(\frac{\partial p_k}{\partial f}\frac{\partial q_k}{\partial g} - \frac{\partial q_k}{\partial f}\frac{\partial p_k}{\partial g}\right).$$

It is readily seen that $(\partial p_k/\partial f) = (f,f_k)$, etc., so that the identity of the two forms on infinitesimal translations reduces to the equation

$$\int\int f(x)g(x')D(x-x')d_4x d_4x' = \tfrac{1}{2}\sum_k [(f,f_k)(g,\tilde{f}_k) - (f,\tilde{f}_k)(g,f_k)],$$

which can be verified without difficulty; here (u,v) denotes the unique real Lorentz-invariant inner product between the wave functions u and v (suitably normalized).

To develop further the differential geometry of the function manifold M, we require:

Heuristic Proposition 2. The transformation K_ϕ: $l(x) \to \int D_\phi(x,x')l(x')dx'$, *acting in the tangent plane* T_ϕ *to* M *at* ϕ, *has the property that* $K_\phi^2 = -I_\phi$, *where* I_ϕ *denotes the identity operation in* T_ϕ.

Argument: This is easily seen for the case $\phi = 0$ by the use of Fourier transforms, or by reduction to a similar property for the Hilbert transform in one dimension. Now suppose that ϕ is "small," in the sense that

$$D_\phi(x+su, x'+su) \to D_0(x,x') \quad \text{as} \quad s \to \pm\infty$$

for any timelike vector u; this means that for large times, $D_0(x,x')$ behaves like the "free" commutator function. In the equation

$$\Box l = m^2 l + V(x)l,$$

it is reasonable to suppose that if, say, V is smooth and vanishes outside of a bounded set, then for large times the situation is asymptotic to that for the equation

$$\Box l = m^2 l.$$

Rigorous results of this sort are not yet available in the mathematical literature, but substantial results in this direction in nonrelativistic cases have been established by Kato, Cook, and others (cf. e.g., Kuroda[16] and the literature cited therein), and in any event such results have a high degree of plausibility from the standpoint of theoretical physics, constituting a weakened and classical version of the adiabatic hypothesis of quantum field theory (cf. Yang[17]).

Now let l be any tangent vector at ϕ. Then $l(x+su)$ is aymptotic, for large s and fixed timelike u, to a solution $l_0(x)$ of the free-field equation. Conversely, l may be characterized as that solution of (6) that is asymptotic to the particular free-field wave function l_0 for early times, i.e., it may be regarded as the solution of a Cauchy problem with data given at time $-\infty$. It is evident that

$$l'(x) = \int D(x,x')l(x')dx'$$

is likewise a solution of (6); and

$$l'(x+su) = \int D(x+su, x')l(x')dx' = \int D(x+su, x'+su)l(x'+su)dx'.$$

Now as

$$s \to -\infty, \quad l'(x'+su) \to l_0'(x')$$

and $D(x+su, x'+su) \to D_0(x,x')$. Assuming now that the passage to the limit may be made under the integral sign, it follows that

$$l'(x+su) \to \int D_0(x,x')l_0(x')dx'.$$

Thus

$$\lim_{s\to-\infty}(K_\phi l)(x+su) = K_0 \lim_{s\to-\infty} l(x+su),$$

where K_0 is the transformation on the free-field wave functions with kernel $D_0(x,x')$. If we denote by T the transformation from the solutions of the free-field equation to those of (6) asymptotic to the given free-field wave function at early times, the foregoing result means that

$$T^{-1}K_\phi T = K_0.$$

Hence $T^{-1}K_\phi^2 T = K_0^2$, and since $K_0^2 = -I_0$, it follows that $K_\phi^2 = -I_\phi$.

Now the property $K_\phi^2 = -I_\phi$ is a variety of functional equation having no explicit reference to the size of ϕ; if it is valid for sufficiently small ϕ, then it should be valid as a general rule. For example, if ϕ is a constant, then the result is evidently valid, although the argument for small ϕ certainly is not.

It is now easy to derive:

Heuristic Proposition 3. M becomes endowed with a positive definite Hermitian metric if the following definitions are made:

(1) For any two tangent vectors l and l' at ϕ, the inner product is given by the equation

$$(l,l')_\phi = \sum_\phi (l,l') + i\Omega_\phi(l,l'),$$

[16] S. T. Kuroda, J. Math. Soc. Japan **11**, 247 (1959).

[17] C. N. Yang and D. Feldman, Phys. Rev. **79**, 972 (1950); G. Källén, Arkiv Fysik **2**, 33 (1950).

where

$$\sum_\phi (l,l')=\Omega_\phi(K_\phi l,l').$$

(2) Complex scalars act on tangent vectors l in accordance with the unique extension of the manner in which real scalars act together with the rule

$$il=K_\phi l.$$

Argument: The only point that is not immediate from a quite general argument is the definiteness of the inner product, i.e., that $(l,l)_\phi>0$ if l is not the zero tangent vector. Since Ω_ϕ is skew-symmetric,

$$(l,l)_\phi=\sum_\phi (l,l)=\Omega_\phi(K_\phi l,l)$$

$$=\int\int D(x,y)l(y)D(y,z)l(z)dxdydz$$

$$=\int\left(\int D(x,y)l(y)dy\right)^2 dx.$$

This shows that $(l,l)_\phi\geq 0$, and that the equality holds here only if $\int D(x,y)l(y)dy$ vanishes. But this is $K_\phi l$, which by Proposition 2 can vanish only if $l=0$.

The remainder of the argument is of a simple and familiar algebraic character—cf. Ehresmann[18]—and may be omitted.

It may be illuminating to consider the content of proposition 3 in the case $\phi=0$, which is easily dealt with explicitly. It says that the space of real normalizable solutions of the Klein-Gordon equation may be given the structure of a complex Hilbert space. The action of i is given by the Hilbert transform with respect to the time variable; the imaginary part of the inner product is given by the form whose kernel is the commutator function; and the real part is obtained by replacing one of the entries in this skew-symmetric form by its Hilbert transform with respect to time, obtaining thereby a positive definite real symmetric form. This complex Hilbert space is easily seen to be in one-to-one correspondence, in an essentially unique Lorentz-invariant fashion, with the conventional space of normalizable positive-frequency complex-valued Klein-Gordon wave function (cf. Segal[19]).

To complete the analogy with a classical mechanical system, and make available formally the apparatus developed in Sec. 3, we require further:

Heuristic Proposition 4. The form Ω is closed,—its covariant differential $d\Omega$ vanishes.

Argument: The evaluation of $d\Omega(X,Y,Z)$ involves considerable computation which we shall not carry out here. There is another way of arguing, which while quite heuristic, throws light on the origin of the form Ω.

As noted earlier, our manifold M may be considered as a submanifold of the manifold S of all scalar functions on space-time. Now any form on S gives rise, by restriction of the tangent vectors to tangency to M, to a form on M; and this restricted form will be closed if the original form was such. In particular this is true of the form $\tilde{\Omega}$ on S defined by the equation

$$\tilde{\Omega}\left(\frac{\partial}{\partial f},\frac{\partial}{\partial g}\right)=\int f(\mathbf{x},t)g(\mathbf{x},t')d_3\mathbf{x}\frac{dtdt'}{t-t'},$$

or alternatively, as

$$\tilde{\Omega}=\sum_k dp_k dq_k,$$

where $p_1, p_2, \cdots q_1, q_2\cdots$ are coordinates on S similar to those defined earlier.

We may formulate S as an infinite-dimensional Riemannian manifold by assigning to each tangent space S_ϕ—the general element of which has the form $\partial/\partial\psi$ for some formally unrestricted scalar function ψ—the usual inner product, i.e., $[(\partial/\partial\psi),(\partial/\partial\psi')]=\int\psi(x)\psi'(x)d_4x$. Thus any such tangent space is isomorphic to the real Hilbert space H of all real square-integrable functions over space-time. This space can be decomposed into eigenspaces of the self-adjoint operator $\Box$, as a so-called "direct integral" of (infinitesimal) eigenspaces $H_s(-\infty<s<\infty)$, so there is a corresponding decomposition of S_ϕ into eigenspaces $S_\phi(s)$. Now at the point ϕ, $\tilde{\Omega}$ gives a skew-symmetric bilinear form $\tilde{\Omega}_\phi$ in the vectors of S_ϕ, which may be restricted to any eigenmanifold packet, say that corresponding to the eigenvalues in the interval $(s-\epsilon, s+\epsilon)$, yielding a bilinear skew-symmetric form $\tilde{\Omega}_\phi(s-\epsilon, s+\epsilon)$, in the vectors of this eigenmanifold. Now as $\epsilon\to 0$, the difference quotient $(2\epsilon)^{-1}\tilde{\Omega}_\phi(s-\epsilon, s+\epsilon)$ has a limit, which is a form $\tilde{\Omega}_\phi(s)$ in the vectors of the eigenspace corresponding to the eigenvalue s. It can be explicitly verified, by recourse to Fourier transforms, that if $s=m^2$, this form is the same as that introduced above with kernel $D_\phi(x,x')$, for the case $p(\phi)=0$, the eigenspace $S(m^2)$ being identical with the T_ϕ defined above.

Now S_ϕ may also be decomposed into eigenspaces of the self-adjoint operator $\Box-p'(\phi)$, and a similar form density $\tilde{\Omega}_\phi(s;p)$, which is a bilinear skew-symmetric form in the eigenspace of this operator with eigenvalue s, obtained. This eigenspace is identical with the T_ϕ, the tangent space to M at ϕ discussed before, and if we permit ourselves to use the plausible conjecture that the two intrinsically defined skew-symmetric bilinear forms on this space, $\tilde{\Omega}_\phi(m^2;p)$ and Ω, agree in general, as they do in the case $p=0$, then it follows (formally) that Ω is closed, being a limit of closed forms.

B. Subsumption of the Conventional Field-Theoretic Formalism

We now assume that we have the manifold M of all solutions of Eq. (6) set up as a Hermitian manifold with fundamental form Ω, and that linearly associated with each vector field X on M we have an operator

[18] C. Ehresmann, Proc. Int. Congr. Math. 1950 (Providence, 1952).

[19] I. E. Segal, Phys. Rev. **109**, 2191 (1958).

$R(X)$, the following commutation relations being satisfied:

$$[R(X),R(Y)]=R([X,Y])+\Omega(X,Y).$$

(The closure of Ω enters primarily as a means of assuring the consistency of these relations.) We wish now to define the "quantum field" $\tilde{\phi}(x)$ so that the conventional commutation relations and transformation properties are derivable.

For any scalar function f on space-time ("weighting function"), which is smooth and vanishes at infinity, we consider the vector field X_f, whose value at ϕ is the tangent vector $\int D(x,x')f(x')dx'$. Evidently, X_f depends linearly on f, and hence $R(X_f)$ does so also, so that we may write formally

$$R(X_f)=\int\tilde{\phi}(x)f(x)dx,$$

for some operator-valued function $\tilde{\phi}$ on space-time. We may also write

$\tilde{\phi}(x)=R(X_f)$ for $f=$a delta-function at x (formally).

We wish to show that $\tilde{\phi}(x)$ satisfies the conventional commutation relations. To this end, note to begin with that the solution of the Eq. (6), with Cauchy data $\varphi(x)=f(\mathbf{x})$ and $(\partial/\partial t)\varphi(x)=g(\mathbf{x})$ at $t=t_1$, is

$$\int_{x_0'=t_1}\left[\frac{\partial D(x,x')}{\partial x_0'}f(\mathbf{x}')+D(x,x')g(\mathbf{x}')\right]d_3\mathbf{x}',$$

for it is evident that this is a solution of (6); that for $t=t_1$, it attains the value $f(\mathbf{x})$, by the Cauchy data defining $D(x,x')$; and using the fact proved earlier that

$$\left[\frac{\partial^2}{\partial x_0\partial x_0'}D(x,x')\right]_{x_0=x_0'}=0,$$

it follows similarly that its time derivative for $t=t_1$ is $g(\mathbf{x})$. Now consider the one-parameter group of motions on M which takes a given wave function ϕ into one having the same values for $t=t_1$, but with $(\partial\phi/\partial x_0)$ displaced by $sg(\mathbf{x})(-\infty<s<\infty)$. The generator of this group of motions will be a vector field on M whose value at ϕ will be the solution of Eq. (6) having corresponding Cauchy data on the line $t=t_1$; it is, accordingly,

$$\int D(x,x')g(\mathbf{x}')d_3\mathbf{x},\quad(x_0'=t_1),$$

or X_f, where $f(x)=g(\mathbf{x})\delta(t-t_1)$. Thus

$$R(X_f)=\int\tilde{\phi}(\mathbf{x},t_1)g(\mathbf{x})d_3\mathbf{x}.$$

If we take another function $g'(\mathbf{x})$ and consider the one-parameter group of transformations of M which it determines in the same manner as $g(\mathbf{x})$, then it is clear that this group commutes with the group determined by g, since they both act additively on the Cauchy data at $t=t_1$. The corresponding vector fields therefore commute, and substitution in the fundamental commutation relation, after choosing g and g' as delta functions, gives the equation

$$[\tilde{\phi}(\mathbf{x},t),\tilde{\phi}(\mathbf{x}',t)]=\Omega(X_f,X_{f'}),$$

where $f(y)=\delta(\mathbf{y}-\mathbf{x})\delta(y_0-t)$ and $f'(y)$ is the same with $\mathbf{x}$ replaced by $\mathbf{x}'$. Substitution now in the equation defining Ω now gives for the right-hand side of the foregoing equation the value $D(\mathbf{x},t,\mathbf{x}',t)$, which vanishes by the definition of D. Thereby so-called "local commutativity" (or "microcausality") is established.

To evaluate $[\tilde{\phi}(\mathbf{x},t),(\partial/\partial t)\tilde{\phi}(\mathbf{x}',t)]$, consider $[\tilde{\phi}(x),\tilde{\phi}(x')]$, where $x_0'=x_0+\epsilon$, ϵ being small. Directly from the fundamental commutation relations we have

$$[\tilde{\phi}(x),\tilde{\phi}(x')]=R([X_{\delta x},X_{\delta x'}])+\Omega(X_{\delta x},X_{\delta x'}).$$

By an observation made earlier, $X_{\delta y}$ has at ϕ the value

$$\int_{x_0'=y_0}D(x,x')\delta(\mathbf{x}'-\mathbf{y})d_3\mathbf{x}',$$

or $D(x,y)$, as a function of x. Thus

$$\Omega(X_{\delta x},X_{\delta x'})=\int\int D(u,x)D(v,x')D(u,v)dudv.$$

Now proposition 2 may be restated as

$$\int D_\phi(x,x')D_\phi(x',x'')l(x'')dx'dx''=-l(x)$$

if l is in T_ϕ. In particular, putting $l(x)=D_\phi(x,y)$ with y fixed, it results that

$$\int D(x,x')D(x',y')D(y',y)dx'dy'=-D(x,y).$$

It follows that $\Omega(X_{\delta x},X_{\delta x'})=-D(x,x')$.

To evaluate $[X_{\delta x},X_{\delta x'}]$, recall that $X_{\delta x'}$ is the generator of the one-parameter transformation group on M, with the parameter s, which takes a general element ϕ of M into that element ϕ' such that

$$\left.\begin{aligned}\phi'(x)&=\phi(x)\\(\partial\phi'/\partial t)&=(\partial\phi/\partial t)+s\delta(\mathbf{x}-\mathbf{x}')\end{aligned}\right\}\text{ at }x_0=x_0'.$$

From this characterization we shall show that it commutes with $X_{\delta x'}$ within terms of order ϵ^2. It is perhaps clearer to deal more generally with a manifold M defined by an equation of the form

$$(\partial u/\partial t)=L(t)u,$$

where $L(t)$ is a nonlinear operator (i.e., $L(t)$ depends on t, but does not involve differentiations with respect

to t). From this equation it follows that

$$u(t')=u(t)+(t'-t)L(t)u(t)+0[(t'-t)^2].$$

If we now consider a one-parameter group dependent on a parameter s, which displaces M so that $u(t)\rightarrow u(t)+sv(t)$, t and v being held fixed, then the corresponding displacement of $u(t')$ may be computed as follows:

$$\begin{aligned}u(t')&\rightarrow[u(t)+sv(t)]+(t'-t)L(t)[u(t)+sv(t)]\\&\qquad+0[(t'-t)^2],\\&=u(t')+(t'-t)[\delta L(t)]_{u(t)}v(t)+sv(t)+0(s^2)\\&\qquad+0[(t'-t)^2],\\&=u(t')+sWv(t)+0(s^2)+Q[(t'-t)^2],\end{aligned}$$

where W is a certain *linear* operator (dependent on t and t', but not on s). That is, the infinitesimal displacement of $u(t)$ by an amount sv, displaces $u(t+\epsilon)$ by an amount $sWv+0(\epsilon^2)$; and the displacement of $u(t+\epsilon)$ acts similarly on $u(t)$. Since vector translations commute, it results that $X_{\delta x}$ and $X_{\delta x'}$ commute within terms of order ϵ^2.

Such terms contribute nothing to the commutator

$$\left[\tilde{\phi}(\mathbf{x},t),\frac{\partial}{\partial t}\phi(\mathbf{x}',t)\right]=\frac{\partial}{\partial t'}[\phi(x),\phi(x')]_{t=t'}.$$

The sole contribution is then

$$\frac{\partial}{\partial t}[-D(x,x')]_{t=t'}=-\delta(\mathbf{x}-\mathbf{x}').$$

Now consider the transformation properties of $\tilde{\phi}(x)$. Designating as a *contact transformation* one that preserves the fundamental form Ω on M, it is clear from the covariance of the construction of $\tilde{\phi}(x)$ that for any such transformation T, the field $\tilde{\psi}(x)=\tilde{\phi}(Tx)$ satisfies the same commutation relations. In a formal way the existence of an operator $U(T)$ with the property that $\tilde{\psi}(x)=U(T)\tilde{\phi}(x)U(T)^{-1}$ is clear, for $U(T)$ may be taken as the operator taking a formally square-integrable functional $f(x)$ over M into the functional $f(T^{-1}x)$. It is evident that in this quite formal sense the map $T\rightarrow U(T)$ is a unitary representation of the group of contact transformations, but this is not strictly the case even for free fields unless T is suitably restricted, e.g., to be an isometry (cf. footnote reference 2).

In conventional field theory it is assumed that the quantized field "satisfies" the original field equation. This is an equation involving local products of fields, and so has no definite mathematical meaning. It has also no empirical physical meaning. The present formalism eliminates these fundamentally objectionable features of the conventional theory, but this advantageous feature in itself limits the possibility of showing complete formal equivalence to conventional theory. It can be stated that the quantized field $\tilde{\phi}(x)$ is here derived in a covariant and unique manner from the classical system; but the equation that states that it "satisfies" the original differential equation has no clear-cut mathematical or physical meaning, and cannot be stated in the present formalism.

C. Convergence Considerations

Although local products of fields do not occur in the formulation of the dynamics of the present quantum fields, so that what have been regarded as the crucial divergences do not occur at least in the very formulation of the theory, some substantial emendations to Sec. 3 are required to provide a rigorous framework for the case of a system of infinitely many degrees of freedom.

Probably the most obvious difficulty is that the space $L_2(M)$ of square-integrable functions over M is not really well defined in the infinite-dimensional case, so that the dynamical variables $R(X)$ are not operators on any well-defined state vectors. There are two approaches possible here: (i) the extension of the integration theory in function space presented in a rigorous fashion for the linear case in footnote reference 7; (ii) the adaptation of the representation-independent formalism of footnote reference 2, in which the dynamical variables are essentially elements in a well-defined algebra of observables, which however are not operators in any *ad hoc* Hilbert space (states being treated through their expectation value functionals, i.e., as suitable linear forms on the observable algebra).

The latter approach is simpler from a theoretical point of view, but it does not so readily lead to an explicit construction for the vacuum state, as does the former approach. In addition, much of what is involved in developing approach (ii) is parallel to part of the development of approach (i). It should therefore suffice here to describe (i).

The main idea is to use the approximation of the infinite system *in a physically meaningful sense* by finite systems. For example, when M is a Hilbert space, it is approximated in a way by subspaces of large finite dimension; the relevant functionals on the Hilbert space are those which are essentially carried by a finite-dimensional submanifold (depend only on a finite number of coordinates), or can be approximated by such in an invariant fashion (cf. footnote reference 2); and the relevant vector fields are principally those generating translations, and so are carried by finite subsystems. In the case of a general Hermitian manifold M we may assume, virtually as a definition of a nonpathological manifold, that it may be approximated by finite-dimensional Hermitian manifolds, in the following sense: There exist phase manifolds N of finite dimension, and maps F of M onto such an N preserving the Hermitian structure (i.e., the induced map dF from the tangent space of M onto that of N is isometric in the finite-dimensional orthocomplement of the subspace of the tangent space on which dF

vanishes), forming a "directed set"; for two of the approximations (N,F) and (N',F'), there is another approximation (N'',F') which may be interposed between M and each of the two, and being ample in the sense that for no tangent vector to M do all the dF vanish. A *tame functional* on M may then be defined as one of the form $f[F(x)]$, for some function f on N in the conventional sense. The sum and product of tame functionals is again such, and an integral on M may be defined in the manner of footnote reference 7 once we have a well-defined linear functional on the collection of all tame functionals that is appropriate—specifically, is nonnegative on nonnegative-valued functionals, and normalized to be unity on the unit function, identically one on M.

The requisite functional may be obtained from the ground state of the generalized harmonic oscillator treated in Sec. 3, assuming the approximating finite-dimensional manifolds satisfy the conditions given there. This is an intrinsic definition, and in the relativistic free-field case is known to yield the conventional theory. The $R(X)$ may then be formulated as operators in the Hilbert space $L_2(M)$ if the X are now restricted to be "tame," in the sense of being carried by a finite-dimensional manifold: for some (N,F), X corresponds to a vector field on N. This gives a covariant class of canonical variables of which the conventional ones are formal functions. In this way all relevant questions concerning analysis on M may be brought back to corresponding questions concerning approximating finite-dimensional manifolds, which do not involve any nontrivial divergences.

Probably the next most important difficulty is a purely classical and mathematical one. There is available at this time virtually no rigorous theory concerning the global solutions of a nonlinear hyperbolic equation, so that the manifold M used above of all classical solutions of Eq. (5) is a rather vague mathematical object. As noted earlier, such a manifold may be defined as an integral manifold of a certain distribution of elements of contact, which are defined by linear equations, and so accessible by existing methods. To a noteworthy degree, the manifold itself is not required, but only such tangent planes to it. On the other hand, for a complete theory, the problem of the rigorous formulation of M cannot be evaded. It is not important to formulate M as a point set, actually, but only as a certain variety of (inverse) limit of finite-dimensional manifolds.

The considerable mathematical difficulties here are in part of an altogether different character from those which seem relevant to the basic difficulties of quantum field theory. The relevant solutions of linear equations such as (6) must be expected to be not ordinary functions, nor even distributions in the sense of Schwartz, but quite highly generalized functions whose character cannot be described in an *a priori* explicit manner. These difficulties are connected with the determination of the precise character of the eigenfunctions associated with the continuous spectrum of a given linear partial differential operator, a problem which is fairly well understood and to a considerable extent resolved in the case of an elliptic operator, although not as yet in the case of a hyperbolic operator. Such rather technical problems may be avoided by the simple and physical expedient of smearing over the mass, in nonlinear analogy with the conventional treatment of the continuous spectrum through the use of packets of eigenfunctions.

The operator $\Box - p'(\phi)$ will be a self-adjoint one when properly formulated in Hilbert space, and will have a certain spectral decomposition into eigenspaces, one of which is defined by (6). If we replace this eigenspace by an eigenmanifold (=eigenspace packet) corresponding to the masses in the range $(m-\epsilon, m+\epsilon)$, we obtain a tangent space whose elements are bona fide square-integrable functions. There seems no reason to doubt that in a quite rigorous and rather straightforward sense, the corresponding distribution of elements of contact will admit an integral manifold, which will be locally a Hilbert-space of functions. Formally this manifold is obtained by joining together all of the manifolds defined by (6) with m in the range $(m-\epsilon, m+\epsilon)$; the manifold M of solutions of (6) is in a rough sense a limit of the more accessible and well-defined manifolds M_ϵ just described.[20] It may be noted incidentally that the global construction of this manifold should give, in combination with the developments of the first part of Sec. 4, concrete and nontrivial examples of quantum fields satisfying axioms similar to those axioms of Källén and Wightman[21] which do not pertain to vacuum expectation values, and in addition the canonical commutation relations for equal times.

The results of the preceding section may be summarized as

Principle III. The quantization of a given nonlinear hyperbolic partial differential equation may be accomplished by utilizing the intrinsic Hermitian structure, as a differentiable manifold, of the manifold M of all classical wave functions for the equation, in formal accordance with principle II. The infinite-dimensionality of M is dealt with by suitable approximation of M by finite-dimensional image manifolds, to which principle II is directly applicable. The field operators are among the canonical variables introduced in Sec. 3. The vacuum state is characterized as that invariant under the group of isometries of M leaving fixed the vanishing classical field, and in suitable cases may be more explicitly described as a limit of ground states of the approximating finite-dimensional systems.

It ought to be noted that *the foregoing isometry group will include effectively the Lorentz group, in the case of a*

[20] Cf. the suggestive work of Dirac in a linear case in Proc. Roy. Soc. (London) **A183**, 284 (1945).

[21] G. Källén and A. Wightman, Kgl. Danske Videnskab. Selskab. Mat.-fys. Skrifter **1**, No. 6, 58 pp, (1958).

Lorentz-invariant equation involving only real masses. Any Lorentz transformation then transforms the elements of M in a fashion leaving invariant the fundamental form and the infinitesimal complex structure, as well as the vanishing classical field, and so determines an element of the group in question. The necessity of the real mass condition is clear from the impossibility of making a covariant separation of a free field into positive and negative frequency components, in the imaginary mass case. Because of the close connection of this separation with the infinitesimal complex structure defined above, the latter will not be Lorentz-invariant in the imaginary mass case. On the other hand, the assumption that only real masses are involved is physically plausible and should be mathematically demonstrable, when suitably formulated in rigorous terms, for the relevant equations.

5. PARTICLES, INTERACTION, AND MODELS

A. Quanta of Fields

The correlation of any quantum field theory with empirical results depends in a practically essential way on the possibility of giving a particle interpretation for the theory. However, if we start with such an equation as

$$\Box\phi = m^2\phi + l\phi^3 (l \neq 0),$$

and quantize the system and obtain its physical vacuum in accordance with the preceding section, the Hilbert space of states of the incoming field is entirely determined (cf. footnote reference 2) mathematically; and it is open to considerable question whether it contains any vectors transforming like the solutions of a Klein-Gordon equation of mass m, or some other mass, let alone is equivalent to a free Bose-Einstein field of Klein-Gordon particles. The justification of an assumption of this type must at this time be essentially empirical; its success in renormalization theory validates it as a physically motivated maneuver in applied mathematics, but neither bears directly on the mathematical question involved, nor does it seem to involve a heuristic principle likely to lead into an effective mathematical development.

At this stage in the present theory we can only give a *formal* analysis of the states of the quantum field in terms of the particles whose wave functions are the tangent vectors at some fixed classical field ϕ_0; there is no mathematical reason to expect this analysis to be convergent or rigorizable, in fact there are indications for the opposite; in a sense the present theory does not so much remove the field-theoretic divergences as isolate them in the practice of giving an *ad hoc* elementary particle analysis of the states of the field.

A fixed classical field ϕ_0 may be thought of as the background field of a particular observer, who will be able to observe directly only small deviations from ϕ_0, in the first instance. That is, classically he does not observe the manifold M of all states, but rather the tangent plane T_{ϕ_0} to M at ϕ_0, the vectors of which represent fields deviating only slightly from the background field, these being the only fields his apparatus will be able to prepare, without interfering significantly with the object of his observations, i.e., without producing quantum effects. For him a quantum of the field is naturally represented by a vector in T_{ϕ_0}, and the field variables most accessible to him are notably the occupation numbers for such quanta. To set up such occupation numbers in a formal theoretical way, let us suppose that the exponential map of the tangent plane T_{ϕ_0} into the manifold M is globally without singularities and applicable to the infinite-dimensional case. Uncertain as this assumption is, it is not the most questionable assumption needed, which is that, at least locally, the measure on M obtained by transforming by the exponential map the canonical measure on T_{ϕ_0} is comparable with ("absolutely continuous with respect to") the physical vacuum measure on M. That this is a harmless assumption when M is finite-dimensional arises from the fact that any two measures compatible with the manifold structure of M are comparable (mathematically, any two measures whose null sets are invariant under translation are comparable); in the infinite-dimensional case this is very far from being true, even very "small" transformations (e.g., $x \to lx$, for any $l \neq \pm 1$) taking the free-field vacuum measure into incomparable ones (cf. Segal[22] for a rigorous treatment of this question). But if the two measures are comparable, then a development similar to that given in Sec. 3 is possible, and for every unitary transformation U in T_{ϕ_0}, there will be a corresponding transformation $\Gamma(U)$ on the state vector space of the field, the map $U \to \Gamma(U)$ being intrinsically defined, and a representation $[\Gamma(UU') = \Gamma(U)\Gamma(U')]$. Occupation numbers may then be defined as in footnote reference 2, pp. 27–31, as the infinitesimal generators of groups $\Gamma(U_t)$ for appropriate phase transformations U_t; they will then have integral proper values, annihilate the vacuum, etc.

The isometry group G_0 acts naturally as a group of linear transformations in T_{ϕ_0}, as in any Hermitian manifold; in the case of the manifold defined, e.g., by the equation $\Box\phi = m^2\phi + \phi^3$, with $\phi_0 = 0$, this includes the usual action of the Lorentz group on the real solutions of the equation $\Box\phi = m^2\phi$. If the action of G_0 is irreducible, as in this case, or more generally if disjoint invariant subspaces are orthogonal, then a complete set of group-theoretic quantum numbers of the usual variety may be set up. In this case the preceding paragraph gives in a formal way a complete analysis of the states of the field in terms of elementary particle occupation numbers, the particles being described by such quantum numbers (cf. footnote reference 2, pp. 27–31).

[22] I. E. Segal, Trans. Am. Math. Soc. 88, 12 (1958).

The *rigorous* validity of an analysis of this type is both mathematically dubious and physically somewhat counter to current lines of thought skeptical of any absolute meaning to the notion of "elementarity" of an empirical particle. In any event, the foregoing analysis appears to exhaust the formally simple particle interpretations applicable to general hyperbolic equations, although for suitable special equations an essentially rigorous notion of elementary empirical particle may conceivably exist (no solid evidence in either direction being presently known). That is to say, there may be in some cases a Lorentz-invariant transformation of the observables of a certain linear field into functions of the observables of the (interacting) field associated with M, although in even the most favorable of the cases of conventional theory, quantum electrodynamics, this seems unlikely, except as an approximation.

B. Covariant Definition of the Interaction

A puzzling feature of conventional theory has been its dependence upon an apparently artificial and unphysical separation of the total Hamiltonian (or Lagrangian) into "free-field" and "interaction" constituents (cf. e.g., van Hove[23]). Such a separation appears in the present theory as the concomitant of the classical observer's limitation to the examination of relatively small displacements from his background field ϕ_0. If ϕ_0 is time-independent, then time will act naturally in a *linear*, "noninteracting," essentially kinematical fashion on $T_{\phi 0}$; the actual dynamics, however, refers to the action of time on M, which will be *nonlinear*, when M is formally coordinatized by $T_{\phi 0}$, by the use, e.g., of the exponential map described earlier. These are classical motions; the corresponding quantum mechanical motions may be represented linearly in the function spaces over $T_{\phi 0}$ and M, respectively. The latter motion is formally equivalent to a motion in the function space over $T_{\phi 0}$, by virtue of the correspondence between $T_{\phi 0}$ and M, making the assumption of comparability of the measures involved, as in the foregoing. Thus are obtained two one-parameter groups of operators in the Hilbert space of square-integrable holomorphic functions over $T_{\phi 0}$. One of these is mathematically rather well defined, arising from the linear action of G_0 on $T_{\phi 0}$, and has as its generator the so-called "free-field" energy (relative to ϕ_0). The other is only formally defined, and in fact the available evidence rather strongly indicates that it lacks rigorous existence, arising from the nonlinear action of G_0 on M, and having as generator the "total" energy. In a formal way the S operator is thereby well defined by the conventional limit. This analysis applies both to renormalizable and nonrenormalizable theories; whether any useful numerical results can be obtained by a maneuver based on the use of partially empirical considerations depends of course, as in renormalization theory, on the equation and the ingenuity of the maneuver.

In the case $\phi_0=0$, where the background field vanishes, G_0 includes the Lorentz group when the defining partial differential equation is Lorentz-invariant and involves only real masses, and the foregoing paragraph applies then not only to translations in time but to the entire Lorentz group.

In simple conventional terms the foregoing indicates the following prescription for the separation of a total Lagrangian into "free-field" and "interaction" parts. The free-field constituent is the Lagrangian for the hyperbolic partial differential equation defining the first-order variation, in the vicinity of the vanishing field, to the manifold of all classical wave functions for the total Lagrangian.

C. Models

The short-lived character of the many attempts to classify in a systematic and economical way elementary particles on the basis of the Lorentz and conventional space-time, with or without an independent internal symmetry group, indicates that a broader attack, on a physically more conservative and theoretically more radical basis, would be desirable. One logical approach is that contemplating the use of alternative symmetry groups and/or space-time manifolds. However, if this is to have a reasonably clear-cut physical interpretation, it must be based on an adequately general field theory. The present theory, while extensively heuristic, is quite independent of the assumption that fields must be described by nonlinear partial differential equations in space-time, or, in fact, of the physical existence of quantum fields at all. Any infinite-dimensional phase manifold may be used as a basis, and other types of examples of such manifolds having symmetry groups of the proper orders of magnitudes are easily given. An example is the set of all smooth maps from a measure space into a finite-dimensional Hilbert manifold, a natural generalization of the much used linear function spaces of smooth square-integrable functions.

The fundamental symmetry group G_0 will be that leaving a designated point ϕ_0 of the basic manifold invariant. The primary elementary particles of the theory are then represented by the vectors in the irreducibly invariant subspaces of $T_{\phi 0}$ under the naturally induced action of G. Group-theoretical quantum numbers will then be definable in the fashion indicated earlier. For example, the elementary particle models described in footnote reference 5 set up certain representations and quantum numbers for symmetry groups G having the group constituted by the Lorentz group together with space-time position coordinates as a degenerate limiting case. From the present standpoint this means that G is a subgroup of the isotropy group leaving fixed the vanishing field; and the stated

[23] L. van Hove, Physica 21, 901 (1955).

action is the natural linear action of G on the infinitesimal classical fields. The description of the manifold M is necessarily a good deal more complicated than this, and is not required in the first instance for particle classification purposes.

It ought to be noted that the manifold M can in principle be built up from the knowledge of the tangent space in the vicinity of each background field. It thus has a certain conceptual quasi-empirical existence. Insofar as a relatively arbitrary classical background field can be experimentally maintained, and the response of the system to relatively arbitrary small disturbances ascertained, these tangent spaces are experimentally approximable to an arbitrary degree of accuracy, quite without any *ad hoc* assumptions as to the particle interpretation of the field, the necessity of a basic partial differential equation, etc. The structure of the tangent space at the vanishing physical field is of great interest in itself, being the basis for the classification of "free" particles; conversely, any empirical linear description of the free particles can be regarded as an approximate description of this tangent plane. A given type of conventional quantum field will in general have no direct empirical classical analog, but this may be ascribed to a lack of closure on the part of the corresponding theory. Ultimately all measurements are reducible to classical ones, and the classical analog to the field of *all* elementary particles may be considered to be the set of all classical fields, in speculative theory constructible as a manifold through the examination of the response to all possible small classical disturbances of an arbitrary background classical field. There appears to be no practical possibility of setting up a useful empirical manifold M in this fashion, but the foregoing conceptual experimentation serves at least to indicate that the manifold M has a certain fairly direct intuitional connection with physical experience that is lacking in the Lagrangian.

The *quantum*, as contrasted with the *classical*, field, plays primarily a formal part in our analysis, and serves mainly only to connect the present formulation with the conventional one. Quite without its use the total energy of the field and the vacuum state, e.g., are well-defined (through the use of principle III). In view of the apparently inevitably dubiously physical character of the quantum field, the possibility that it may well be theoretically expendable is not very surprising.

ACKNOWLEDGMENTS

We are much indebted to the following mathematicians and physicists for informative and stimulating conversations: S. Helgason, D. Shale, W. F. Stinespring; K. Gottfried, W. Heisenberg, G. Källén, L. Rosenfeld.

Kemmer Wave Equation in Riemann Space

ADEL DA SILVEIRA
Colégio Pedro II and NEPEC, Rio de Janeiro, E. G., Brazil
(Received July 1, 1960)

The Kemmer equation is written in Riemann space. The form of the equation, the supplementary relations and the form of the covariant derivatives of the operators β^μ are considered. As a check of the equation it is shown that the equations for the values zero and one of the spin can be obtained from this generalized Kemmer equation with the help of generalized Fujiwara operators. In particular the equation for photons in interaction with the gravitational field is obtained.

I. INTRODUCTION

THE relativistic wave equation of the electron in Riemann space was analyzed recently by P. A. M. Dirac.[1] Special attention was given to the covariance of the equation under the rotations of the four-legs h_r whose components $h_{r\mu}$ are connected with the components $g_{\mu\nu}$ of the metric tensor by means of the well-known relations

$$h_\mu{}^r h_{r\nu}=g_{\mu\nu}.$$

The relativistic electron wave equation in Riemann space was studied by several authors and is different from the Dirac equation of special relativity. One has

$$\gamma^\mu[(\partial/\partial x^\mu)-\Gamma_\mu]\psi+m\psi=0,$$

where γ^μ are matrices which depend on the coordinates. The matrices Γ_μ are traditionally obtained from the value of the covariant derivative of the γ^μ.[2-4]

Another method of handling the theory was indicated by Dirac in the paper previously mentioned. He assumed a form for the equation and a connection between the Lorentz coefficients and the matrix of transformation of the wave function. The form of the matrix Γ_μ was assumed independently of any other hypothesis and the covariance of the equation was then proved.

This procedure of Dirac can be applied to the formulation of the Kemmer equation for spin-zero and spin-one particles[5] in Riemann spaces. In this case, as we will show in the present paper, the matrix corresponding to Γ_μ is null, this value being a consequence of conditions that one can impose in this case. It is worth mentioning that these conditions cannot be admitted in the Dirac theory.

The physical interpretation and the supplementary conditions are indicated in Sec. III. The separation of the equations for the two values of the spin is made with the help of generalized Fujiwara operators[6,7] and is considered in Sec. IV.

[1] P. A. M. Dirac, Festschr. Max Planck **1958**, p. 339.
[2] W. Pauli, Ann. Physik **18**, 337 (1933).
[3] B. S. De Witt and C. M. De Witt, Phys. Rev. **87**, 116 (1952).
[4] D. R. Brill and J. A. Wheeler, Revs. Modern Phys. **29**, 465 (1957).
[5] N. Kemmer, Proc. Roy. Soc. (London) **A173**, 91 (1939).
[6] I. Fujiwara, Progr. Theoret. Phys. Kyoto **10**, 6 (1953).
[7] H. Umezawa, *Quantum Field Theory* (North-Holland Publishing Company, Amsterdam, 1956), Chap. V.

II. KEMMER EQUATION

The relativistic wave equation of Kemmer for spin-zero and spin-one particles has the form

$$\beta^r\partial\psi/\partial x^r+m\psi=0, \tag{1}$$

where $x'=x$, $x^2=y$, $x^3=z$, $x^4=ict$ and the matrices β^r satisfy the relations

$$\beta^r\beta^s\beta^t+\beta^t\beta^s\beta^r=\beta^r\delta^{st}+\beta^t\delta^{sr}, \tag{2}$$

where δ^{st} is the Kronecker delta.

As in the case of the Dirac theory, the Lorentz transformation with coefficients $a_s{}^r$ is connected with a transformation of the wave function

$$\psi=S\psi'.$$

For the covariance of the equation, we assume the following relation between such an S and the $a_s{}^r$:

$$a_t{}^r S^{-1}\beta^t S=\beta^r. \tag{3}$$

The introduction of a phase factor in S does not alter the relation (3). The matrix S is not determined, however, uniquely by (3), even if we choose the value of this factor. We can consider, therefore, the following conditions, instead of (3):

$$a_s{}^r S^+\eta^4\beta^s S=\eta^4\beta^r, \tag{4a}$$

$$S^+\eta^4 S=S\eta^4 S^+=\eta^4, \tag{4b}$$

$$a_s{}^r S^+\eta^4\eta^s S=\eta^4\eta^r,$$

where S^+ is the H.C. of S.

The η^r ($r=1, 2, 3$) are matrices of square one. We will assume, however, that

$$(\eta^4)^2=-I.$$

In the Dirac theory there are relations analogous to (4a) and (4b):

$$a_s{}^r S^+\gamma^4\gamma^s S=\gamma^4\gamma^r, \quad S^+\gamma^4 S=S\gamma^4 S^+=\gamma^4.$$

These conditions can be obtained from

$$a_s{}^r S^{-1}\gamma^s S=\gamma^r$$

on account of reality conditions of the $a_s{}^r$ and with the help of the Schur lemma.[8] In the case of the Kemmer theory however, there are three matrices that commute

[8] R. H. Good, Jr., Revs. Modern Phys. **27**, 187 (1955).

with all the β^r so Eqs. (4a) and (4b) are not equivalent to (3).

In order to write the Kemmer equation in the Riemann spaces, we will introduce a set of four orthonormal vectors $h_{r\mu}$. The Latin indices characterize the vectors and the Greek ones are connected with their components. These components, on the other hand, are related with the components of the metric tensor $g_{\mu\nu}$ by means of

$$h_\mu{}^r h_{r\nu} = g_{\mu\nu}. \tag{5a}$$

The orthonormal relations are

$$h_r{}^\mu h_{s\mu} = \delta_{rs}. \tag{5b}$$

A rotation of the four-legs characterized by $a_s{}^r$ with r and s running from one to four, alters their components according to

$$h_r{}'^\mu = a_r{}^s h_s{}^\mu. \tag{7}$$

Let us assume now the following form for the Kemmer equations in Riemann space:

$$h_r{}^\mu \beta^r \psi_{,\mu} + m\psi = 0, \tag{7}$$

where the comma denotes partial derivative, as usual. This equation is evidently covariant under general transformation of coordinates because of its tensor form with respect to μ. To demonstrate its covariance under the rotations of the $h_{r\mu}$, let us consider the following transformation

$$\psi = S\psi', \tag{8}$$

associated with the rotation. S is now a function of the position and the relations

$$h_r{}^\mu S^+ \eta^4 \beta^r S = h_r{}'^\mu \eta^4 \beta^r, \tag{9a}$$

$$h_s{}^\mu S^+ \eta^4 \eta^s S = h_r{}'^\mu \eta^4 \eta^r, \tag{9b}$$

among the components $h_r{}^\mu$ and $h_r{}'^\mu$ before and after the transformation, are consequences of (4a), (4c), and (6).

If we substitute (8) and (7) and take (4b) and (9a) into account, we obtain

$$h_r{}'^\mu \beta^r (\psi_{,\mu}{}' - \eta^4 S^+ \eta^4 S_{,\mu} \psi') + m\psi' = 0.$$

To prove the covariance, one has to demonstrate that

$$\eta^4 S^+ \eta^4 S_{,\mu} = 0.$$

Let us consider now the quantities $\Gamma_\mu{}'$ defined by

$$\Gamma_\mu{}' = \tfrac{1}{2} h_a{}'^\rho h_{b\rho;\mu}{}' \eta^a \eta^b. \tag{10}$$

The semicolon denotes covariant derivative with respect to x^μ. We have

$$\eta^4 \Gamma_\mu{}' = -\tfrac{1}{2} h_a{}'^\rho h_{b\rho;\mu}{}' \eta^4 \eta^a \eta^4 \eta^4 \eta^b = -\tfrac{1}{2} h_a{}^\rho S^+ \eta^4 \eta^a S \eta^4 (h_{b\rho} S^+ \eta^4 \eta^b S)_{;\mu} \tag{11}$$

on account of (9b). Quantities with Latin indices are scalars with respect to coordinate transformations. The covariant derivative indicated in (11) has three terms. The first is zero because

$$S \eta^4 S^+{}_{,\mu} \tag{12}$$

is necessarily anti-Hermitian, on account of (4b). As this expression (12) contains η^4, it is equal to the sum of the matrices $\eta^4\beta^i$, $\eta^4\eta^i\eta^j\eta^k$, $\beta^4\beta^i\beta^j$ multiplied by convenient coefficients, the indices i, j, k, and 4 being all different. In view of the properties

$$\eta^i \beta^k = -\beta^k \eta^i$$

for i different from k, and

$$\beta^i \eta^i = \eta^i \beta^i = \beta^i,$$

for equal indices (without sum), we have, in view of (5b),

$$\eta^4 \eta^a S \eta^4 S^+{}_{,\mu} \eta^4 \eta_a = 0.$$

The other terms of the covariant derivative of (11) can be evaluated with the help of (4b) and (5b) and the values of the squares of η^r. We have, therefore,

$$\eta^4 \Gamma_\mu{}' = S^+ \eta^4 S_{,\mu} + \eta^4 \Gamma_\mu, \tag{13}$$

where Γ_μ is the analog of $\Gamma_\mu{}'$ in the unprimed system.

On the other hand, we have in view of the commutation properties of the η^r and the null value of the covariant derivative of the metric tensor $g_{\mu\nu}$ with respect to any x^σ:

$$\Gamma_\mu{}' = \tfrac{1}{4} (h_a{}'^\rho h_{b\rho;\mu}{}' + h_b{}'^\rho h_{a\rho;\mu}{}') \eta^4 \eta^b = 0.$$

The value of Γ_μ can be calculated in the same way. Therefore, we can write, because of (13),

$$S^+ \eta^4 S_{,\mu} = 0.$$

III. PHYSICAL INTERPRETATION AND SUPPLEMENTARY CONDITIONS

The form of the generalized Kemmer equation (7) suggests the value of the covariant derivative of the β^μ defined by

$$\beta^\mu{}_{;\sigma} = h_r{}^\mu \beta^r. \tag{15}$$

The form is simpler than the corresponding form of the γ^μ and is

$$\beta^\mu{}_{;\sigma} = \beta^\mu{}_{,\sigma} + \left\{ \begin{matrix} \mu \\ \sigma\alpha \end{matrix} \right\} \beta^\alpha. \tag{16}$$

The value of this derivative can be taken as

$$\beta^\mu{}_{;\sigma} = 0 \tag{17}$$

and is covariant under the rotations of the $h_r{}^\mu$ as we can see easily with the help of (9a) and (14). Equation (14) affords the introduction of the current density j^μ as well as the proof of its conservation law. Let us consider the quantity

$$j^\mu = h_r{}^\mu \psi^+ \eta^4 \beta^r \psi = \psi^+ \eta^4 \beta^\mu \psi.$$

These j^μ behave like a vector under coordinate transformations and do not alter under rotations of the

$h_r{}^\mu$ for, in view of (9a) and (4a), we have

$$h_r{}^\mu\psi^+\eta^4\beta^\mu S=h_r{}'^\mu\psi'^+\eta^4\beta^r\psi'.$$

The conservation law

$$j^\mu{}_{;\mu}=0$$

is a consequence of (7), (18), and of the adjoint equation

$$(\partial\bar{\psi}/\partial x^\mu)\beta^\mu-m\bar{\psi}=0,$$

where

$$\bar{\psi}=\psi^+\eta^4.$$

Another consequence of the value of the covariant derivative of the β^μ is an additional relation analogous to Eq. (6) of the Kemmer paper. If we multiply the wave equation (7) by $\beta^\lambda\beta_\alpha\partial_\lambda$ and take account of (16) and (17), we have

$$\partial_\alpha\psi=\beta^\lambda\beta_\alpha\partial_\lambda\psi.$$

This proof depends strongly on the hypothesis that the mass of the particle is different from zero.

IV. SEPARATED WAVE EQUATIONS

(a) Spin Zero

To obtain the wave equation for zero-spin particles, we shall introduce the operators[6]

$$P=(\beta_1)^2(\beta_2)^2(\beta_3)^2(\beta_4)^2,\quad P_r=P\beta_r, \tag{18}$$

and

$$P_\mu=P\beta_\mu=h_\mu{}^rP\beta_r.$$

The P is independent of the position. We shall assume as in the case of the β^μ that

$$P^\nu{}_{;\sigma}=P^\nu{}_{,\sigma}+\left\{\begin{matrix}\nu\\ \sigma\alpha\end{matrix}\right\}P^\alpha=0.$$

The relation

$$P_\mu\beta_\nu=Pg_{\mu\nu}, \tag{20}$$

is a consequence of (5a), (15), and of

$$P_r\beta_s=P\delta_{rs}. \tag{21}$$

The first equation for these particles is obtained by multiplying Eq. (7) by P. We have, because of (16) and (17),

$$\partial_\mu(P\beta^\mu\psi)+\left\{\begin{matrix}\mu\\ \mu\alpha\end{matrix}\right\}P\beta^\alpha\psi+mP\psi=0. \tag{22}$$

If we introduce the notation

$$P\beta^\mu\psi=U^\mu,\quad\text{and}\quad P\psi=U,$$

we can write (22) in the form

$$\partial_\mu U^\mu+\left\{\begin{matrix}\mu\\ \mu\alpha\end{matrix}\right\}U^\alpha+mU=0.$$

We get the second equation by multiplying Eq. (7) by P_ν, taking into account the value of $P_{;\sigma}{}^\nu$ and by applying (20) and (18). We obtain

$$\partial_\nu U+mU_\nu=0.$$

(b) Spin One

In this case we shall introduce the operator

$$R_\mu=h_{r\mu}R^r,$$

where R^r is defined by the relations

$$R^r=-(\beta_1)^2(\beta_2)^2(\beta_3)^2\beta^r\beta_4\quad\text{for } r=1,2,3,$$

$$R^r=(\beta_1)^2(\beta_2)^2(\beta_3)^2(1-\beta_4)^2\quad\text{for } r=4.$$

The following properties can easily be proved

$$R_\mu\beta_\nu=-R_\nu\beta_\mu, \tag{23}$$

$$R_\mu\beta_\nu\beta_\lambda=g_{\nu\lambda}R_\mu-g_{\mu\lambda}R_\nu. \tag{24}$$

The relations (16), (17), (23), and (24) are consistent with the following definition:

$$\partial_\alpha R_\mu-\left\{\begin{matrix}\sigma\\ \alpha\mu\end{matrix}\right\}R_\sigma=0. \tag{25}$$

We obtain the first equation by applying R_ν to Eq. (7), if we take into account (16), (17), and (25) as well as the equation

$$g^{\mu\alpha}{}_{;\nu}=0.$$

We obtain

$$g^{\mu\alpha}\partial_\mu F_{\nu\alpha}-g^{\mu\epsilon}\left\{\begin{matrix}\alpha\\ \epsilon\mu\end{matrix}\right\}F_{\nu\alpha}-g^{\mu\alpha}\left\{\begin{matrix}\sigma\\ \mu\nu\end{matrix}\right\}F_{\sigma\alpha}+m^2U_\nu=0 \tag{26a}$$

or

$$g^{\mu\alpha}F_{\nu\alpha;\mu}+m^2U_\nu=0, \tag{26b}$$

where the following notation was used:

$$U_\alpha=R_\alpha\psi,\quad F_{\nu\alpha}=mR_\nu\beta_\alpha\psi.$$

The second equation results from the application of the operator $R_\nu\beta_\lambda$ to Eq. (7). If we use (16), (17), (24), and (25), we obtain

$$F_{\nu\lambda}=\partial_\nu U_\lambda-\partial_\lambda U_\nu. \tag{27}$$

Equations (26) and (27) are to be used for particles of spin one in interaction with gravitation, if the mass of the particle is different from zero. We can assume now Eqs. (26) and (27), independently of the process that we have used to obtain them, for the analogous equations in the free case can be introduced independently of the Kemmer equation. In this case, we have equations for photons in interaction with gravitation if we take the mass as being zero.

ACKNOWLEDGMENTS

The author is indebted to Dr. Armando D. Tavares, Dr. Aldízio F. Costa, Dr. Carlos M. do Amaral, and Dr. Edson Rodrigues for valuable discussions.

Hamiltonian Formalism and the Canonical Commutation Relations in Quantum Field Theory*

H. ARAKI†
Palmer Physical Laboratory, Princeton University, Princeton, New Jersey
(Received June 6, 1960)

Cyclic representations of the canonical commutation relations and their connection with the Hamiltonian formalism are studied. The vacuum expectation functional $E(f)=(\Psi_0,e^{i\varphi(f)}\Psi_0)$ turns out to be a very convenient tool for the discussion. The uniqueness of a translationally invariant state (vacuum) is proved under the assumption of the cluster decomposition property for $E(f)$. The existence and near uniqueness of the Hamiltonian in cyclic representations of the canonical commutation relations are established. The conditions for the relativistic invariance of the theory are stated in terms of vacuum expectation values at a fixed time. It is shown that $E(f)$ is the Fourier transform of a quasi-invariant nonnegative measure on the space of all linear functionals of the test functions.

1. INTRODUCTION

IN this paper we shall discuss Hamiltonian formalism and canonical commutation relations for a self-interacting neutral scalar field in a mathematically rigorous way. Generalizations to one or more tensor fields are easy and do not add any essentially new feature. Not so obvious is the extension of the formalism to the case of spinor fields, which is outside the scope of this paper.

In the conventional approach, one takes the field $\varphi(\mathbf{x})$ and its canonical conjugate $\pi(\mathbf{x})$ at time $t=0$ as the basic variables which allow a complete description of the system, and one assumes the canonical commutation relations

$$[\varphi(\mathbf{x}),\varphi(\mathbf{x}')]=[\pi(\mathbf{x}),\pi(\mathbf{x}')]=0,$$
$$[\varphi(\mathbf{x}),\pi(\mathbf{x}')]=i\delta(\mathbf{x}-\mathbf{x}'). \tag{1.1}$$

The time-development of the system is determined by the Hamiltonian which is given as a function of φ and π,

$$H=H(\varphi,\pi). \tag{1.2}$$

It has been pointed out[1] that the relations (1.1) are not sufficient to define the field operators (up to unitary equivalence) even if irreducibility[2] is assumed, in contrast with the case of particle mechanics.[3] There is an immense manifold of inequivalent representations of the relations (1.1) and therefore the problem arises: Which one of them is appropriate for a particular model? By "model" we mean an explicit expression (1.2) for the Hamiltonian. The point is that a particular expression for the Hamiltonian will define a bona fide operator only if an appropriate representation of the canonical variables is chosen.[4,5]

There have been various mathematically rigorous attempts at establishing a classification scheme for inequivalent representations of (1.1).[6–9] There has also been a heuristic discussion of the relation between the form of the Hamiltonian and the appropriate representation.[10] Here we intend to treat this latter problem in a mathematically rigorous fashion, restricting our discussion to cyclic representations of the canonical variables. (The cyclicity as well as the irreducibility are discussed in Sec. 3.)

The main objective of this paper is to prove that, under certain conditions, the vacuum expectation value of $\exp i\int\varphi(\mathbf{x})f(\mathbf{x})d\mathbf{x}$, which is denoted by $E(f)$, is sufficient to determine (up to unitary equivalence) the Hilbert space (Sec. 4), the representation of both φ (Sec. 4) and π (Sec. 7) in this Hilbert space, the vacuum state (namely, the unique translationally invariant state, see Sec. 6), the form of the Hamiltonian (Sec. 8), and the representations of transformation groups (Euclidean group in Sec. 5 and Lorentz group in Sec. 9). The positive definiteness of the Hilbert space as well as the invariance requirements will give restrictions on $E(f)$. It is also shown in Sec. 6 that, for a given cyclic representation of φ, there is at most one $E(f)$ satisfying translational invariance and cluster decomposition property and hence the representation of φ

* Supported in part by the Air Force Office of Scientific Research, Air Research and Development Command. Condensed from a part of the thesis submitted to the Faculty of Princeton University for partial fulfillment of Ph.D. requirements.

† Present address: Department of Nuclear Engineering, Kyoto University, Kyoto, Japan.

[1] K. O. Friedrichs, Communs. Pure Appl. Math. 5, 367, 383 (1952); L. van Hove, Physica 18, 145 (1950); A. S. Wightman and S. S. Schweber, Phys. Rev. 98, 812 (1955).

[2] See Sec. 3.

[3] J. von Neumann, Math. Ann. 104, 570 (1931).

[4] If the Hamiltonian is invariant under spatial translations and has a nondegenerate discrete eigenstate which is different from the free vacuum, then the Fock representation is not appropriate for this Hamiltonian. See R. Haag, Kgl. Danske Videnskab. Selskab, Math.-fys. Medd. 29, No. 12 (1955).

[5] Apart from the problem of choosing an appropriate representation, there may be other problems to be settled before one can give an exact meaning to a given formal expression for the Hamiltonian. For example, the expression $\int\varphi(\mathbf{x})^2d\mathbf{x}$ is by itself not well defined.

[6] L. Gårding and A. S. Wightman, Proc. Natl. Acad. Sci. U. S. 40, 622 (1954).

[7] I. E. Segal, Trans. Am. Math. Soc. 88, 12 (1958).

[8] J. Lew, thesis, Princeton University, Princeton, New Jersey, 1960.

[9] H. Fukutome, Progr. Theoret. Phys. (Kyoto) 23, 989 (1960).

[10] F. Coester and R. Haag, Phys. Rev. 117, 1137 (1960).

also determines π, the vacuum state, the Hamiltonian, and the representation of transformation groups.

In Sec. 10, we shall show that $E(f)$ is the Fourier transform of a quasi-invariant positive measure. This establishes a connection between the work of Segal,[7] Lew,[8] and Fukutome[9] and our approach. The main result of this last section was also obtained by Lew[8] and by Fukutome.[9]

Applications of the present formalism to special models will be treated in a separate paper.

2. DEFINITION OF REPRESENTATIONS OF THE CANONICAL COMMUTATION RELATIONS[11]

The quantities $\varphi(\mathbf{x})$ and $\pi(\mathbf{x})$ cannot be considered as ordinary operators because of the δ function in (1.1). Hence we consider the smoothed-out fields

$$\varphi(f)=\int \varphi(\mathbf{x})f(\mathbf{x})d\mathbf{x}, \tag{2.1}$$

$$\pi(g)=\int \pi(\mathbf{x})g(\mathbf{x})d\mathbf{x}, \tag{2.2}$$

where $f(\mathbf{x})$ and $g(\mathbf{x})$ are real-valued functions belonging to certain classes of functions, D_1 and D_2, respectively. Equation (1.1) implies

$$[\varphi(f_1),\varphi(f_2)]=[\pi(g_1),\pi(g_2)]=0, \tag{2.3}$$

$$[\varphi(f),\pi(g)]=i(f,g), \tag{2.4}$$

where

$$(f,g)=\int f(\mathbf{x})g(\mathbf{x})d\mathbf{x}. \tag{2.5}$$

In order to avoid complications due to the domain question of the unbounded operators $\varphi(f)$ and $\pi(g)$, we introduce the unitary operators[12]

$$U(f)=\exp i\varphi(f), \tag{2.6}$$

$$V(g)=\exp i\pi(g). \tag{2.7}$$

The linear dependence of $\varphi(f)$ and $\pi(g)$ on f and g together with (2.3) implies

$$U(f_1)U(f_2)=U(f_1+f_2), \tag{2.8}$$

$$V(g_1)V(g_2)=V(g_1+g_2). \tag{2.9}$$

Equation (2.4) is replaced by

$$U(f)V(g)=V(g)U(f)\exp -i(f,g). \tag{2.10}$$

These three relations together with the unitarity of $U(f)$ and $V(g)$ imply

$$U(0)=V(0)=1, \tag{2.11}$$

$$U(f)^*=U(-f), \tag{2.12}$$

$$V(g)^*=V(-g). \tag{2.13}$$

[11] Cf. also footnote references 6 through 9.

[12] This formulation of the canonical commutation relations was first used by H. Weyl [Z. Physik **46**, 1 (1927)].

Furthermore, $U(tf)$ and $V(tg)$ are continuous in the real variable t.

Conversely, if one has $U(f)$ and $V(g)$ satisfying (2.8) through (2.13) and if $U(tf)$ and $V(tg)$ are continuous in t, then the infinitesimal generators defined by

$$\varphi(f)=\frac{1}{i}\frac{d}{dt}U(tf)\Big|_{t=0}, \tag{2.14}$$

$$\pi(g)=\frac{1}{i}\frac{d}{dt}V(tg)\Big|_{t=0}, \tag{2.15}$$

are linear in f and g, self-adjoint, and satisfy (2.3) and (2.4) on a dense domain of the Hilbert space.[13]

Thus, we define a representation of the canonical commutation relations in the following way.

Definition

Given two real linear spaces D_1 and D_2 with an inner product (f,g) defined between an arbitrary $f\epsilon D_1$ and an arbitrary $g\epsilon D_2$. By a representation of the canonical commutation relations for the pair (D_1, D_2), we mean a pair of mapping U and V from D_1 and D_2, respectively, into the set of unitary operators in a Hilbert space such that (2.8) through (2.10) [and hence, (2.11) through (2.13)] are satisfied for f, f_1, $f_2\epsilon D_1$ and g, g_1, $g_2\epsilon D_2$, and such that $U(tf)$ and $V(tg)$ are weakly continuous in t for fixed $f\epsilon D_1$ and $g\epsilon D_2$.

For unitary operators, the weak continuity assumed here is equivalent to strong continuity. We also note that

$$U(\sum_{i=1}^{n} t_i f_i) \quad \text{and} \quad V(\sum_{i=1}^{n} t_i g_i)$$

are continuous in t_i for fixed $f_i\epsilon D_1$ and $g_i\epsilon D_2$ $(i=1,\cdots,n)$.

The elements of D_1 and D_2 are called test functions for the field and its conjugate. In the following, we do not specify D_1 and D_2. For a theory of a self-interacting neutral scalar field, they are sufficiently large classes of sufficiently smooth, real functions on three-dimensional space and one expects that the exact choice of D_1 and D_2 is immaterial for the physical content of the theory.[14] For a theory of several interacting fields, we can take as D_1 and D_2 the direct sum of the D_1 and D_2 for individual fields. We note that the work of Gårding and Wightman[6] refers to the case where $D_1=D_2$ consists of all finite linear combinations of a chosen basis f_i.

3. CYCLICITY AND IRREDUCIBILITY

In this section we will define cyclicity and irreducibility of a representation and discuss their physical meaning. The cyclicity of representations will be assumed in later discussions.

[13] J. von Neumann, Math. Z. **30**, 3 (1929); L. Gårding, Proc. Natl. Acad. Sci. U. S. **33**, 331 (1947). It is not known whether one can take the same domain for all f and g.

[14] The choice may be important in order to obtain a convenient mathematical scheme.

We denote by $\mathfrak{U}$, $\mathfrak{V}$, and $\mathfrak{UV}$, the sets of operators $\{U(f)|f\epsilon D_1\}$, $\{V(g)|g\epsilon D_2\}$, and $\{U(f)V(g)|f\epsilon D_1, g\epsilon D_2\}$, respectively. We denote by $\mathfrak{A}(X)$ the algebra generated by the operators of a set X and the unit operator. The most general element of $\mathfrak{A}(\mathfrak{U})$, for example, is

$$\sum_{i=1}^{n} c_i U(f_i), \tag{3.1}$$

where $f_i\epsilon D_1$ and the c_i are complex numbers. We denote by $\mathfrak{A}(X)\Psi_0$ the linear subset of the total Hilbert space $\mathfrak{H}$ which consists of all the vectors obtained by the application of the operators in $\mathfrak{A}(X)$ on a fixed vector Ψ_0. We denote by $\mathfrak{H}(X,\Psi_0)$ the closure of $\mathfrak{A}(X)\Psi_0$.

If $\mathfrak{H}(X,\Psi_0)$ is the total space $\mathfrak{H}$, then the vector Ψ_0 is said to be cyclic relative to X. If there is at least one cyclic vector relative to X, then the Hilbert space is said to be cyclic relative to X. Thus, a cyclic representation of $\mathfrak{U}$ is a representation of $\mathfrak{U}$ where the whole Hilbert space is spanned by the vectors of the form

$$\sum_{i=1}^{n} c_i U(f_i)\Psi_0.$$

Later, we are going to assume that the vacuum state is a cyclic vector.

The commutant X' of a set X of bounded operators is the set of all the bounded operators which commute with all the elements of X. The bicommutant $(X')'$ will be denoted by X''. X is contained in X'' and $(X')''$ is always equal to X'. If X is a self-adjoint set and is equal to X'', X is called a W^*-algebra (ring of operators in the original terminology of von Neumann). It is known[15] that for a self-adjoint set X containing the unit operator, X'' is the weak and at the same time the strong closure of X. Thus, in the case of a self-adjoint set X one may say that X'' consists of functions of the operators in X.

A self-adjoint set X of bounded operators is said to be maximal Abelian if $X'=X''$. In this case X is obviously Abelian and there are no bounded operators other than function of its elements, which can be added to X to form a larger set of commuting bounded operators. Therefore, this is what Dirac calls a complete set of commuting observables.[16]

The following theorem relates a complete set of commuting observables with cyclicity of the Hilbert space relative to a set of commuting bounded operators, giving the latter a physical meaning.

Theorem 3.1

If a self-adjoint set X of commuting bounded operators has a cyclic vector, then X is maximal Abelian. Conversely, if X is maximal Abelian and if the Hilbert space is separable, then X has a cyclic vector. (See Segal[17] for the proof.)

The relation between the most general representation and cyclic representations can be understood through the Gelfand's theorem[18] which states that any representation of a c^*-algebra is a direct sum of cyclic representations.[19]

A subspace $\mathfrak{H}_1$ of a Hilbert space $\mathfrak{H}$ is called invariant under a set X of linear operators if, for any $\Psi\epsilon\mathfrak{H}_1$ and $A\epsilon X$, $A\Psi$ is in $\mathfrak{H}_1$. A Hilbert space is said to be irreducible relative to X, if it has no invariant proper subspace relative to X.

If a Hilbert space $\mathfrak{H}$ is irreducible relative to a set X of bounded operators, then every vector is cyclic relative to X. The converse is also true. A necessary and sufficient condition for the irreducibility relative to a self-adjoint set of bounded operators is that X' consists only of multiples of the identity operator (Schur's lemma.[20]) This lemma shows that if the space is irreducible relative to X, then every operator is expressible in terms of the elements of X.

Let us call a self-adjoint set X of bounded operators in a Hilbert space $\mathfrak{H}$ a "determining set of observables" if the knowledge of the expectation values of all the operators of $\mathfrak{A}(X)$ in a state determines the state up to a multiplicative c number, namely, if

$$(\Phi,R\Phi)/(\Phi,\Phi)=(\Psi,R\Psi)/(\Psi,\Psi)$$

for all $R\epsilon\mathfrak{A}(X)$ implies $\Phi=\lambda\Psi$ with some complex number λ.

The following theorem states the equivalence of the notion of irreducibility with that of a "determining set of observables," giving the former a physical meaning.

Theorem 3.2

A self-adjoint set X of bounded operators is determining if and only if the space is irreducible relative to X.[21]

An irreducible representation of the canonical commutation relations is necessarily cyclic relative to $\mathfrak{UV}$. However, it is not necessarily cyclic relative to $\mathfrak{U}$, nor does the cyclicity relative to $\mathfrak{U}$ imply the irreducibility

---

[15] For example, J. Diximier, *Les Algebres d'operateurs dans l'espace Hilbertien* (Gauthier-Villars, Paris, France, 1957), p. 44, corollaire 1.

[16] P. Dirac, *The Principle of Quantum Mechanics* (Oxford University Press, New York, 1947), 3rd ed., p. 57.

[17] I. Segal, Mem. Am. Math. Soc. No. 9, II (1951), corollaries 1.1 and 1.2.

[18] M. Neumark, *Sowjetische Arbeiten zur Funktionalanalysis* (Verlag Kultur und Fortschritt, Berlin, 1954); "Involutive Algebren," p. 114, Sec. 5.2, IV.

[19] A c^*-algebra is a complete normed algebra with adjoint operation. Although $\mathfrak{A}(\mathfrak{U})$ is not a c^*-algebra, one can assign a unique (uniform) norm to the elements of $\mathfrak{A}(\mathfrak{U})$ due to the presence of the operators $V(g)$ (assuming that D_2 separates D_1; see footnote reference 8). Hence, one can consider the completion $c^*(\mathfrak{U})$ of $\mathfrak{A}(\mathfrak{U})$ relative to this norm instead of considering $\mathfrak{A}(\mathfrak{U})$. Note that the cyclicity relative to $\mathfrak{A}(\mathfrak{U})$ is equivalent to that relative to $c^*(\mathfrak{U})$.

[20] M. Neumark, footnote reference 18, p. 114, Sec. 5.3, V.

[21] M. Neumark, reference 18, p. 122, Sec. 6.4, theorem 4; H. Araki, thesis, Princeton University, 1960, p. 25.

relative to $\mathfrak{U}\mathfrak{B}$.[22] Irreducibility relative to $\mathfrak{U}\mathfrak{B}$ is automatic for a cyclic representation with a cyclic vector Ψ_0 if the Hamiltonian H is a function of $\mathfrak{U}\mathfrak{B}$[23] and if Ψ_0 is a nondegenerate eigenvector of H.

4. CYCLIC REPRESENTATIONS AND VACUUM EXPECTATION VALUES

In this section, we will relate the problem of cyclic representations with the problem of vacuum expectation values.[8,10,24]

For any normalized vector Ψ_0 and for any self-adjoint set X of bounded operators, the functional on $\mathfrak{A}(X)$ defined by

$$E(R)=(\Psi_0,R\Psi_0), \quad R\epsilon\mathfrak{A}(X) \tag{4.1}$$

satisfies the following conditions,

$$E(\lambda_1R_1+\lambda_2R_2)=\lambda_1E(R_1)+\lambda_2E(R_2), \tag{4.2}$$

$$E(R^*)=E(R)^*, \tag{4.3}$$

$$E(R^*R)\geqslant 0, \tag{4.4}$$

$$E(1)=1, \tag{4.5}$$

where λ_1 and λ_2 are complex numbers and R_1, R_2, and R belong to $\mathfrak{A}(X)$.

The following theorem states the converse.

Theorem 4.1

For a given self-adjoint algebra $\mathfrak{A}$ containing 1, and for a given linear functional $E(R)$ on $\mathfrak{A}$ satisfying (4.2) through (4.5), there exists a Hilbert space $\mathfrak{H}$, a normalized vector Ψ_0 in $\mathfrak{H}$ and a cyclic representation of $\mathfrak{A}$ by operators (not necessarily bounded) defined on a dense domain containing the cyclic vector Ψ_0, such that (4.1) is satisfied.[25]

Applying the foregoing theorem to the representation of the canonical commutation relations, we get the following two theorems. It should be noted that, in these cases, the algebra $\mathfrak{A}$ is automatically represented by bounded operators.

Theorem 4.2

The necessary and sufficient condition for a functional $E(f)$, $f\epsilon D_1$, to be represented as

$$E(f)=(\Psi_0,U(f)\Psi_0) \tag{4.6}$$

with a cyclic normalized vector Ψ_0 in a Hilbert space $\mathfrak{H}$ and with a representation of $\mathfrak{U}$ on $\mathfrak{H}$ [namely, unitary operators $U(f)$ on $\mathfrak{H}$ satisfying (2.8)] is that

$$E(f)^*=E(-f), \tag{4.7}$$

$$E(0)=1, \tag{4.8}$$

$$\sum_{i,j=1}^{n} c_ic_j^*E(f_i-f_j)\geqslant 0, \tag{4.9}$$

for an arbitrary integer n, an arbitrary set of complex numbers c_i, $i=1,\cdots,n$ and an arbitrary set of elements f_i in D_1. In order that $U(f)$ is continuous on f relative to some topology in D_1 [for example, the continuity of $U(tf)$ with respect to the real variable t], it is necessary and sufficient that $E(f+f_0)$ be continuous on f for a fixed f_0. Such a representation of $\mathfrak{U}$ with Ψ_0 as a cyclic vector is unique up to unitary equivalence.

For the proof, we apply theorem 4.1 takine $\mathfrak{A}=\mathfrak{A}(\mathfrak{U})$. As the functional $E(R)$, we use $E(f)$ for $R=U(f)$ and Eq. (4.2) for general $R\epsilon\mathfrak{A}(\mathfrak{U})$. Equation (4.9) corresponds to (4.4), and (4.8) corresponds to (4.3) and the unitarity of $U(f)$. To prove the continuity of $U(f)$, suppose $f_n\to f$. If $\chi\epsilon\mathfrak{A}(\mathfrak{U})\Psi_0$, $\Psi_n=U(f_n)\chi$, and $\Psi=U(f)\chi$, then from the assumed continuity of $E(f)$, we see that $(\Psi_n,\Psi)\to(\Psi,\Psi)$. Since $\|\Psi_n\|=\|\Psi\|=\|\chi\|$, we have $\|\Psi_n-\Psi\|\to 0$. For general $\Phi\epsilon\mathfrak{H}$, and for a given positive number ϵ, we can always find $\chi\epsilon\mathfrak{A}(\mathfrak{U})\Psi_0$ such that $\|\Phi-\chi\|<\epsilon/3$ because of the cyclicity. For this χ and ϵ we can find an integer N such that, for $n>N$,

$$\|U(f_n)\chi-U(f)\chi\|<\epsilon/3.$$

Then

$$\|U(f_n)\Phi-U(f)\Phi\|<\epsilon$$

for $n>N$. This completes the proof of the theorem.

Theorem 4.3

The necessary and sufficient condition that a functional $E(f,g)$, $f\epsilon D_1$, $g\epsilon D_2$, be represented as

$$E(f,g)=(\Psi_0,U(f)V(g)\Psi_0) \tag{4.10}$$

with cyclic normalized vector Ψ_0 and with unitary operators $U(f)$ and $V(g)$ satisfying (2.8) through (2.10), is that

$$E(f,g)^*=E(-f,-g)e^{i(f,g)}, \tag{4.11}$$

$$E(0,0)=1, \tag{4.12}$$

$$\sum_{i,j=1}^{n} c_ic_j^*E(f_i-f_j,g_i-g_j)e^{i(g_j,f_j)-i(g_j,f_i)}\geqslant 0 \tag{4.13}$$

for any integer n, for any set of complex numbers c_i, and for any collection of test functions $f_i\epsilon D_1$ and $g_i\epsilon D_2$. In order that $U(f)$ and $V(g)$ be continuous on f and g relative to some topologies in D_1 and D_2, respectively, it is necessary and sufficient that $E(f+f_0,g+g_0)$ be continuous on f and g separately. (The proof is similar to theorem 4.2.)

[22] For an example, H. Araki, footnote reference 21.
[23] A "function" of $\mathfrak{U}\mathfrak{B}$ means $H=\int\lambda dP(\lambda)$, $P(\lambda)\epsilon(\mathfrak{U}\mathfrak{B})''$.
[24] I. Segal (preprint).
[25] The corresponding theorem for a complete normed algebra is well known. (M. Neumark, footnote reference 18, pp. 119–121, Sec. 6, 3, VI.) A similar proof can be given to this theorem. (For example, H. Araki, footnote reference 21.) Because of the absence of norm in the algebra $\mathfrak{A}$, representing operators are not necessarily bounded.

We add the following lemma which is useful for some of the later discussions.

Lemma. $U(f_i)$, $i=1\cdots n$ for distinct f_i in a cyclic representation of $\mathfrak{U}$ are linearly independent if $V(g)$ can be defined and if D_2 separates D_1. If Ψ is a cyclic vector relative to $\mathfrak{U}$, then the $U(f_i)\Psi$ are also linearly independent.

Suppose $\sum c_iU(f_i)=0$. We multiply this equation with $V(tg)$ and $V(-tg)$ from both sides and with a function $h(t)$ and integrate over t. We choose g such that (f_i,g) is distinct which is possible if D_2 separates D_1.[26] We choose $h(t)$ such that its Fourier transform vanishes at (f_i,g) except for $i=i_0$. Then we obtain $c_{i_0}=0$, which proves the linear independence of the $U(f_i)$.

5. INVARIANCE UNDER TIME-REVERSAL AND EUCLIDEAN TRANSFORMATIONS

The operator T defined on $\mathfrak{A}(\mathfrak{U})\Psi_0$ through

$$T\sum c_iU(f_i)\Psi_0=\sum c_i^*U(-f_i)\Psi_0 \tag{5.1}$$

is antiunitary on $\mathfrak{A}(\mathfrak{U})\Psi_0$. Hence, it can be extended to an antiunitary operator on the whole space if Ψ_0 is cyclic relative to $\mathfrak{U}$. T, thus defined, has the following properties:

$$T^2=1, \tag{5.2}$$

$$T\Psi_0=\Psi_0, \tag{5.3}$$

$$T\varphi(f)T^{-1}=\varphi(f);\quad [TU(f)T^{-1}=U(-f)]. \tag{5.4}$$

Conversely, an antiunitary operator with the properties (5.2) through (5.4) is unique as can be seen from the foregoing construction of T. Interpreting Ψ_0 as the vacuum state we see that such an operator T has the physical meaning of time reversal operator[27] if it satisfies, in addition,

$$T\pi(g)T^{-1}=-\pi(g);\quad [TV(g)T^{-1}=V(g)]. \tag{5.5}$$

When such an antiunitary operator T exists, the representation of the canonical commutation relations is said to be invariant under time reversal relative to Ψ_0. The time reversal invariance does not impose any restriction on $E(f)$ but gives a restriction (5.5) on V which is to be defined for a given $E(f)$.

For a given $E(f,g)$ of (4.10), the necessary and sufficient condition for the time reversal invariance (relative to Ψ_0) is

$$E(f,g)^*=E(-f,g). \tag{5.6}$$

If (5.6) holds, T can be defined in a cyclic representation of $\mathfrak{UB}$ by

$$T\sum_{i=1}^n c_iU(f_i)V(g_i)\Psi_0=\sum_{i=1}^n c_i^*U(-f_i)V(g_i)\Psi_0. \tag{5.7}$$

Conversely if T exists then (5.6) holds because

$$E(f,g)^*=(T\Psi_0,TU(f)V(g)\Psi_0)=E(-f,g).$$

In the space of a functions of a three vector $\mathbf{x}$ there is a representation of the three-dimensional Euclidean group $\{(\mathbf{a},R)\,|\,\mathbf{a}$: amount of translation, R: rotation$\}$ by linear mappings $L(\mathbf{a},R)$ defined by

$$[L(\mathbf{a},R)f](\mathbf{x})=f(R^{-1}(\mathbf{x}-\mathbf{a})). \tag{5.8}$$

A representation of the canonical commutation relations is said to be invariant under Euclidean group relative to Ψ_0 if there is a set of unitary operators $U(\mathbf{a},R)$ which satisfies

$$U(\mathbf{a}_1,R_1)U(\mathbf{a}_2,R_2)=U(\mathbf{a}_1+R_1\mathbf{a}_2,\ R_1R_2), \tag{5.9}$$

$$U(\mathbf{a},R)U(f)U(\mathbf{a},R)^{-1}=U(L(\mathbf{a},R)f), \tag{5.10}$$

$$U(\mathbf{a},R)V(g)U(\mathbf{a},R)^{-1}=V(L(\mathbf{a},R)g), \tag{5.11}$$

$$U(\mathbf{a},R)\Psi_0=\Psi_0 \tag{5.12}$$

and is continuous in $(\mathbf{a},R)$.

If a cyclic representation of $\mathfrak{UB}$ is invariant under the Euclidean group, then $E(L(\mathbf{a},R)f+f_0,\ L(a,R)g+g_0)$ is continuous in $(\mathbf{a},R)$ and

$$E(f,g)=E(L(\mathbf{a},R)f,L(\mathbf{a},R)g). \tag{5.13}$$

Conversely, if $E(f,g)$ has these properties, then $U(\mathbf{a},R)$ defined by

$$U(\mathbf{a},R)\sum_{i=1}^n c_iU(f_i)V(g_i)\Psi_0$$

$$=\sum_{i=1}^n c_iU(L(\mathbf{a},R)f_i)V(L(\mathbf{a},R)g_i)\Psi_0 \tag{5.14}$$

can be extended to a unitary operator satisfying (5.9) through (5.12). The continuity of $U(a,R)$ can be proved as in theorem 4.2.

The necessary and sufficient condition for the invariance of a cyclic representation of $\mathfrak{U}$ under the Euclidean group is

$$E(f)=E(L(\mathbf{a},R)f), \tag{5.15}$$

and the continuity of $E(L(\mathbf{a},R)f+f_0)$ on $(\mathbf{a},R)$. (5.11) serves as a restriction on V which is to be defined for a given $E(f)$.

6. UNIQUENESS OF THE VACUUM

From now on we shall assume that $E(f)$ [and $E(f,g)$] satisfies the Euclidean invariance requirement (5.15) [and (5.13)]. In this section we shall make an additional requirement, namely, the cluster decom-

[26] If D_2 separates D_1, then for each $f_{ij}=f_i-f_j$ there is g_{ij} such that $(f_{ij},g_{ij})\neq 0$ (no summation). Then the functions

$$h_{kl}(t)=\sum_{i,j}(f_{kl},g_{ij})t_{ij}$$

are not identically zero and hence, there is a $t=\{t_{ij}\}$ such that $h_{kl}(t)\neq 0$ for all kl. Then $g=\sum g_{ij}t_{ij}$ satisfies $(f_{ij},g)\neq 0$ for $i\neq j$ and hence (f_i,g) are distinct.

[27] E. P. Wigner, Nachr. Akad. Wiss. Götingen Math. physik. Kl. (1932) 546.

position property of $E(f)$. This states that correlations do not extend over infinite distances[28,29]:

$$\lim_{\lambda\to\infty}[E(f_0+f_\lambda)-E(f_0)E(f_\lambda)]=0, \tag{6.1a}$$

$$\lim_{\lambda\to\infty}[E(f_0+f_\lambda, g_0+g_\lambda)-E(f_0,g_0)E(f_\lambda,g_\lambda)]=0, \tag{6.1b}$$

where

$$f_\lambda=L(\lambda\mathbf{a},1)f, \quad g_\lambda=L(\lambda\mathbf{a},1)g, \tag{6.2}$$

f_0, $f\epsilon D_1$, g_0, $g\epsilon D_2$ and $\mathbf{a}$ is an arbitrary nonzero translation vector. Equations (6.1) are a very weak form of the cluster decomposition property because we have not specified how fast the left-hand sides decrease with increasing λ.

In the preceding section we have seen how the Euclidean transformation operator $U(\mathbf{a},R)$ may be constructed if $E(f)$ is given. Using the assumed cluster decomposition property, we shall show in theorem 6.1 that there is no other vector besides Ψ_0 which is invariant (even up to a factor) under the so constructed representation of the Euclidean group.

The second problem to be considered in this section is the following. If a cyclic representation of $U(f)$ is given [either in terms of $E(f)$ or in some other way], then corresponding to each cyclic vector Ψ_0' in this space, one can define $E'(f)=(\Psi_0',U(f)\Psi_0')$ and if $E'(f)$ also satisfies the conditions of invariance and cluster decomposition, then one can construct different theories for different $E'(f)$ in the same given representation of $U(f)$. We shall show in Theorem 6.2 that this is not the case, namely, there is at most one $E(f)$ fulfilling (5.15) and (6.1).

Theorem 6.1

Given $E(f)$ satisfying (6.1) and (5.15). Construct $U(\mathbf{a},R)$ as in Sec. 5. Then, any state Ψ_0' invariant under $U(\mathbf{a},1)$ up to a factor is a multiple of Ψ_0. The same conclusion holds for $E(f,g)$ satisfying (6.2) and (5.13) provided that D's are such that $e^{i(f,g)}$ has a cluster decomposition property.[30]

We will prove it for the case of a cyclic representation of $\mathfrak{U}$. Suppose

$$U(\mathbf{a},1)\Psi_0'=e^{i\mu(a)}\Psi_0', \tag{6.3}$$

$$\|\Psi_0'\|=1. \tag{6.4}$$

By (5.9) and the continuity of $U(\mathbf{a},R)$, $\mu(\mathbf{a})=(\mathfrak{y},\mathbf{a})$ for some vector $\mathfrak{y}$. By the cyclicity, there exists for a given $\epsilon>0$ a state Ψ_ϵ of the form

$$\Psi_\epsilon=\sum_{i=1}^{n} c_iU(f_i)\Psi_0 \tag{6.5}$$

such that

$$\|\Psi_\epsilon-\Psi_0'\|<\epsilon. \tag{6.6}$$

(6.4) and (6.6) imply

$$|\,\|\Psi_\epsilon\|-1\,|<\epsilon. \tag{6.7}$$

Define

$$\Psi_\epsilon'=N^{-1}\sum_{k=1}^{N} U(\lambda\mathbf{a}_k,1)\exp[-i\lambda(\mathfrak{y},\mathbf{a}_k)]\Psi_\epsilon \tag{6.8}$$

with a distinct set of vectors $\mathbf{a}_k$. Then by (6.3) and the unitarity of $U(\mathbf{a},1)$, we have

$$\begin{aligned}\|\Psi_\epsilon'-\Psi_0'\| &\leqslant N^{-1}\sum_{k=1}^{N}\|U(\lambda\mathbf{a}_k,1)\exp[-i\lambda(\mathfrak{y},\mathbf{a}_k)]\Psi_\epsilon-\Psi_0'\| \\ &=\|\Psi_\epsilon-\Psi_0'\|<\epsilon.\end{aligned}$$

Hence

$$|\,\|\Psi_\epsilon'\|-1\,|<\epsilon. \tag{6.9}$$

On the other hand,

$$\|\Psi_\epsilon'\|^2=N^{-2}\sum_{k,l=1}^{N}\sum_{i,j=1}^{n} c_j^*c_iE[L(\lambda\mathbf{a}_k,1)f_i - L(\lambda\mathbf{a}_l,1)f_j]\exp[i\lambda(\mathfrak{y},\mathbf{a}_l-\mathbf{a}_k)].$$

For sufficiently large λ, due to (6.1) and (5.15),

$$\begin{aligned}\|\Psi_\epsilon'\|^2 &\sim N^{-2}\sum_{k\neq l}\sum_{i,j=1}^{n} c_j^*c_iE(f_i)E(-f_j) \\ &\quad\times\exp[i\lambda(\mathfrak{y},\mathbf{a}_l-\mathbf{a}_k)]+N^{-2}\sum_{k=1}^{N}\sum_{i,j=1}^{n} c_j^*c_iE(f_i-f_j) \\ &=N^{-2}|\,(\Psi_0,\Psi_\epsilon)\,|^2\sum_{k\neq l}^{N}\exp[i\lambda(\mathfrak{y},\mathbf{a}_l-\mathbf{a}_k)]+N^{-2}\|\Psi_\epsilon\|^2.\end{aligned}$$

Since N and λ are arbitrary and Ψ_ϵ is independent of them, (6.7) and (6.9) together with the foregoing equation imply

$$|\,1-|\,(\Psi_0,\Psi_\epsilon)\,|\,|<\epsilon, \quad \mathfrak{y}=0. \tag{6.10}$$

Using (6.6) we have

$$|\,1-|\,(\Psi_0,\Psi_0')\,|\,|<2\epsilon.$$

Since ϵ is an arbitrary positive number, we have $|\,(\Psi_0,\Psi_0')\,|=1$ which implies that Ψ_0' is a multiple of Ψ_0.

Theorem 6.2

Given $E(f)$ satisfying (6.1) and (5.15). If a set of unitary operators $U'(a,1)$ and a cyclic vector Ψ_0' satisfy equations similar to (5.9), (5.10), and (5.12) in the cyclic representation of $\mathfrak{U}$, then there exists a

[28] For discussions of the relevance of the cluster decomposition property in field theory, see R. Haag, Phys. Rev. **112**, 669 (1958); F. Coester and R. Haag, footnote reference 10.

[29] In a relativistic field theory of particles with nonzero mass, the cluster decomposition property may be proved in a much stronger form than (6.1) from other basic principles. See G. F. Dell'Antonio and P. Gulmanelli, Nuovo cimento 7, 38 (1959); H. Araki, Ann. Phys. (to be published).

[30] This is true if functions in D's tend to zero at infinity.

unitary operator S such that

$$SU(f)S^{-1}=U(f), \tag{6.11}$$

$$SU(\mathbf{a},1)S^{-1}=U'(\mathbf{a},1), \tag{6.12}$$

$$S\Psi_0=\Psi_0'. \tag{6.13}$$

A similar theorem holds for $E(f,g)$.

For the proof, define

$$E'(f)=(\Psi_0',U(f)\Psi_0'). \tag{6.14}$$

For a given $\epsilon>0$, there exists a vector of the form (6.5) satisfying (6.6). Then $E_1(f)$ defined by

$$E_1(f)=\sum_{i,j=1}^{n} c_j^*c_iE(f_i-f_j+f) \tag{6.15}$$

will approximate $E'(f)$ in the sense that

$$|E'(f)-E_1(f)|<2\epsilon. \tag{6.16}$$

Since $E'(f)$ has the property (5.15), we have

$$|E_1(f)-E_1(L(\lambda\mathbf{a},1)f)|<4\epsilon \tag{6.17}$$

for any $\mathbf{a}$ and λ. By (6.1) and (6.15) we have for sufficiently large λ

$$|E_1(L(\lambda\mathbf{a},1)f)-E(f)\|\Psi_\epsilon\|^2|<\epsilon. \tag{6.18}$$

(6.16), (6.17), (6.18), and (6.7) imply (note that $|E(f)|\leqslant 1$)

$$|E'(f)-E(f)|<9\epsilon+\epsilon^2.$$

Since ϵ is arbitrary, we have $E'(f)=E(f)$. Hence, if we define S by

$$S\sum_{i=1}^{n} c_iU(f_i)\Psi_0=\sum_{i=1}^{n} c_iU(f_i)\Psi_0',$$

it can be extended to a unitary operator satisfying (6.11)–(6.13).

We note that theorem 6.1 implies the irreducibility of the set of operators $\mathfrak{B}=\{U(f)U(\mathbf{a},1)\}$ in the cyclic representation of $\mathfrak{U}$. For, take any B in $\mathfrak{B}'$. $B\Psi_0$ is obviously translationally invariant and hence $B\Psi_0=\lambda\Psi_0$ which implies $B=\lambda$ because of cyclicity. Therefore $\mathfrak{B}$ is irreducible.

7. EXISTENCE AND UNIQUENESS OF $V(g)$ IN A CYCLIC REPRESENTATION OF $\mathfrak{U}$

To get a representation of the canonical commutation relations in a given cyclic representation of $\mathfrak{U}$, we have to define $V(g)$. To guarantee the existence of $\mathfrak{B}$, the expectation functional $E(f)$ must satisfy a certain (not very stringent) condition which will be formulated in Sec. 10. In the present section, we shall show that if $\mathfrak{B}$ exists, then it may be chosen so that it satisfies (5.5) and (5.11). Furthermore we demonstrate that (5.5) together with the commutation relations completely determines all the matrix elements of $\pi(g)$ between states of $\mathfrak{A}(\mathfrak{U})\Psi_0$ (which lie dense in $\mathfrak{H}$) so that V satisfying (5.5) is nearly uniquely determined by $E(f)$. The determination of V is not entirely unique due to domain questions of $\pi(g)$.

If at least one representation of $\mathfrak{B}$ exists, then we define a mapping h_g of $\mathfrak{U}''$ into itself by

$$h_g(X)=V(g)XV(g)^{-1},\quad X\epsilon\mathfrak{U}''. \tag{7.1}$$

Obviously $h_g(X)\epsilon\mathfrak{U}'=\mathfrak{U}''$. $h_g(U(f))$ does not depend on the choice of the representation of $\mathfrak{B}$ because of (2.10). Since every element of $\mathfrak{U}''$ is a strong limit of a linear combination of $U(f)$, the mapping h_g is independent of the choice of $\mathfrak{B}$. We note that

$$h_{g1}h_{g2}=h_{g1+g2},\quad h_0=1. \tag{7.2}$$

As we will see in Sec. 10, if $V(g)$ exists then there exists a unique, positive definite self-adjoint operator A_g which is affiliated with $\mathfrak{U}''$ and satisfies

$$(A_g^{\frac{1}{2}}\Psi_0,h_g(X)A_g^{\frac{1}{2}}\Psi_0)=(\Psi_0,X\Psi_0). \tag{7.3}$$

Furthermore $V_0(g)$ defined by

$$V_0(g)X\Psi_0=h_g(X)A_g^{\frac{1}{2}}\Psi_0,\quad X\epsilon\mathfrak{U}'' \tag{7.4}$$

has an extension which is unitary and satisfies (2.9) and (2.10). Here we will show that $V_0(g)$ satisfies (5.5) and (5.11).

For any X in $\mathfrak{U}''$

$$Th_g(X)T^{-1}=h_g(TXT^{-1}), \tag{7.5}$$

$$TXT^{-1}=X^*, \tag{7.6}$$

because these are true for $X=U(f)$. (Note that h_g is a unitary transformation and hence conserves a strong limit.) (7.6) is true for $A_g^{\frac{1}{2}}$ which is affiliated with $\mathfrak{U}''$. Hence,

$$TV_0(g)T^{-1}X\Psi_0=Th_g(X^*)A_g^{\frac{1}{2}}\Psi_0=V_0(g)X\Psi_0,\quad X\epsilon\mathfrak{U}''$$

which proves (5.5) for $V_0(g)$.

Next, for any X in $\mathfrak{U}''$

$$U(\mathbf{a},R)h_g(X)U(\mathbf{a},R)^{-1}=h_{L(\mathbf{a},R)g}[U(\mathbf{a},R)XU(\mathbf{a},R)^{-1}] \tag{7.7}$$

because this is true for $X=U(f)$. Furthermore

$$\begin{aligned}&(U(\mathbf{a},R)A_g^{\frac{1}{2}}U(\mathbf{a},R)^{-1}\Psi_0,h_{L(\mathbf{a},R)g}(X)\\&\quad\times U(\mathbf{a},R)A_g^{\frac{1}{2}}U(\mathbf{a},R)^{-1}\Psi_0)\\&\quad=(A_g^{\frac{1}{2}}\Psi_0,h_g[U(\mathbf{a},R)^{-1}XU(\mathbf{a},R)]A_g^{\frac{1}{2}}\Psi_0)=(\Psi_0,X\Psi_0).\end{aligned}$$

Since $U(a,R)A_g^{\frac{1}{2}}U(a,R)^{-1}$ commute with any X in $\mathfrak{U}''$ and is positive semidefinite, we have from the uniqueness of A_g

$$U(\mathbf{a},R)A_g^{\frac{1}{2}}U(\mathbf{a},R)^{-1}=A_{L(\mathbf{a},R)g}^{\frac{1}{2}}. \tag{7.8}$$

Hence, for any X in $\mathfrak{U}''$,

$$\begin{aligned}&U(\mathbf{a},R)V_0(g)U(\mathbf{a},R)^{-1}X\Psi_0\\&\qquad=U(\mathbf{a},R)h_g(U(\mathbf{a},R)^{-1}XU(\mathbf{a},R))A_g^{\frac{1}{2}}\Psi_0\\&\qquad=V_0(L(a,R)g)X\Psi_0,\end{aligned}$$

which proves (5.11) for $V_0(g)$.

We now show the uniqueness of matrix elements of $\pi(g)$ between states of $\mathfrak{A}(\mathfrak{U})\Psi_0$. The transformation property (5.5) is crucial. From (2.10)

$$(\Psi_0,U(f)\pi(g)\Psi_0)-(\Psi_0,\pi(g)U(f)\Psi_0)=-(f,g)E(f).$$

On the other hand, from (5.4), (5.5), and (2.12),

$$(\Psi_0,U(f)\pi(g)\Psi_0)=(TU(f)\pi(g)\Psi_0T\Psi_0)$$
$$=-(\Psi_0,\pi(g)U(f)\Psi_0).$$

Therefore,

$$(\Psi_0,U(f)\pi(g)\Psi_0)=-(\tfrac{1}{2})(f,g)E(f), \tag{7.9}$$

$$(U(f_1)\Psi_0,\pi(g)U(f_2)\Psi_0)=\tfrac{1}{2}[(f_1,g)+(f_2,g)]E(f_2-f_1). \tag{7.10}$$

This uniqueness does not necessarily imply the uniqueness of $V(g)$ because Ψ_0 may not be in the domain of $\pi(g)$, and even if it is there may be many different self-adjoint extensions of $\pi(g)$.

Suppose that there are two $V(g)$, say $V_1(g)$ and $V_2(g)$. Then

$$W(g)=V_1(g)^{-1}V_2(g) \tag{7.11}$$

is unitary and commutes with $U(f)$ and T. This implies that $W(g)\epsilon\mathfrak{U}''$ and

$$W(g)^*=TW(g)T^{-1}=W(g), \tag{7.12}$$

where the first equality is due to (7.6). Hence, $W(g)$ is idempotent

$$W(g)^2=1. \tag{7.13}$$

Since $W(g)\epsilon\mathfrak{U}''$, they commute with each other. From (2.9)

$$W(g_1+g_2)=W(g_1)h_{-g_1}W(g_2). \tag{7.14}$$

From (2.11)

$$W(0)=1. \tag{7.15}$$

Conversely, for any Hermitian unitary $W(g)$ in $\mathfrak{U}''$ satisfying (7.14) and (7.15), it is easy to show that $V_1(g)W(g)$ has the properties (2.9) and (2.10) for $V(g)$.

An obvious solution of (7.14) and (7.15) is given by

$$W(g)=W_0h_{-g}(W_0) \tag{7.16}$$

where W_0 is some fixed Hermitian unitary operator in $\mathfrak{U}''$. In this case and only in this case the two $V(g)$ connected with the multiplier $W(g)$ are related by a unitary transformation in $\mathfrak{U}''$.

Examples where more than one $V(g)$ satisfying (5.5) and (5.11) but not connected by a unitary transformation in $\mathfrak{U}''$ exist in a cyclic representation of $\mathfrak{U}$ can be constructed.[31]

If $\mathfrak{U}\mathfrak{B}$ is irreducible for one definition of $V(g)$, then there is no $W(g)$ of the form (7.16) except 1. For, suppose $W(g)$ of the form (7.16) exists, and both $V(g)$ satisfy (5.11). Then

$$U(\mathbf{a},1)W(g)U(\mathbf{a},1)^{-1}=W(L(\mathbf{a},1)g). \tag{7.17}$$

[31] Examples will be discussed in a separate paper.

Set $X=W_0U(\mathbf{a},1)W_0U(\mathbf{a},1)^{-1}$. $X\epsilon\mathfrak{U}''$ and from (7.17) $X=h_{L(a,1)g}(X)$ for all g. These imply that $X\epsilon(\mathfrak{U}\mathfrak{B})'$. Hence by irreducibility assumption X is a multiple of the identity operator. Since $X^2=1$,[32] and $X_{\mathbf{a}=0}=1$, we have $X=1$. Therefore $W_0\epsilon\mathfrak{B}'$ which implies $W_0=\pm1$.[33]

8. EXISTENCE AND UNIQUENESS OF HAMILTONIAN

We will restrict our attention to the models in which π is the time derivative of φ,

$$[H,\varphi(f)]=-i\pi(f).^{34} \tag{8.1}$$

In terms of $U(f)$ this is equivalent to

$$[H,U(f)]=U(f)\pi(f)+(\tfrac{1}{2})(f,f)U(f). \tag{8.2}$$

In addition we assume that the cyclic vector Ψ_0 satisfies

$$H\Psi_0=0. \tag{8.3}$$

We now show that (8.2) and (8.3) are sufficient to determine all matrix elements of H between states of $\mathfrak{A}(\mathfrak{U})\Psi_0$. Take the matrix element of (8.2) between $U(f_1)\Psi_0$ and Ψ_0 and use (7.10) and (8.3). The result is

$$(U(f_1)\Psi_0,HU(f_2)\Psi_0)=(\tfrac{1}{2})(f_1,f_2)E(f_1-f_2). \tag{8.4}$$

The class of models with the property (8.1) corresponds formally to Hamiltonian functions of the form

$$H=(\tfrac{1}{2})\int\pi(\mathbf{x})^2d\mathbf{x}+H'(\varphi) \tag{8.5}$$

where H' is a functional of φ only. Another characterization of these models which avoid the use of $\pi(g)$ is

$$[U(f_1),[H,U(f_2)]]=-(f_1,f_2)U(f_1+f_2). \tag{8.6}$$

Equation (8.6) is equivalent to (8.2) provided that H is Hermitian and invariant under time reversal[35]:

$$H^*=H, \tag{8.7}$$

$$THT^{-1}=H. \tag{8.8}$$

Starting from (8.6) we derive the matrix elements of H in the following manner. Owing to (8.3) we have

$$(\Psi_0,U(f_1)HU(f_2)\Psi_0)+(\Psi_0,U(f_2)HU(f_1)\Psi_0)$$
$$=-(f_1,f_2)E(f_1+f_2).$$

On the other hand, by (8.7) and (8.8),

$$(\Psi_0,U(f_1)HU(f_2)\Psi_0)=(TU(f_1)HU(f_2)\Psi_0,T\Psi_0)$$
$$=(\Psi_0,U(f_2)HU(f_1)\Psi_0).$$

From the foregoing two equations we obtain (8.4).

[32] Note that $U(a,1)W_0U(a,1)^{-1}$ commutes with W_0 because both are in $\mathfrak{U}''$. Note also that $W_0^2=1$.

[33] We assume here the cluster decomposition property of $E(f)$ and use the results of Sec. 6.

[34] We assume that D_1 and D_2 have sufficiently large intersection, so that expressions like $\pi(f)$ and (f_1,f_2) are meaningful for sufficiently many f's. We may assume that $D_1=D_2$.

[35] For the derivation of (8.2) from (8.6)–(8.8), see H. Araki, thesis, Princeton University, 1960.

We now show the existence of a positive semidefinite, self-adjoint H satisfying (8.4). First we note that if the equality in (4.9) occurs only when $c_i=0$ then the state $\sum_{i=1}^{n} c_i U(f_i)\Psi_0$ vanishes only when $c_i=0$ and hence

$$\left(\sum_{j=1}^{m} d_j U(f_j')\Psi_0, H\sum_{i=1}^{n} c_i U(f_i)\Psi_0\right) = \tfrac{1}{2}\sum_{i,j}^{n,m} c_i d_j^*(f_j',f_i)E(f_i-f_j') \quad (8.9)$$

is a consistent definition of a Hermitian form H on $\mathfrak{A}(\mathfrak{U})\Psi_0$.

Furthermore H is positive semidefinite on $\mathfrak{A}(\mathfrak{U})\Psi_0$, namely,

$$(\Psi,H\Psi)=\tfrac{1}{2}\sum_{i,j}^{n} c_j^* c_i(f_j,f_i)E(f_i-f_j)\geqq 0, \quad (8.10)$$

where

$$\Psi=\sum_{i=1}^{n} c_i U(f_i)\Psi_0.$$

To prove this we need the following lemma.

Lemma. If two (finite dimensional) Hermitian matrices A and B are positive semidefinite, then the matrix C, whose matrix elements are

$$C_{ij}=A_{ij}B_{ij} \quad (8.11)$$

in any fixed orthonormal basis, is also positive semidefinite.

Since C is a restriction of the Kronecker product of A and B to a special subspace, it is positive semidefinite. In a less abstract way, A and B can be written as

$$A_{ij}=\sum_k a_k u_{ki}^* u_{kj}, \quad a_k\geqq 0;$$
$$B_{ij}=\sum_k b_k v_{ki}^* v_{kj}, \quad b_k\geqq 0.$$

Hence

$$\sum_{i,j} x_i^* C_{ij} x_j=\sum_{k,l} a_k b_l\left|\sum_i u_{ki}v_{li}x_i\right|^2\geqq 0,$$

which shows the positive semidefiniteness of C.

We take $E(f_j-f_i)$ as A_{ij} and (f_i,f_j) as B_{ij}. A is positive semidefinite due to (4.9) and B is obviously so. Hence H is also positive semidefinite due to the above lemma.

We now use the following theorem of Friedrichs.[36]

Theorem 8.1

A positive semidefinite Hermitian form $\{\Psi_1,\Psi_2\}$ defined on a dense linear set $\mathfrak{R}$ in a Hilbert space $\mathfrak{H}$ can be extended by continuity to a positive semidefinite Hermitian form on a larger linear set $\mathfrak{R}_1\supset\mathfrak{R}$ which consists of elements Ψ of $\mathfrak{H}$ such that, for some sequence of elements Ψ_n of $\mathfrak{R}$, $\|\Psi-\Psi_n\|\to 0$ and $\{\Psi_n-\Psi_m, \Psi_n-\Psi_m\}\to 0$. Furthermore there exists a positive semidefinite self-adjoint operator A on $\mathfrak{H}$ with its domain D in $\mathfrak{R}_1$ such that

$$\{\Psi_1,\Psi_2\}=(\Psi_1,A\Psi_2)$$

for any $\Psi_1\epsilon\mathfrak{R}_1$ and $\Psi_2\epsilon D$.

Owing to this theorem, we have the following theorem.

Theorem 8.2

For a given cyclic representation of $\mathfrak{U}$, in which $U(f)$ is linearly independent for distinct f, there always exists a positive semidefinite, self-adjoint Hamiltonian H with domain D contained in $\mathfrak{R}_1$, which has an extension to a Hermitian form on $\mathfrak{R}_1$ such that the extended form satisfies (8.4) on $\mathfrak{A}(\mathfrak{U})\Psi_0$, where $\mathfrak{R}_1$ is constructed from $\mathfrak{R}=\mathfrak{A}(\mathfrak{U})\Psi_0$ as stated in the theorem 8.1. D is invariant under T and $U(a,R)$ and H satisfies (8.8) and

$$U(\mathbf{a},R)HU(\mathbf{a},R)^{-1}=H. \quad (8.12)$$

For the proof of the latter half, we have only to note that $\mathfrak{R}_1$ is invariant under T and $U(\mathbf{a},R)$ and (8.8) and (8.12) holds on $\mathfrak{R}_1$ as an equation for the Hermitian form H and operators T and $U(\mathbf{a},R)$ as one can easily verify.

As for the uniqueness, if one assumes that the domain of H is contained in $\mathfrak{R}_1$, then H is unique. Without any condition on the domain, the extension is not unique and Krein's method[37] gives all the positive semidefinite self-adjoint extension. There can also be a self-adjoint extension of H which is not positive semidefinite.[38]

The Hamiltonian H has the following noteworthy property. Denote by $\mathfrak{U}_C$ the collection of $U(f)$, where $f\epsilon D_1$ has its support in a fixed region C of x space. Suppose that C_i are a finite number of disjoint regions and A_i, $A_i'\epsilon\mathfrak{A}(\mathfrak{U}_{C_i})$. Then states in $\mathfrak{A}(\mathfrak{U}_{C_1})\Psi_0$ for different i are orthogonal to each other relative to the Hermitian form $\{\Psi_1,\Psi_2\}=(\Psi_1,H\Psi_2)$, namely,

$$\left(\sum_i A_i'\Psi_0, H\sum_j A_j\Psi_0\right)=\sum_i(A_i'\Psi_0,HA_i\Psi_0). \quad (8.13)$$

As a final remark, the requirement that H is "local," i.e., of the form

$$H=\tfrac{1}{2}\int \pi(\mathbf{x})^2 d\mathbf{x}+\int H'(\mathbf{x})d\mathbf{x}, \quad (8.14)$$

where $H'(\mathbf{x})$ depends only on $\varphi(\mathbf{x})$ and its space derivatives, can be expressed by

$$[V(g_1),[H,V(g_2)]]=0 \quad (8.15)$$

whenever the supports of g_1 and g_2 are disjoint.

[36] K. O. Friedrichs, Math. Ann. **109**, 465, 685 (1934); *ibid.* **110**, 777 (1935); R. Riesz and B. Sz-Nagy, *Lecons d'analyse Fonctionnelle* (Akadémiai Kiadó, Budapest, 1953), Chap. VIII, Sec. 124.

[37] M. Krein, Rec. (Sbornik) Math. Moscow **20**, 431 (1947); *ibid.* 365 (1947); F. Riesz and B. Sz-Nagy, footnote reference 36, Sec. 125.

[38] For an example, see H. Araki, footnote reference 21, p. 57.

9. RELATIVISTIC INVARIANCE

The infinitesimal generator K corresponding to an infinitesimal Lorentz transformation

$$x'^{\mu}=x^{\mu}+(\lambda_{\nu}{}^{\mu}x^{\nu}+a^{\mu}) \tag{9.1}$$

can be written as

$$K=\tfrac{1}{2}M^{\mu\nu}\lambda_{\mu\nu}+P^{\mu}a_{\mu}, \tag{9.2}$$

$$M_{\mu\nu}=-M_{\nu\mu}. \tag{9.3}$$

P^i are momentum operators, $P^0=H$, M^{ij} are angular momentum operators, and M^{0i} are infinitesimal generators for pure Lorentz transformations. Greek indices run from 0 to 3 and Roman indices run from 1 to 3. The signature of the metric is $(1,-1,-1,-1)$ and $c=1$. As is well known,

$$[P^{\mu},P^{\nu}]=0, \tag{9.4}$$

$$i[M^{\mu\nu},P^{\lambda}]=P^{\mu}g^{\nu\lambda}-P^{\nu}g^{\mu\lambda}, \tag{9.5}$$

$$i[M^{\mu\nu},M^{\rho\sigma}]=-g^{\mu\rho}M^{\nu\sigma}+g^{\mu\sigma}M^{\nu\rho}-g^{\nu\sigma}M^{\mu\rho}+g^{\nu\rho}M^{\mu\sigma}. \tag{9.6}$$

The vacuum Ψ_0 is assumed to be invariant,

$$P^{\mu}\Psi_0=M^{\mu\nu}\Psi_0=0. \tag{9.7}$$

In a given cyclic representation of $\mathfrak{U}$, we define a four-dimensional field ψ by

$$\psi(h)=\int dt e^{iHt}\varphi(h)e^{-iHt}, \tag{9.8}$$

where h is a four-dimensional test function whose restriction on a spacelike hyperplane is in D_1, and $\varphi(h)$ still depends on t. We assume that $\psi(h)$ is a scalar field[39] and $\pi(g)$ is the time derivative of ψ when $h\to g(\mathbf{x})\delta(t)$. Then the following commutation relations hold:

$$i[P^i,\varphi(f)]=-\varphi(\partial^i f), \tag{9.9}$$

$$i[H,\varphi(f)]=\pi(f), \tag{9.10}$$

$$i[M^{ij},\varphi(f)]=\varphi([x^i\partial^j-x^j\partial^i]f), \tag{9.11}$$

$$i[M^{0i},\varphi(f)]=\pi(x^i f), \tag{9.12}$$

$$i[P^i,\pi(f)]=-\pi(\partial^i f), \tag{9.13}$$

$$i[H,\pi(f)]=\dot{\pi}(f), \tag{9.14}$$

$$i[M^{ij},\pi(f)]=\pi([x^i\partial^j-x^j\partial^i]f), \tag{9.15}$$

$$i[M^{0i},\pi(f)]=\dot{\pi}(x^i f)+\varphi(\partial^i f), \tag{9.16}$$

where [40] (9.14) defines $\dot{\pi}$ and

$$[x^i\partial^j f](\mathbf{x})=x^i\partial^j f(\mathbf{x})=-x^i(\partial/\partial x^j)f(\mathbf{x}). \tag{9.17}$$

In a cyclic representation of $\mathfrak{U}$, we have already discussed the condition for the existence and uniqueness of P^i, M^{ij} and H. P^i and M^{ij} are infinitesimal generators of $U(\mathbf{a},R)$. We will now consider the uniqueness and existence of M^{0i}. These operators have to satisfy (9.12), (9.16), and those equations among (9.4)–(9.6) which contain M^{0i}, namely,

$$i[M^{0i},H]=-P^i, \tag{9.18}$$

$$i[M^{0i},P^j]=-\delta_{ij}H, \tag{9.19}$$

$$i[M^{0i},M^{jk}]=-\delta_{ij}M^{0k}+\delta_{ik}M^{0j}, \tag{9.20}$$

$$i[M^{0i},M^{0j}]=-M^{ij}. \tag{9.21}$$

We will first show the uniqueness of matrix elements of M^{0i} between states of $\mathfrak{A}(\mathfrak{U})\Psi_0$.[41] From (9.12) we have

$$[M^{0i},U(f)]=U(f)\pi(x^i f)+\tfrac{1}{2}U(f)(f,x^i f). \tag{9.22}$$

By (7.10) we have

$$(U(f_1)\Psi_0,M^{0i}U(f_2)\Psi_0)=\tfrac{1}{2}(f_1,x^i f_2)E(f_2-f_1). \tag{9.23}$$

Next we consider the problem of existence neglecting all domain questions. We shall reduce all the conditions on M^{0i} defined by (9.23) to a single equation. First, if (9.18) holds, then due to (9.14) and the Jacobi identity, (9.16) also holds. Conversely, if (9.16) holds, one can easily verify that $i[M^{0i},H]+P^i$ commutes with $U(f)$ and annihilates the vacuum. Hence we conclude that (9.18) holds at least on $\mathfrak{A}(\mathfrak{U})\Psi_0$. Similarly if (9.16) is true, (9.19)–(9.21) holds on $\mathfrak{A}(\mathfrak{U})\Psi_0$. Of course we do not know whether $\mathfrak{A}(\mathfrak{U})\Psi_0$ is in the domain of operators in these equations.

Thus, assuming that the operator M^{0i} defined by (9.23) exists and $\mathfrak{A}(\mathfrak{U})\Psi_0$ is contained in the domain of all the relevant operators, the condition for the relativistic invariance of the theory is given either by (9.16) or by (9.18).

We can further reduce this condition to a condition on $E(f,g)$ or more specifically on the matrix elements of the product of two $\pi(g)$ between states of $\mathfrak{A}(\mathfrak{U})\Psi_0$. Namely, since H and M^{0i} annihilate the vacuum,

$$(U(f_1)\Psi_0,M^{0i}HU(f_2)\Psi_0)=([M^{0i},U(f_1)]\Psi_0,[H,U(f_2)]\Psi_0).$$

Using a similar equation for HM^{0i}, (9.22), (8.2), and (2.10), we see that (9.18) is equivalent to

$$\begin{aligned}(\Psi_0,\,&U(f_2-f_1)[\pi(x^i f_1)\pi(f_2)-\pi(f_1)\pi(x^i f_2)]\Psi_0)\\&=i\frac{d}{dt}E\left(f_2-f_1+t\frac{\partial f_2}{\partial x^i}\right)\Big|_{t=0}+\tfrac{1}{4}\{(f_2,x^i f_2)\\&\times[(f_1,f_1)-(f_1,f_2)]+(f_1,x^i f_2)[(f_2,f_2)-(f_1,f_1)]\\&+(f_1,x^i f_1)[(f_1,f_2)-(f_2,f_2)]\}E(f_2-f_1).\end{aligned} \tag{9.24}$$

If the free field is the only relativistic quantum field theory with canonical commutation relation, then (9.24) might be useful to prove it.

[39] This means that $U(a,\Lambda)\psi(h)U(a,\Lambda)^{-1}=\psi[L(a,\Lambda)h]$, where $U(a,\Lambda)$ is a unitary representation of the inhomogeneous Lorentz group and $[L(a,\Lambda)h](x^{\mu})=h[(\Lambda^{-1})_{\nu}{}^{\mu}(x^{\nu}-a^{\nu})]$.

[40] Note that $\varphi(\partial f/\partial x^i)=-\int f(\mathbf{x})\partial\varphi(\mathbf{x})/\partial x^i d\mathbf{x}$ symbolically.

[41] This is, of course, meaningful only when such matrix element exists.

10. $E(f)$ AND QUASI-INVARIANT NONNEGATIVE MEASURE

Among many ways to represent the Hilbert space for a single oscillator, the representation which diagonalizes the position variable is useful because the wave function enables one to visualize various situations on the one hand, and because a highly developed technique of partial differential equations is available for the method of solution on the other.

An analogous representation for the case of quantum field theory would be one which diagonalizes the field $\varphi(\mathbf{x})$. The difficulty lies in finding a manageable way in which the scalar product between states

$$(\Psi_1,\Psi_2)=\int \Psi_1(\chi)^*\Psi_2(\chi)d\mu(\chi) \tag{10.1}$$

and the vacuum expectation functional

$$E(f)=\int e^{i(f,\chi)}|\Psi_0(\chi)|^2 d\mu(\chi) \tag{10.2}$$

can be defined and computed.

A systematic study in this direction using the representation theory of algebras has been made by Lew.[8] A heuristic discussion has been made by Coester and Haag.[10] The reader is also referred to footnote references 7, 9, and 24.

In this section we shall show that $E(f)$ is a Fourier transform of a nonnegative quasi-invariant measure. This means that $E(f)$ can be expressed as in (10.2). However our measure space seems to be too large for practical purposes.

Let us take an arbitrary functional $E(f)$ of Theorem 4.2. Then the function of $\mathbf{t}=(t_1\cdots t_n)$ defined by

$$e(\mathbf{t})=E(\sum_{i=1}^{n} t_i f_i) \tag{10.3}$$

is continuous in t and positive in Schwartz's sense.[42] Hence by Bochner's theorem[43] it is a Fourier transform of a non-negative measure

$$e(\mathbf{t})=\int \exp[i(\mathbf{t},\mathbf{p})]d\mu(\mathbf{p}). \tag{10.4}$$

Because of (4.8)

$$\int d\mu(\mathbf{p})=1. \tag{10.5}$$

Since this measure depends on the f's, we write the measure of a Borel p set A as

$$\mu(\mathbf{p}\epsilon A; f_1\cdots f_n).$$

Obviously μ has the following property,

$$\mu(\mathbf{p}\epsilon A; f_1\cdots f_n)=\mu(\mathbf{p}\epsilon A\otimes R^{m-n}; f_1\cdots f_n), \tag{10.6}$$

$$\mu(\mathbf{p}\epsilon A; f_1\cdots f_n)=\mu(\mathbf{p}\epsilon A^L; f_1{}^L\cdots f_n{}^L), \tag{10.7}$$

where $m\geqslant n$, L is a nonsingular matrix, $f_i{}^L=\sum_j L_{ij}f_j$ and

$$A^L=\{(\textstyle\sum_j L_{ij}p_j)\,|\,\mathbf{p}\epsilon A\}.$$

We now apply the following theorem of Kolmogorov.[44]

Theorem 10.1

Let T be any infinite aggregate, and the space Ω be that of the real valued functions of $t\epsilon T$. Let $t_1\cdots t_n$ be a finite subset of T and let A be an n-dimensional Borel set. The condition $[\xi(t_1)\cdots\xi(t_n)]\epsilon A$ for t function $\xi(\cdot)$ defines a subset of Ω. Let $\mathfrak{F}_0$ be the class of all such subsets of Ω. ($\Omega\epsilon\mathfrak{F}_0$.) Suppose that a set function q is defined on the sets of $\mathfrak{F}_0$ with the following property. For each finite fixed t set $(t_1\cdots t_n)$, $q(\{\xi(\cdot)\,|\,[\xi(t_1)\cdots\xi(t_n)]\epsilon A\})$ is a nonnegative measure of the Borel set A. In addition, in order that q is single-valued, q should satisfy

$$q(\{\xi(\cdot)\,|\,[\xi(t_1)\cdots\xi(t_n)]\epsilon A\}) = q(\{\xi(\cdot)\,|\,[\xi(t_1)\cdots\xi(t_m)]\epsilon A\otimes R^{m-n}\}), \tag{10.8}$$

$$q(\{\xi(\cdot)\,|\,[\xi(t_1)\cdots\xi(t_n)]\epsilon A\}) = q(\{\xi(\cdot)\,|\,[\xi(t_{P1})\cdots\xi(t_{Pn})]\epsilon A^P\}), \tag{10.9}$$

where $m\geqslant n$, $P_1\cdots P_n$ is a permutation of $1\cdots n$ and A^P is obtained from A by permutation P of coordinates. Then q can be extended to a complete nonnegative measure defined on the sets of $\bar{B}(\mathfrak{F}_0)$ where $B(\mathfrak{F}_0)$ is a Borel field generated by $\mathfrak{F}_0$ and $\bar{B}(\mathfrak{F}_0)$ is the completion of $B(\mathfrak{F}_0)$.

We introduce a Hamel basis $\{f_\alpha\}$ in D_1 and identify T with $\{f_\alpha\}$. Since a function on T is in one-to-one correspondence with a linear functional on D_1, one can identify Ω with the set D_1' on all linear functionals on D_1. We define q by

$$q(\{\xi(\cdot)\,|\,[\xi(t_1)\cdots\xi(t_n)]\epsilon A\})=\mu(\mathbf{p}\epsilon A; f_1\cdots f_n) \tag{10.10}$$

with $t_i=f_i$. Then due to (10.6) and (10.7), the Eqs. (10.8) and (10.9) are satisfied. Then we have a measure on the set of $\bar{B}(\mathfrak{F}_0)$ which we will call μ. Suppose that χ is an element of Ω and

$$f=\sum_{i=1}^{n} c_i f_i.$$

Then $e^{i(\chi,f)}$ is a measurable and integrable function and

$$\int_{D_1'} e^{i(\chi,f)}d\mu(\chi)=\int_{D_1'} e^{i\Sigma c_i(\chi,f_i)}d\mu(\chi) = \int e^{i\Sigma c_i p_i}d\mu(\mathbf{p}; f_1\cdots f_n)=E(f). \tag{10.11}$$

[42] L. Schwartz, *Theorie des Distributions II* (Hermann & Cie, Paris, 1951), p. 130.

[43] S. Bochner, *Vorlesungen über Fouriesche integrale* (Akademische Verlagsgesellschaft, Leibzig, Germany, 1932); L. Schwartz, footnote reference 42, p. 132, theorem XVIII.

[44] A. Kolmogorov, Grundbegriffe der Wahrscheinlichkeitsrechnung, Ergeb. Math. 2, No. 3, 27–30 (1933).

Finally, we will show that the measure μ does not depend on the Hamel basis $\{f_\alpha\}$. For this it suffices to show that for any set M in $\mathfrak{F}_0$, $\mu(M)$ is independent of the Hamel basis. Suppose that

$$M=\{\chi(\cdot)|[\chi(f_1)\cdots\chi(f_n)]\epsilon A\}=\{\chi(\cdot)|[\chi(f_1')\cdots\chi(f_{m'})]\epsilon A'\},$$

where $\{f_\alpha\}$ and $\{f_\alpha'\}$ are different Hamel bases. Since the only restriction on χ is the linearity, by using Eq. (10.8) one can reduce the form of M into

$$M=\{\chi(\cdot)|[\chi(f_1)\cdots\chi(f_l)]\epsilon A_1\}=\{\chi(\cdot)|[\chi(f_1')\cdots\chi(f_l')]\epsilon A_2\},$$
$$f_i'=\sum_j L_{ij}f_j,\quad A_2=A_1{}^L,\quad m+n\geqslant l\geqslant m,n$$

where L is a nonsingular matrix. Then by Eq. (10.7) we see that $\mu(M)$ is independent of the choice of the Hamel basis.

Conversely, any $E(f)$ of the form (10.11) satisfies (4.7)–(4.9). For example,

$$\sum c_j{}^*E(f_j-f_i)c_i=\int|\sum c_ie^{i(\chi,f_i)}|^2d\mu(\chi)\geqslant 0.$$

The continuity of $E(tf+f_0)$ in t follows from Bochner's theorem.[43] Thus we have the following theorem.

Theorem 10.2

In order that $E(f)$ satisfies conditions of theorem 4.2, it is necessary and sufficient that $E(f)$ be a Fourier transform of a nonnegative measure μ on the space of all linear functional on D_1. The cyclic representation of $\mathfrak{A}(\mathfrak{U})$ in theorem 4.2 is unitarily equivalent to the multiplication algebra of the functionals $\sum_{i=1}^n c_ie^{i(\chi,f_i)}$ on the Hilbert space of all L_2 functionals with respect to the measure μ.

For the proof of the latter half of the theorem, we identify $\sum c_ie^{i(\chi,f_i)}$ with elements of $\mathfrak{A}(\mathfrak{U})\Psi_0$ in the Hilbert space $\mathfrak{H}=L_2(D_1',\mu)$ of all L_2 functionals with the inner product

$$(\Psi_1,\Psi_2)=\int\Psi_1(\chi)^*\Psi_2(\chi)d\mu(\chi)\qquad(10.12)$$

(1 corresponds to Ψ_0). Then the elements of $\mathfrak{H}(\mathfrak{U},\Psi_0)$ consist of the functional belonging to the closure of the set of functionals of the type $\sum c_je^{i(\chi,f_j)}$ with respect to L_2 norm. This includes, in particular, the characteristic function of any set of belonging to $\mathfrak{F}_0$ and hence, the characteristic function of any set in $B(\mathfrak{F}_0)$. Therefore all L_2 functionals (more precisely their equivalence classes) are in $\mathfrak{H}(\mathfrak{U},\Psi_0)$. This completes the proof of the theorem.

Next we will consider the condition on the measure for the existence of $V(g)$. If D_1 separates D_2, then D_2 can be considered as a subset of D_1'. For a subset B of D_1' we define $(B+g)=\{\chi+g|\chi\epsilon B\}$. A measure μ is said to be D_2-quasi-invariant if $\mu(B)=0$ implies $\mu(B+g)=0$ for any g in D_2. We will prove the following theorem.

Theorem 10.3

A necessary and sufficient condition for the measure μ in order that $V(g)$ may be defined in $L_2(D_1',\mu)$ is that μ be D_2-quasi-invariant.

For the sufficiency proof, we use the Radon Nikodym theorem.[45] If μ is D_2-quasi-invariant, then there exists the derivative

$$A_g(\chi)=d\mu(\chi+g)/d\mu(\chi)\geqslant 0\qquad(10.13)$$

such that

$$\int F(\chi+g)A_g(\chi)d\mu(\chi)=\int F(\chi)d\mu(\chi).\qquad(10.14)$$

We define $V(g)$ by

$$[V(g)\Psi](\chi)=\Psi(\chi+g)[A_g(\chi)]^{\frac{1}{2}}.\qquad(10.15)$$

First, by (10.14) $V(g)$ is unitary,

$$\int|[V(g)\Psi](\chi)|^2d\mu(\chi)=\int|\Psi(\chi+g)|^2A_g(\chi)d\mu(\chi)=\int|\Psi(\chi)|^2d\mu(\chi).$$

Second, since[46]

$$A_g(\chi+g')A_{g'}(\chi)=A_{g+g'}(\chi)$$

we obtain (2.9) in the following way:

$$[V(g)V(g')\Psi](\chi)=[V(g')\Psi](\chi+g)A_g(\chi)^{\frac{1}{2}}=\Psi(\chi+g+g')[A_{g'}(\chi+g)A_g(\chi)]^{\frac{1}{2}}=[V(g+g')\Psi](\chi).$$

Third, (2.10) is trivially satisfied.

For the necessity proof, we note that any operator in $\mathfrak{A}(\mathfrak{U})$ is represented by the multiplication of

$$R(x)=\sum_{i=1}^n c_ie^{i(\chi,f_i)}\qquad(10.16)$$

and satisfies

$$[V(g)RV(-g)](\chi)=R(\chi+g).\qquad(10.17)$$

Denoting by P_B the projection operator defined by multiplication of the characteristic function $\varphi_B(\chi)$ of a set B, we prove

$$V(g)P_BV(-g)=P_{B-g}.\qquad(10.18)$$

If B is a periodic, finite dimensional cylinder function, then $\varphi_B(\chi)$ is a uniform limit of $R(\chi)$ and hence we get (10.18) from (10.17). By taking the limit of an infinite period, the (10.18) for any set B in $\mathfrak{F}_0$ holds as a strong limit of the (10.18) for periodic B. $B(F_0)$ is

[45] P. Halmos, *Measure Theory* (D. Van Nostrand Company, Inc., Princeton, New Jersey, 1950), Sec. 31.

[46] P. Halmos, footnote reference 45.

the closure of $\mathfrak{F}_0$ under the operations $B \to B^c$, $\{B_1,B_2\} \to B_1 \bigcap B_2$, and $B_1 \subset B_2 \cdots \subset B_n \subset \cdots \to \bigcup B_n \equiv B$. In the last operation $P_{B_n} \to P_B$ strongly. Hence we see that (10.18) holds for B in $B(F_0)$. Since $P_B=0$ and $\mu(B)=0$ is equivalent, μ is D_2-quasi-invariant.

11. CONCLUDING REMARKS

We have seen that the vacuum expectation functional $E(f)$ determines essentially all the content of a theory in Hamiltonian formalism under the assumption that the Hilbert space is cyclic relative to $\mathfrak{U}$. Namely, it determines the Hilbert space and the representation of $U(f)$ (theorem 4.2), the time-reversal operator T and a unitary representation of Euclidean group (Sec. 5), the vacuum state (theorem 6.1), the representation of $V(g)$ (Sec. 7), the Hamiltonian (theorem 8.2), and the representation of Lorentz group if it exists (Sec. 9).

The unitarity of $U(f)$ and the positive definiteness of the Hilbert space give the conditions (4.7)–(4.9) on $E(f)$. The invariance under Euclidean transformation gives (5.15). On physical grounds[28,29] we have assumed the cluster decomposition property (6.1). The existence of $V(g)$ gives a condition which is stated in theorem 10.3, though this is not an explicit condition for $E(f)$. The relativistic invariance gives (9.24) [alternatively (9.16) or (9.18)]. All these conditions are sufficient as well as necessary for the mentioned properties except that (9.24) should be supplemented by conditions concerning domain questions and existence of the operator defined by the Hermitian form (9.23).

We note that the assumed time reversal invariance is crucial for the near uniqueness of $V(g)$ and H.

Theorem 6.2 tells us that for a given representation of $\mathfrak{U}$ or $\mathfrak{U}\mathfrak{B}$ the choice of $E(f)$ (which results from the choice of the cyclic vector Ψ_0) satisfying (5.15) and (6.1) is unique. Combining with theorem 8.2, we see that any specific representation of $\mathfrak{U}$ and $\mathfrak{U}\mathfrak{B}$ which is cyclic relative to $\mathfrak{U}$ is capable of describing essentially one Hamiltonian only. Hence in discussing any specific Hamiltonian the choice of the correct representation of canonical commutation relations is essential.

Theorems (10.2) and (10.3) tell us that $E(f)$ is a Fourier transform of a quasi-invariant measure. If one can find a convenient practical way to handle this measure and the computation of integrals such as (10.11) and (10.12), then this provides a powerful method for constructing examples of $E(f)$ because the condition (4.9) which is most difficult to satisfy is automatically satisfied. For such an application our measure space seems to be too large.

We will discuss some examples of $E(f)$ in a separate paper.

ACKNOWLEDGMENTS

The author would like to thank especially Professor R. Haag, who supervised this work and made many valuable suggestions. The author would like to thank Professor A. S. Wightman for continued interest in the work and for critical discussions, and Dr. J. Cook for valuable information which he has given in connection with mathematical problems in Secs. 3 and 10.

Approximate Solutions of the Bethe-Salpeter Equation*

S. H. Vosko†
Department of Physics, Carnegie Institute of Technology, Pittsburgh, Pennsylvania
(Received June 23, 1960)

A modified representation of the Bethe-Salpeter wave function for scalar particles interacting via a massless scalar field is presented and related to Salpeter's approximate wave function. A procedure for obtaining approximate solutions of the Bethe-Salpeter equation for arbitrary interactions is introduced. The method is based on a variational principle and is capable of high accuracy when used with the trial functions developed here. The choice of trial function is suggested by the important features of the exact solutions for the special case described above. The method is applied to a finite range potential which corresponds to the lowest order approximation to a simple field theory. The results of the calculation suggest that the effect of retardation is large when an interaction is transmitted by a field with mass.

I. INTRODUCTION

A FUNDAMENTAL field theoretic solution of the deuteron problem is one of the long-standing problems of nuclear physics. The possibility of solving this problem was greatly enhanced by the formulation of a relativistic wave equation for two-body systems by several authors.[1,2] The usefulness of the equation has been demonstrated in several accurate calculations of energy levels for hydrogen[3] and for positronium.[4] The validity of the perturbation method used in these calculations is based on special properties of the electromagnetic interaction.

The deuteron differs from the above in two respects:

(a) The interaction is transmitted by a meson field and cannot be accurately approximated by an instantaneous interaction. Therefore the retardation effects associated with the relative time must be treated covariantly. These effects are important because the velocities of the nucleons are large when they are exchanging mesons.

(b) The exact form of the interaction kernel is unknown. Moreover, even if it is assumed that the interaction is transmitted entirely by pions, the interaction kernel cannot be expanded as a rapidly converging series in the meson-nucleon coupling constant. In this work the influence of the retardation effects on the wave function and the interaction strength will be studied. The calculations are done with an interaction kernel which has the form of the lowest order approximation to a simple field theory.

Recently, exact solutions of the Bethe-Salpeter (B-S) equation have been obtained for the case of two scalar particles interacting via a scalar massless field (scalar photons) in the ladder approximation.[5,6] The method used is dependent on certain special properties of the interaction kernel and does not appear to be applicable to other types of interaction kernels. It is the purpose of this paper to introduce methods that are applicable to an arbitrary form of potential. Although a system of two bound scalar particles does not correspond to any system found in nature, it is felt that a thorough understanding of this problem will be useful for solving the spinor problem.

The method introduced here is based on a variational principle. The particular variational principle that is practical from a computational standpoint requires a very judicious choice of trial function to obtain sufficiently precise eigenvalues. A thorough study of the important features of the wave function is necessary to apply the variational principle intelligently. This is done by considering the exact solutions mentioned above. In Sec. II a new representation of the wave function in momentum space is presented and is related to the form used by Wick[5] and the nonrelativistic perturbation approach of Salpeter.[3] The wave function in configuration space is discussed in Sec. III. In particular, the important features of the wave function for nonrelativistic binding energies are presented.

In Sec. IV the accuracy of the variational principle which is especially suited for arbitrary interactions is demonstrated by means of a trial function which is very similar to the wave function obtained from the Salpeter[3] approach. Also, it is shown that the accuracy of the foregoing variational principle is strongly dependent on a cancellation effect when applied to situations where the binding energy is small. This discussion shows how this feature of the principle can be used to obtain high accuracy with rather simple trial functions. Section V is concerned with the details and results of calculations for a potential which has a finite range and is very similar to the Yukawa interaction.

* This work was supported in part by the U. S. Atomic Energy Commission.

† Submitted in December, 1957 in partial fulfillment for the degree of Doctor of Philosophy at Carnegie Institute of Technology. Present address: McMaster University, Hamilton, Ontario.

[1] E. E. Salpeter and H. A. Bethe, Phys. Rev. **84**, 1232 (1951).

[2] J. Schwinger, Proc. Natl. Acad. Sci. **37**, 455 (1951).

[3] E. E. Salpeter, Phys. Rev. **87**, 328 (1952).

[4] R. Karplus and A. Klein, Phys. Rev. **87**, 848 (1952).

[5] G. C. Wick, Phys. Rev. **96**, 1124 (1954); (hereafter referred to as W).

[6] R. E. Cutkosky, Phys. Rev. **96**, 1135 (1954).

II. MOMENTUM SPACE SOLUTIONS

In the notation of W, the B-S equation in momentum space for two scalar particles of equal mass[7] interacting through a massless scalar field in the ladder approximation is

$$F_+F_-\psi(p)=\frac{\lambda}{\pi^2}\int\frac{d^4k\psi(k)}{(p-k)^2}, \tag{1}$$

where p is the relative four-momentum in the center-of-mass frame of reference and the abbreviation

$$F_\pm=(p\pm i\eta)^2+1 \tag{2}$$

has been introduced for convenience. Recall that η is a four-vector with a component in the p_4 direction only and η is related to the binding energy (B.E.) by

$$|\eta|=1-[(\text{B.E.})/2]<1. \tag{3}$$

It was shown in W that the s states solutions of (1) can be represented by

$$\psi(p)=\int_{-1}^{+1}dzg(z)[p^2+2izp\cdot\eta+1-\eta^2]^{-3}, \tag{4}$$

where $g(z)$ satisfies the ordinary differential equation

$$\frac{d^2g}{dz^2}+\frac{\lambda g(z)}{(1-z^2)(1-\eta^2+z^2\eta^2)}=0, \tag{5}$$

with the boundary conditions $g(\pm1)=0$. It should be pointed out that the representation of the wave function by an integral over a single parameter "z" is a property of the particular interaction under consideration. For example, if the interaction is through a scalar field with mass, intergrals over two parameters are required. The unique property which allows a one-parameter representation is that the functional dependence on the momentum is reproduced when the interaction operator [rhs of (1)] is applied to a wave function of the form

$$\psi(p)\sim[p^2+2izp\cdot\eta+1-\eta^2]^{-n}. \tag{6}$$

In the extreme nonrelativistic limit, when $\eta^2\to1$, $g(z)$ for the ground state develops a kink at $z=0$, in fact,

$$g(z)\approx(1-|z|)\quad\text{and}\quad\lambda=2(1-\eta^2)^{\frac{1}{2}}/\pi, \tag{7}$$

which is the nonrelativistic Balmer formula for λ. On substituting (7) for $g(z)$ in (4), the integration over z can be performed resulting in

$$\psi(p)\approx[F_+F_-(p^2+1-\eta^2)]^{-1}, \tag{8}$$

where $F_\pm$ is defined in (2).

Equation (8) suggests a new representation of the wave function

$$\psi(p)=[F_+F_-]^{-1}\int_{-1}^{+1}dzf(z)[p^2+2izp\cdot\eta+1-\eta^2]^{-1}, \tag{9}$$

which is especially appropriate in the nonrelativistic range. Following the method outlined in W, the condition that (9) satisfy (1) requires f to be a solution of

$$\frac{d^2}{dz^2}[Q(z)f(z)]+\frac{\lambda f(z)}{(1-z^2)}=0, \tag{10}$$

where $Q(z)=(1-\eta^2+z^2\eta^2)$ and $f(\pm1)=0$. On comparing Eqs. (5) and (10) one can see that $f(z)=Cg(z)/Q(z)$, where C is a function of η only. Substituting this expression for f in (9), the B-S wave function is

$$\psi(p)=[F_+F_-]^{-1}\int_{-1}^{+1}dzg(z)\times CQ^{-1}(z)[p^2+2izp\cdot\eta+1-\eta^2]^{-1}. \tag{11}$$

In the extreme nonrelativistic limit, $CQ^{-1}(z)$ is approximately $\delta(z)$ and (11) reduces to (8). However, (11) is a more useful representation of the wave function than (4), since a large part of the important asymmetric momentum dependence is exactly accounted for in the $[F_+F_-]^{-1}$ factor of (11). Moreover, for the ground-state solution, which is our main interest, $g(z)$ is a slowly varying function in comparison to $Q^{-1}(z)$. Thus, even when $\eta^2=0.95$ (which is quite relativistic), with any reasonable choice for $g(z)$, (11) gives a very accurate wave function.

In the light of (11) it is interesting to examine the wave functions used by Salpeter[3] and by Karplus and Klein[4] in their respective work involving the electromagnetic interaction. The basis of their method is to note that essentially all of the binding is due to the instantaneous part of the interaction. With this in mind they approximate the true interaction kernel by the instantaneous part for which they are able to obtain a solution of (1). The difference between the instantaneous part and the true interaction kernel is treated by perturbation theory. The assumption of an instantaneous interaction is equivalent to dropping $(p_4-k_4)^2$ on the right-hand side of (1) which can then be written

$$\psi_s(p)=[F_+F_-]^{-1}\frac{\lambda}{\pi^2}\int d^4k\frac{\psi_s(k)}{(\mathbf{p}-\mathbf{k})^2}. \tag{12}$$

After Salpeter[3] we define a three-dimensional wave function,

$$\mathcal{Y}(\mathbf{p})=\int_{-\infty}^{+\infty}dp_4\psi_s(\mathbf{p},p_4). \tag{13}$$

Since the interaction does not depend on k_4, we can immediately integrate the right-hand side of (12) over k_4, which gives

$$\psi_s(p)=[F_+F_-]^{-1}\frac{\lambda}{\pi^2}\int d^3k\frac{\mathcal{Y}(\mathbf{k})}{(\mathbf{p}-\mathbf{k})^2}. \tag{14}$$

This means that the p_4 dependence of ψ_s is given com-

[7] Natural units $\hbar=c=m=1$ are used throughout.

pletely by the $[F_+F_-]^{-1}$ factor in (14). An equation for $\mathcal{Y}(\mathbf{p})$ is obtained by integrating both sides of (14) over p_4. The result is

$$\mathcal{Y}(\mathbf{p})=(\mathbf{p}^2+1)^{\frac{1}{2}}(\mathbf{p}^2+1-\eta^2)^{-1}\frac{\lambda}{2\pi}\int d^3k\frac{\mathcal{Y}(\mathbf{k})}{(\mathbf{p}-\mathbf{k})^2}, \quad (15)$$

which is identical to the Schrödinger equation in momentum space except for the extra factor $(\mathbf{p}^2+1)^{-\frac{1}{2}}$. However, in the extreme nonrelativistic case, where one expects the above approximations to be valid, the important region of $\mathbf{p}^2$ is of order $(1-\eta^2)\ll 1$. Therefore, the $(\mathbf{p}^2+1)^{-\frac{1}{2}}$ can be neglected in a first approximation and then treated as a perturbation. For the ground state the solution of (15) is

$$\mathcal{Y}(\mathbf{p})=\tfrac{1}{2}\pi(\mathbf{p}^2+1-\eta^2)^{-2}. \quad (16)$$

To obtain $\psi_s(p)$, substitute for $\mathcal{Y}$ in (14) which gives

$$\psi_s(p)=[F_+F_-]^{-1}(\mathbf{p}^2+1-\eta^2)^{-1}. \quad (17)$$

Before comparing with the exact solution (11) it is worth nothing that $(1-\eta^2)$ is 1.3×10^{-5} for positronium. For binding energies of this order of magnitude, $CQ^{-1}(z)$ can be approximated by $\delta(z)$ and (11) gives a wave function which is very similar to (17) except that the last factor $(\mathbf{p}^2+p_4{}^2+1-\eta^2)^{-1}$ is replaced by $(\mathbf{p}^2+1-\eta^2)^{-1}$. At first sight this difference might appear significant; however, $[F_+F_-]^{-1}$ is a highly peaked function around $p_4=0$. In fact, it is very nearly proportional to $F(\mathbf{p})\delta(p_4)$. This can be seen by writing

$$F_+F_-=[(\mathbf{p}^2+p_4{}^2+1-\eta^2)^2+4\eta^2p_4{}^2], \quad (18)$$

and noting that the important region of p^2 is of order or less than a few times $(1-\eta^2)$. In this region $[F_+F_-]^{-1}$ decreases much more rapidly in the p_4 direction than in the $\mathbf{p}$ direction. This shows that the wave function given by the Salpeter approximation is quite good in the extreme nonrelativistic range.

III. CONFIGURATION SPACE WAVE FUNCTIONS

The wave function in configuration space will be denoted by $\mathfrak{X}(x)$ and is related to the momentum wave function by the Fourier transform

$$\mathfrak{X}(x)=(2\pi)^{-2}\int d^4pe^{ip\cdot x}\psi(p), \quad (19)$$

where $x=(\mathbf{r},x_4)$ and $p\cdot x=\mathbf{p}\cdot\mathbf{r}+p_4x_4$. To investigate the solutions around the origin in configuration space we must consider the partial differential equation for $\mathfrak{X}(x)$. Taking the Fourier transform of (1) gives

$$\{[-\Box+(1-\eta^2)]^2-4\eta^2\partial^2/\partial x_4{}^2\}\,\mathfrak{X}(x)=\lambda V(R)\mathfrak{X}(x), \quad (20)$$

where the potential $V(R)$ is

$$V(R)=4/R^2, \quad R=(x_\mu x_\mu)^{\frac{1}{2}}. \quad (21a)$$

If the field transmitting the interaction has a mass μ, the potential is

$$V(R)=4\mu R^{-1}K_1(\mu R), \quad (21b)$$

K_1 being a modified Hankel function. The singularity of V at the origin is given by (21a) for both cases. Owing to the asymmetry of the differential operator, the solution must be expressed as a sum of four-dimensional spherical harmonics C_n,

$$\mathfrak{X}(x)=\textstyle\sum_n H_n(R)C_n. \quad (22)$$

The radial functions satisfy a system of coupled fourth-order differential equations. However, the term which produces the coupling is second order, so that the indicial equation which determines the nature of the solution around the origin comes from the fourth-order part of the operator. The equation we must analyze is

$$\left\{\left[\frac{1}{R^3}\frac{\partial}{\partial R}\left(R^3\frac{\partial}{\partial R}\right)\right]^2+\mathcal{O}(R^{-2})+\mathcal{O}(R^0)\right\}H(R)=0. \quad (23)$$

Equation (23) is of the Fuchsian type. A general discussion is given in Ince.[8] The leading terms of the four independent solutions are

$$H_1(R)=R^2(1+b_2R^2+b_4R^4+\cdots) \quad (24a)$$

$$H_2(R)=1+\cdots R^2+\cdots R^2\ln R+\cdots \quad (24b)$$

$$H_3(R)=\ln R+\cdots R^2(\ln R)^2 \quad (24c)$$

$$H_4(R)=(\ln R)^2+\cdots R^2(\ln R)^3. \quad (24d)$$

On assuming that the function and its first derivative must be defined at $R=0$, the solution is then a linear combination of $H_1(R)$ and $H_2(R)$. It is useful to note that the derivative of $H_1(R)$ and $H_2(R)$ vanishes at the origin. Since this is a general result, we shall make our approximate functions in Sec. V satisfy this condition.

To obtain more insight into the important characteristics of nonrelativistic wave functions so that good choices can be made for the variational calculation it is useful to study the configuration space wave function corresponding to (4). The wave function $\mathfrak{X}(x)$ follows directly from the Fourier transform which yields

$$\mathfrak{X}(x)=8^{-1}\int_{-1}^{+1}dzg(z)Q^{-\frac{1}{2}}(z)e^{z(x\cdot\eta)}RK_1(Q^{\frac{1}{2}}(z)R), \quad (25)$$

where $K_1(y)$ is the modified Hankel function normalized according to

$$K_1(y)=y^{-1} \quad \text{as} \quad y\to 0 \quad (26a)$$

with asymptotic behavior

$$K_1(y)\approx(\pi/2)^{\frac{1}{2}}y^{-\frac{1}{2}}e^{-y} \quad \text{for large } y. \quad (26b)$$

[8] E. L. Ince, *Ordinary Differential Equations* (Dover Publications, New York, 1950).

Our problem is to find the region of "z" which is most significant for obtaining a proper nonrelativistic wave function. To answer this question a criterion for an acceptable wave function must be established. Recall that in the nonrelativistic theory of the two-body problem the relative time does not enter. The wave function decays like $\exp[-(1-\eta^2)^{\frac{1}{2}}r]$ for all values of the relative time. Thus a reasonable requirement on the wave function obtained from the B-S equation is that it decays like $\exp[-(1-\eta^2)^{\frac{1}{2}}r]$ over a wide range of relative time values (i.e., for $|x_4| \lesssim (1-\eta^2)^{-\frac{1}{2}}$). This means that it tends to zero much more slowly than $\exp[-(1-\eta^2)^{\frac{1}{2}}x_4]$ in the relative time direction.

For the particular point which we wish to bring out it is best to rewrite (25) as

$$\mathfrak{X}(x)=8^{-1}\int_{-1}^{+1} dz[g(z)Q^{-1}(z)]e^{z(x\cdot\eta)}\varphi(y), \quad (27)$$

where

$$\varphi(y)=yK_1(y) \quad \text{and} \quad y=Q^{\frac{1}{2}}(z)R. \quad (28)$$

The wave function $\mathfrak{X}(x)$ can be interpreted as a sum of the functions $e^{z(x\cdot\eta)}\varphi(y)$, indexed by the parameter z, with amplitudes $g(z)Q^{-1}(z)$. The advantage of writing the wave function in the form (27) is that each of the functions $e^{z(x\cdot\eta)}\varphi(y)$ has the value unity when $R=0$. Thus when we inquire which region of z gives the main contribution to the wave function for $R<(1-\eta^2)^{-\frac{1}{2}}$, we immediately see that the functions corresponding to z values less than a few times $(1-\eta^2)^{\frac{1}{2}}$ are most important because the amplitude factor $g(z)Q^{-1}(z)$ is very peaked around $z=0$. This establishes the importance of the small values of z. However, we must investigate to see whether these values of z are sufficient to describe the wave function for large values of R [i.e., $r>(1-\eta^2)^{-\frac{1}{2}}$]. The exponential part of the asymptotic expression of $\mathfrak{X}(x)$ follows directly from (26b);

$$\mathfrak{X}(x)\sim\int_{-1}^{+1} dz[g(z)Q^{-1}(z)] \times\exp\{-[Q^{\frac{1}{2}}(z)-z\eta\cos\theta_4]R\}, \quad (29)$$

where $\cos\theta_4=x_4/R$. Since $Q^{\frac{1}{2}}(z)$ is always greater than $(z\eta)$, the wave function satisfies the boundary condition $\mathfrak{X}\to 0$ as $R\to\infty$. For a specified direction ($\cos\theta_4$ const) the component with slowest exponential decay corresponds to the value of z which makes $[Q^{\frac{1}{2}}(z)-z\eta\cos\theta_4]$ a minimum. The value of z which achieves this is

$$z_m\eta=(1-\eta^2)^{\frac{1}{2}}\cot\theta_4. \quad (30)$$

However, if $\cos\theta_4>\eta$, z_m as given by (30) is outside the interval of permissible z values and the minimum occurs for $z_m=\pm 1$. The two cases can be summarized as follows:

(i) $$|\cos\theta_4|<\eta; \quad z_m\eta=(1-\eta^2)^{\frac{1}{2}}\cot\theta_4, \quad (31a)$$

$$Q^{\frac{1}{2}}(z_m)-z_m\eta\cos\theta_4=(1-\eta^2)^{\frac{1}{2}}\sin\theta_4, \quad (31b)$$

$$\mathfrak{X}(x)\sim\exp\{-(1-\eta^2)^{\frac{1}{2}}R\sin\theta_4\} =\exp\{-(1-\eta^2)^{\frac{1}{2}}r\}. \quad (31c)$$

(ii) $$|\cos\theta_4|>\eta; \quad z_m=\pm 1, \quad (32a)$$

$$Q^{\frac{1}{2}}(z_m)-z_m\eta\cos\theta_4=1-\eta|\cos\theta_4|, \quad (32b)$$

$$\mathfrak{X}(x)\sim\exp\{-[1-\eta|\cos\theta_4|]R\}. \quad (32c)$$

For the extreme nonrelativistic limit (i.e., $\eta\gtrsim 0.995$), (31) holds for all angles θ_4 except a narrow arc $\Delta\theta_4\sim(1-\eta^2)^{\frac{1}{2}}$ around the relative time direction, and (31c) satisfies the general criterion discussed above. Thus to produce a wave function with the proper asymptotic form over most of the plane, values of $|z\eta|\lesssim$ a few times $(1-\eta^2)^{\frac{1}{2}}$ must be included. On recalling the previous discussion for small R, we conclude that these values of z are sufficient to produce a wave function for all values of R. These conclusions can be corroborated by considering the wave function in momentum space.

It is interesting to note that the asymmetry of the wave function (25) is given by the rather simple $e^{z(x\cdot\eta)}$ factor. We will make use of the particular way the asymmetric term enters in choosing trial functions in Sec. V.

In W it was mentioned that an adiabatic separation of the configuration space Eq. (20) was possible in the extreme nonrelativistic range. Actually the method does not give the correct form of λ vs η [i.e., Eq. (7)]; however, because of its pedagogical value the method will be presented. With the potential (21a) and the space and relative time coordinates written explicitly (20) becomes

$$\left\{\left[-\nabla^2-\frac{\partial^2}{\partial x_4^2}+(1-\eta^2)\right]^2-4\eta^2\frac{\partial^2}{\partial x_4^2}\right\}\mathfrak{X}(\mathbf{r},x_4) =\frac{4\lambda}{r^2+x_4^2}\mathfrak{X}(\mathbf{r},x_4). \quad (33)$$

To motivate the separation of variables, let

$$\mathbf{r}=(1-\eta^2)^{\frac{1}{2}}\boldsymbol{\varrho} \quad \text{and} \quad x_4=(1-\eta^2)^{\frac{1}{2}}\tau. \quad (34)$$

Then in terms of the new variables (33) can be written as

$$\left\{\left[-\nabla_\varrho^2-\frac{\partial^2}{\partial\tau^2}+1\right]^2-\frac{4\eta^2}{1-\eta^2}\frac{\partial^2}{\partial\tau^2}\right\}\mathfrak{X}(\boldsymbol{\varrho},\tau) =\frac{4\lambda}{(1-\eta^2)(\varrho^2+\tau^2)}\mathfrak{X}(\boldsymbol{\varrho},\tau). \quad (35)$$

In the nonrelativistic limit when $\eta^2\to 1$, the coefficient of $\partial^2/\partial\tau^2$ is a large number while all other coefficients

are of order one. The situation is analogous to the problem of separating the electronic and nuclear motion in molecules. This suggests a solution of the form

$$\chi(\varrho,\tau)=v(\varrho)u(\varrho,\tau), \tag{36}$$

where u is the solution of

$$-\frac{4\eta^2}{(1-\eta^2)}\frac{\partial^2 u}{\partial\tau^2}=\frac{4\lambda}{(1-\eta^2)(\varrho^2+\tau^2)}u-\xi(\rho)u. \tag{37}$$

In molecules the nuclear motion is adiabatic with respect to the electronic motion; here we treat the spatial motion as adiabatic with respect to the relative time motion.

The ρ dependence of $\xi(\rho)$ follows directly from (37) by use of the transformation $\tau=\rho y$ and is

$$\xi(\rho)=\frac{4}{1-\eta^2}\frac{C(\lambda)}{\rho^2}. \tag{38}$$

The differential equation for u in the new variable y is

$$\frac{d^2u}{dy^2}+\frac{\lambda u}{1+y^2}=Cu, \tag{39}$$

where the η^2 in the numerator of the left-hand side of (37) has been set equal to one. In the spirit of an adiabatic separation, when (41) is inserted into (35) to obtain the differential equation for $v(\rho)$, certain terms are neglected. The derivatives of u with respect to ρ are dropped and

$$\left[-\nabla_\rho^2-\frac{\partial^2}{\partial\tau^2}+1\right]$$

is replaced by

$$[-\nabla_\rho^2+1].$$

Then using (37) and (38), the v equation is

$$[-\nabla_\rho^2+1]^2v=\frac{4C}{(1-\eta^2)\rho^2}v. \tag{40}$$

Although the foregoing approximations appear reasonable and consistent, they are not correct in the region around the origin (i.e., $R\lesssim 1$). This will be shown by solving (39) and (40) and then examining the wave function $\chi(\mathbf{r},t)$ obtained from their solutions.

The reduction of (40) into the form of a familiar eigenvalue problem can be done most easily in momentum space. Let $\omega(\varrho)$ be the three-dimensional Fourier transform of $v(\varrho)$. Then in momentum space (40) is

$$[\mathbf{p}^2+1]^2\omega(\mathbf{p})=\Lambda(4\pi)^{-1}\int d^3k\omega(\mathbf{k})/|\mathbf{p}-\mathbf{k}|, \tag{41}$$

where $\Lambda=4C/(1-\eta^2)$. Noting that the kernel of (41) is the Green's function of the Laplacian operator, the integral equation can be reduced to a partial differential equation by applying the Laplacian to both sides of (41). This results in

$$\nabla_p^2\{[\mathbf{p}^2+1]^2\omega(\mathbf{p})\}=-\Lambda\omega(\mathbf{p}). \tag{42}$$

Since we are concerned with the ground state of the B-S equation, only s states need be considered which reduce (42) to

$$\frac{1}{p}\frac{d^2}{dp^2}\{p[p^2+1]^2\omega(p)\}+\Lambda\omega(p)=0. \tag{43}$$

To complete the determination of Λ the boundary conditions on $\omega(p)$ must be investigated.

The boundary conditions on the function $v(\varrho)$ are the usual ones associated with the Schrödinger equation. That is, v decays exponentially for large values of ρ, and v is defined for $\rho=0$. The former condition implies that ω is defined for $p=0$; therefore,

$$p\omega(p)=0 \quad \text{when} \quad p=0. \tag{44}$$

The boundary condition for large values of p can be obtained from the integral Eq. (41) and the condition that $[p\omega(p)]$ tends to zero for large values of p, which is necessary for $v(p)$ can be defined at the origin. It then follows that

$$\{p[p^2+1]^2\omega(p)\}\rightarrow \text{const as } p\rightarrow\infty. \tag{45}$$

With these boundary conditions the solution of (43) is

$$\omega(p)=[p^2+1]^{-\frac{1}{2}} \tag{46}$$

corresponding to the eigenvalue $\Lambda=3$. The wave function v in configuration space is

$$v(\mathbf{r})=rK_1[(1-\eta^2)^{\frac{1}{2}}r], \tag{47}$$

where we have returned to the original coordinates.

To complete the determination of the wave function and the eigenvalue λ, Eq. (38) must be considered. It can be shown by a straight-forward analysis that (38) has two regular singular points ($y=\pm i$), and an essential singularity at $y=\infty$. Thus the equation cannot be solved in terms of well-known functions. However, it is easy to see that for small values of λ

$$C\approx\pi^2\lambda^2/4+\mathcal{O}(\lambda^4), \tag{48}$$

and for large values of y the function u behaves like

$$u(y)\approx\exp[-C^{\frac{1}{2}}|y|]. \tag{49}$$

On combining (48) with the relation connecting C and Λ, one obtains

$$\lambda=(\tfrac{3}{4})^{\frac{1}{2}}2(1-\eta^2)^{\frac{1}{2}}/\pi \tag{50}$$

where higher-order terms of $(1-\eta^2)$ have been neglected. This is a very curious result; λ has the correct functional dependence on η^2 but the coefficient is wrong by the factor $(\frac{3}{4})^{\frac{1}{2}}$. The corresponding wave function is

$$\chi(\mathbf{r},x_4)=\{rK_1[(1-\eta^2)^{\frac{1}{2}}r]\}u(x_4/r), \tag{51}$$

where u is an even function and behaves like

$$u \approx \exp\{-(\tfrac{3}{4})^{\frac{1}{2}}(1-\eta^2)^{\frac{1}{2}}|x_4|/r\} \quad \text{for} \quad |x_4| \gg r. \tag{52}$$

The wave function is reasonable in the region $r \gtrsim |x_4|$ but is very poor in the region $r < |x_4|$. It is especially bad around the origin where the function should be very nearly spherically symmetric. Instead it decreases sharply in the x_4 direction. As pointed out earlier the error arises from dropping the $\partial^2/\partial\tau^2$ term to obtain the v Eq. (42). This approximation, which was made to separate Eq. (35), is good for large values of R since there the asymmetric term

$$\{[4\eta^2/(1-\eta^2)]\partial^2/\partial\tau^2\}$$

dominates, but this is not the case near the origin. These statements can be corroborated by considering the momentum space wave functions of Sec. II. The asymmetric p_4 term dominates for $p^2 \lesssim (1-\eta^2)$; this corresponds to large values of R. However, for large values of p the wave function is spherically symmetric. Therefore the wave function in configuration space should be spherically symmetric around the origin.

IV. VARIATIONAL METHOD

The reduction of the B-S equation into a one-dimensional problem is only possible for the particular interaction due to a scalar massless field. To investigate solutions for arbitrary interactions one must resort to other procedures. The method introduced in Sec. V is based on a variational principle and is applicable to any form of potential.

There are several variational principles which are related through the variation-iteration method[9] for solving eigenvalue problems. Their formulation is simple and straightforward.[10] The difficulties arise in performing the necessary integrations with reasonable trial functions. Although the iterated forms can be used with crude trial functions, the integrals become completely out of hand with even the simplest choice for the B-S equation. Thus the primary concern in choosing the variational principle is that it gives integrals which can be evaluated, without excessive labor, when used with nontrivial trial functions. For arbitrary interactions, only the orthodox (noniterated) variational principle satisfies this condition. The disadvantage of this approach is that the main emphasis is put on the choice of trial wave functions. In this respect the analysis of wave functions in Secs. II and III will be very useful.

For purposes of discussion it is most convenient to employ general operator notation. The B-S equation is

$$\mathcal{F}\psi = \lambda \mathcal{U}\omega, \tag{53}$$

where $\mathcal{F}$ represents the operator on the left-hand side of (1) or (20) and $\mathcal{U}$ is the interaction operator. Since $\mathcal{F}$ and $\mathcal{U}$ are positive-definite and self-adjoint it follows that

$$\lambda = \text{minimum}\{(\psi,\mathcal{F}\psi)/(\psi,\mathcal{U}\psi)\} \tag{54}$$

is a variational principle for λ.

To appreciate the usefulness of the variational principle, its accuracy can be tested for the scalar massless field. A basis for comparison can be established by solving the ordinary differential equation (5). One must resort to numerical methods because the equation has four singular points. All calculations will be done for $\eta^2 = 0.99$ which is a rather critical test since neither the extreme relativistic nor the extreme nonrelativistic approximation is valid. Direct numerical integration gives[10]

$$\lambda = 0.0838 \pm 0.0001.$$

It is also of interest to consider the solution of (5) by a variational procedure. With the simple trial function

$$g_1(z) = (1-|z|) + b(1-z^2), \tag{55}$$

the minimum occurs for $b = 0.491$ yielding a value for λ of 0.0858 which is about $2\frac{1}{2}\%$ too large. If the extreme nonrelativistic choice is made for g_1 (i.e., $b=0$) the value obtained for λ is 0.0888. A better trial function which eliminates the cusp around $z=0$ of the function $(1-|z|)$ is

$$g_2(z) = \begin{cases} \left(1 - \dfrac{a}{2} - \dfrac{z^2}{2a}\right) & |z| < a, \\ (1-|z|) & |z| > a. \end{cases} \tag{56}$$

Minimizing with respect to "a" gives an upper bound of 0.08414 to λ with $a = 0.2$ which is a considerable improvement over $g_1(z)$.

To test the accuracy of (54) we shall use the extreme nonrelativistic wave function (8). By a straightforward integration the numerator of (54) can be evaluated and yields

$$\pi^2[4\eta^3(1-\eta^2)^{\frac{1}{2}}]^{-1}\{\sin^{-1}\eta - \eta(1-\eta^2)^{\frac{1}{2}}\}. \tag{57}$$

The denominator is more complicated and the integrals cannot be expressed in terms of simple functions. It can be expressed as

$$\pi^2[4\eta^3(1-\eta^2)^{\frac{1}{2}}]^{-1} \times \left\{[\eta(1-\eta^2)^{-\frac{1}{2}}\sin^{-1}\eta + \ln 2(1-\eta^2)^{\frac{1}{2}}]\sin^{-1}\eta - \int_0^1 \frac{dz}{z(1+z)} \tan^{-1}\frac{z\eta}{(1-\eta^2)^{\frac{1}{2}}}\right\}. \tag{58}$$

In the extreme nonrelativistic limit, $\eta^2 \to 1$, the arc sine terms dominate in both expressions. On taking $\sin^{-1}\eta \approx \pi/2$, the quotient tends to $2(1-\eta^2)^{\frac{1}{2}}/\pi$ as it should. The terms dropped represent relativistic correc-

[9] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill Book Company, Inc., New York, 1953).

[10] For numerical details see S. H. Vosko, Ph.D. thesis, Carnegie Institute of Technology, 1957.

tions. To compare the variational result for $\eta^2=0.99$ with the exact value given in (54), the integral in (58) was evaluated by Gauss' method using the six-point formula. The final result is 0.0843, which is only 0.6% larger than the exact result. It is interesting to note that this is very much better than using the function $g(z)=(1-|z|)$ to obtain λ from (5) by a variational procedure even though the wave function (8) corresponds to this choice for $g(z)$. This shows that the variational principle expressed in terms of the B-S wave function is not very sensitive to the choice of $g(z)$, and is capable of giving very accurate results.

Before considering trial functions which are particularly suitable for the variational principle it is important to understand how the minimization is achieved. It will be shown that the minimization is strongly dependent upon a cancellation of positive and negative terms in the numerator of (54). How this cancellation comes about is not obvious since the operators are positive-definite. For the purposes of illustration we shall consider the scalar massless field which is an extreme example, since the interaction strength tending to zero with the binding energy is dependent on this cancellation effect. Substitution of (4) into (54) gives

$$\lambda=\text{minimum}\left\{\int_{-1}^{+1}dz\int_{-1}^{+1}d\zeta g(z)\mathfrak{F}(z,\zeta)g(\zeta)\Big/\int_{-1}^{+1}dz\int_{-1}^{+1}d\zeta g(z)\mathfrak{V}(z,\zeta)g(\zeta)\right\}, \quad (59)$$

where

$$\mathfrak{F}(z,\zeta)=\int d^4p[p,-z]^{-3}[F_+F_-][p,\zeta]^{-3} \quad (60)$$

$$\mathfrak{V}(z,\zeta)=\pi^{-2}\int d^4p\int d^4k\times[p,-z]^{-3}[(p-k)^{-2}][k,\zeta]^{-3} \quad (61)$$

and the abbreviation

$$[p,z]=p^2+2izp\cdot\eta+1-\eta^2$$

has been used. It follows directly from these definitions that the kernels $\mathfrak{F}(z,\zeta)$ and $\mathfrak{V}(z,\zeta)$ are symmetric and positive definite. Equation (59) can be regarded as a variational principle for λ in terms of the functional $g(z)$. The η dependence of the kernels can be exhibited explicitly by the change of variables

$$z\eta=x(1-\eta^2)^{\frac{1}{2}},\quad \zeta\eta=y(1-\eta^2)^{\frac{1}{2}}, \quad (62)$$

and the change of the p and k scale by $(1-\eta^2)^{\frac{1}{2}}$. Then $[p,z]$ becomes $(1-\eta^2)(p^2+2ixp_4+1)$. The change in variables from (z,ζ) to (x,y) in Eq. (59) introduces a factor $\eta^{-2}(1-\eta^2)$ and the limits of integration become $\pm\epsilon^{-1}$ where $\epsilon=(1-\eta^2)^{\frac{1}{2}}/\eta$. Collecting powers of $(1-\eta^2)$, the numerator and the denominator of (59) can be expressed in the form

$$\eta^{-2}(1-\eta^2)^{-2}\int dx\int dy g(x)\mathfrak{F}_1(x,y)g(y)+\eta^{-2}(1-\eta^2)^{-1}\int dx\int dy g(x)\mathfrak{F}_2(x,y)g(y), \quad (63)$$

and

$$\eta^{-2}(1-\eta^2)^{-2}\int dx\int dy g(x)\mathfrak{V}_1(x,y)g(y), \quad (64)$$

respectively, where all the kernels are positive definite. For the extreme nonrelativistic case where $\epsilon^{-1}\gg1$, the major contributions to (63) and (64) come from the region $|x|\lesssim1$ and $|y|\lesssim1$, where $g(x)$ is essentially constant. Then it would appear that both quantities were of order $(1-\eta^2)^{-2}$ and the quotient tended to a constant. This contradicts the previous result of $\lambda\sim(1-\eta^2)^{\frac{1}{2}}$. Comparing (63) and (64) with (57) and (58) we see that the difficulty must be in (63). To reconcile (63) with the previous calculation one must require that

$$\int dx\int dy\mathfrak{F}_1(x,y)\lesssim\epsilon, \quad (65)$$

where the limits of integration are $\pm\epsilon^{-1}$. If the limits $\pm\infty$ are used ($\epsilon\to0$), then the integral should vanish. By direct integration this can be shown to be the case.

The diagonal matrix elements, $\mathfrak{F}(x,x)$, must be positive since $\mathfrak{F}$ is positive-definite. Therefore the off-diagonal elements of $\mathfrak{F}(x,y)$ must be negative so that the positive contribution from the region $x\approx y$ may be canceled. The sources of these negative matrix elements may be understood by examining Eq. (60). Since we are concerned with the ground state our considerations can be specialized to $g(z)$ even. Then it is most convenient to work with a kernel which has this condition explicitly exhibited. Define

$$\mathfrak{F}_s(z,\zeta)=\frac{1}{2}\int d^4p\{[p,-z]^{-3}+[p,z]^{-3}\}[F_+F_-]\times\{[p,-\zeta]^{-3}+[p,\zeta]^{3}\} \quad (66a)$$

which is related to $\mathfrak{F}(z,\zeta)$ by

$$\mathfrak{F}_s(z,\zeta)=\mathfrak{F}(z,\zeta)+\mathfrak{F}(z,-\zeta). \quad (66b)$$

Similarly, the interaction kernel becomes

$$\mathfrak{V}_s(z,\zeta)=\mathfrak{V}(z,\zeta)+\mathfrak{V}(z,-\zeta). \quad (67)$$

By collecting the denominators of an individual bracketed term in (66a), it can be seen that the integrand of (66a) is positive for all values of p when $(z\eta)^2$ is smaller than $[(1-\eta^2)/3]$. Therefore, $\mathfrak{F}_s(z,\zeta)$ is positive when $(z\eta)^2$ and $(\zeta\eta)^2$ are both smaller than $[(1-\eta^2)/3]$. It may be positive for other values of z and ζ but this depends on the details of the integral. For $\mathfrak{F}_s(z,\zeta)$ to be negative, at least one of the terms $z\eta$

or $\zeta\eta$ must be greater than $[(1-\eta^2)/3]^{\frac{1}{2}}$. The matrix elements of $\mathcal{U}_s(z,\zeta)$ are always positive. Expressions for $\mathcal{F}(z,\zeta)$ and $\mathcal{U}(z,\zeta)$ are given in the Appendix.

The main conclusion which may be drawn from this discussion is that the minimum of (59) is achieved by a delicate cancellation due to the complicated structure of the kernel $\mathcal{F}(z,\zeta)$. This effect is essential for determining eigenvalues for nonrelativistic binding energies.

V. APPROXIMATE WAVE FUNCTIONS

This section will be concerned with the problem of choosing approximate wave functions that can be used in the variational principle. The choice of functions will be guided by the criteria discussed in Secs. III and IV. Particular care is taken so that the cancellation effect described in the previous section can occur. This is achieved by using a linear combination of functions and varying the amplitude of each function so that a minimum is obtained.

To illustrate the procedure consider the scalar massless field. The form (4) of the exact wave function suggests a trial wave

$$\psi_T=\sum_{z_n} a(z_n)\{[p,-z_n]^{-3}+[p,z_n]^{-3}\}, \qquad (68)$$

where $0\leqslant z_n<1$ and the sum is over a finite number of z_n's. Substituting this trial function in (54) gives

$$\lambda=\text{minimum}\left\{\frac{\sum_{\zeta_n}\sum_{z_n} a(z_n)\mathcal{F}_s(z_n,\zeta_n)a(\zeta_n)}{\sum_{\zeta_n}\sum_{z_n} a(z_n)\mathcal{U}_s(z_n,\zeta_n)a(\zeta_n)}\right\}, \qquad (69)$$

where the $a(z_n)$ and z_n are the variational parameters. The ideal procedure would be to vary all parameters; however, the labor involved would be prohibitive. A more practical procedure is to make a judicious choice of z_n and then vary the $a(z_n)$ to produce a minimum. After some experimentation the following values of z_n were arrived at for the special case $(1-\eta^2)=0.01$:

$z_1\eta$	$z_2\eta$	$z_3\eta$	$z_4\eta$	$z_5\eta$
0.000	0.065	0.173	0.300	0.500.

(70)

The choice of the first two points is primarily concerned with the over-all shape of the wave function in accordance with the discussion of Sec. III. The objective governing the choice of the other three points is to minimize the numerator of (69) by exploiting the negative values of the off-diagonal matrix elements of $\mathcal{F}_s$. It should be noted that large values of $z_n(z_n\sim 1)$ are not needed. For other values of η^2 some readjustment of the points would be required.

With the z_n's fixed the problem of determining the $a(z_n)$ can be reduced to solving the matrix eigenvalue problem

$$\sum_{\zeta_n}\mathcal{F}_s(z_n,\zeta_n)a(\zeta_n)=\lambda\sum_{\zeta_n}\mathcal{U}_s(z_n,\zeta_n)a(\zeta_n), \qquad (71)$$

where $\mathcal{F}_s(z_n,\zeta_n)$ and $\mathcal{U}_s(z_n,\zeta_n)$ are positive-definite symmetric matrices and $a(\zeta_n)$ is the eigenvector. The solution of (71) for the ground state is straightforward. The functional form of $\mathcal{F}_s$ and $\mathcal{U}_s$ are given in the Appendix and a tabulation of the matrices for particular values of z_n are given in Tables I and II. For the $(z_n\eta)$'s given in (70) the eigenvector is (1.0000000; 2.5463714; 4.5260774; 1.9318722; 7.9293120) and the corresponding eigenvalue is 0.083920, which is only 0.12% greater than the exact value. It is interesting to note that this method gives more accurate results than the wave function (8) in the variational principle. To investigate whether a value of $(z_n\eta)$ greater than 0.5 would make an appreciable improvement, a sixth value of $(z_n\eta)$ equal to 0.7 was added. The resulting eigenvalue is 0.083875, a very small improvement over the previous result, which shows the values of z_n given in (70) are adequate.

Owing to the similarity of Eqs. (4) and (68) there should be some connection between the values of $a(z_n)$ which produce the minimum in (69), and the function $g(z)$. The values of $a(z_n)$ and the corresponding values of $g(z)$ at the points z_n cannot be compared directly. A reasonable assumption is that the sum in (68) could be thought of as a numerical integration of (4). Then $a(z_n)$ is equal to $g(z_n)\Delta z_n$ where Δz_n varies from point to point. A rough comparison can be made by constructing a histogram from the values of $a(z_n)$, where the height of each block is equal to $a(z_n)/\Delta z_n$. The choice of Δz_n is rather arbitrary but any reasonable choice does not alter the histogram appreciably. The $g(z)$ plotted for comparison in Fig. 1 is from the variational solution (56). The histogram is normalized such that its area is equal to the area under $g(z)$.

TABLE I. $\mathcal{F}_s(z,\zeta)$ matrix for $(1-\eta^2)=0.01$.[a]

$z\eta\backslash\zeta\eta$	0.000	0.065	0.173	0.300	0.500
0.000	0.27733333	0.096959036	−0.039726098	−0.027291486	−0.010824967
0.065	0.096959036	0.054529575	−0.0049262454	−0.011011672	−0.0062225370
0.173	−0.039726098	−0.0049262454	0.016494543	0.0075587743	0.0014103839
0.300	−0.027291486	−0.011011672	0.0075587743	0.0069608986	0.0032060624
0.500	−0.010824967	−0.0062225370	0.0014103839	0.0032060624	0.0025363315
0.700	−0.0050435419	−0.0033574601	0.000032635305	0.0014553906	0.0016710349

[a] *Note.* The factor $\pi^2/4(1-\eta^2)^{\frac{1}{2}}$ has been omitted.

TABLE II. $v_s(z,\zeta)$ matrix for $(1-\eta^2)=0.01$.[a]

$z\eta \backslash \zeta\eta$	0.000	0.065	0.173	0.300	0.500
0.000	0.66666666	0.43489565	0.11847216	0.035149843	0.0094487446
0.065	0.43489565	0.29371344	0.088466198	0.028416016	0.0081631006
0.173	0.11847216	0.088466198	0.035725803	0.014529605	0.0051085866
0.300	0.035149843	0.028416016	0.014529605	0.0072270637	0.0030615355
0.500	0.0094487446	0.0081631006	0.0051085866	0.0030615355	0.0015698808
0.700	0.0037577238	0.0033631595	0.0023489828	0.0015690780	0.00090878084

[a] *Note.* The factor $\pi^2/4(1-\eta^2)^3$ has been omitted.

A parametric representation of the B-S wave function when the interaction is transmitted by a scalar field with mass is suggested by the analogous problem in nonrelativistic wave mechanics. The ground state solution of the Schrödinger equation for the Coulomb potential is the familiar exponential. The analog of the Coulomb potential for the B-S equation is (21a) and the corresponding wave function is (25). The similar feature of the two solutions to note is that their asymptotic behavior in configuration space is characterized by a single decaying exponential which is precisely the same as the corresponding Green's function (see Appendix of W). On the other hand, the solution of the Schrödinger equation for the Yukawa interaction requires a sum of exponentials with different decay factors. That is, the solution can be represented as

$$\int_{+1}^{\infty} d\alpha f(\alpha) \exp[-\alpha(1-\eta^2)^{\frac{1}{2}}r] \tag{72a}$$

in configuration space, or

$$\int_{+1}^{\infty} d\alpha f(\alpha)[\mathbf{p}^2+\alpha^2(1-\eta^2)]^{-2} \tag{72b}$$

in momentum space. The corresponding example for the B-S equation is a scalar field with mass. By analogy, the B-S wave function for this case may be generated from the known solutions for the scalar massless field by including a sum of various decay factors, that is,

$$\chi(x)=8^{-1}\int_{+1}^{\infty} d\alpha \int_{-1}^{+1} dz \times[g(z,\alpha)\alpha^{-2}Q^{-1}(z)]e^{i(x\cdot\eta)}\omega K_1(\omega), \tag{73}$$

where $\omega=\alpha Q^{\frac{1}{2}}(z)R$. This corresponds to the momentum space wave function[11]

$$\psi(p)=\int_{+1}^{\infty} d\alpha \int_{-1}^{+1} dz g(z,\alpha)[p,z,\alpha]^{-3}, \tag{74}$$

where the abbreviation

$$[p,z,\alpha]\equiv(p+iz\eta)^2+\alpha^2Q(z) \tag{74a}$$

has been introduced for convenience. With the proper choice of $g(z,\alpha)$ Eq. (74) would be an exact solution of the B-S equation

$$F_+F_-\psi(p)=\lambda\pi^{-2}\int d^4k\psi(k)/[(p-k)^2+\mu^2]. \tag{75}$$

According to the procedure used in the foregoing, for a variational calculation the integrals over α and z could be replaced by finite sums, for example,

$$\psi_A(p)=\sum_{\alpha l}\sum_{z_n} a(z_n,\alpha_l)[p,z_n,\alpha_l]^{-3}. \tag{76}$$

The sum over z_n depends on the value of η^2 (i.e., binding energy) under consideration. For $\eta^2=0.99$, the values given in (70) are adequate. The sum over α_l should include $\alpha_l=1$ to ensure agreement with the Green's function for large values of R; other values of α_l could be treated as variational parameters. Substituting (76) in the variation principle would result in an expression very similar to (69) except the matrix elements would be designated by four parameters $(z_n,\alpha_l;\zeta_n,\beta_l)$. The determination of the $a(z_n,\alpha_l)$ which minimize the quotient can be reduced to the familiar secular problem as in the previous case. Considering the results obtained for the scalar massless field one might expect to determine the interaction strength λ to

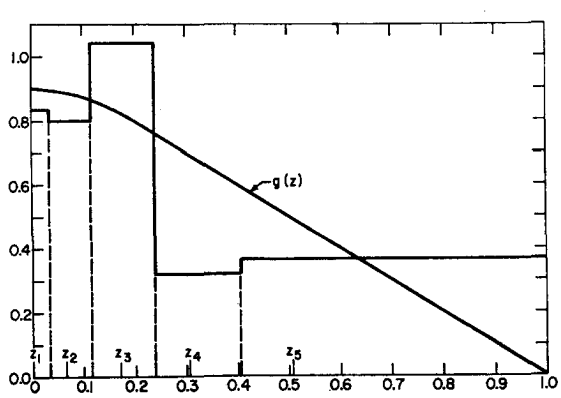

FIG. 1. Histogram for comparison of $a(z_n)$ and $g(z)$. The area under each block is proportional to $a(z_n)$.

[11] A similar parametrization has been used by G. Wanders [Phys. Rev. 104, 1782 (1956)]. Also R. E. Cutkosky and the author have investigated the wave function $[F_+F_-]^{-1}\int d\alpha\int dz \times f(z,\alpha)[p,z,\alpha]^{-1}$; however, the integral equation obtained for $f(z,\alpha)$ is too complicated to be useful.

TABLE III.

P	$PK_1(P)$	Right-hand side of Eq. (77)
0.2	0.955	0.964
0.6	0.782	0.792
1.0	0.602	0.604
2.0	0.280	0.280
3.0	0.120	0.126

within 0.5% for the case (75) using a wave function of the form (76) with two α_l's, of which one would be a variable parameter. Use of the wave function (76) presents a problem because the matrix elements of the potential operator cannot be evaluated in terms of simple functions. At best they can be represented as an integral over a single parameter which could be evaluated numerically if a very precise solution were required. However, Eq. (92) does not warrant such careful treatment since it does not represent any known physical situation.

The effects of a finite range potential can be obtained by a simpler calculation by altering the original potential slightly. The mode of approximation is suggested by writing the potential (21b) in the form

$$4R^{-2}[(\mu R)K_1(\mu R)],$$

and noting that the function in brackets is generally similar to

$$(1+\mu R)e^{-\mu R}.$$

By direct numerical computation it can be shown that for P less than 5

$$PK_1(P)\approx(1+a)^{-1}[(1+P)e^{-P}+a(1+2P)e^{-2P}] \quad (77)$$

where $a=0.66746$. Beyond 5 the value of the function is quite small and relatively unimportant. The accuracy of (94) is demonstrated by the sample values given in Table III. On using (77), the finite range potential which we will consider is

$$U(R)=4R^{-2}(1+a)^{-1}\times[(1+\mu R)e^{-\mu R}+a(1+2\mu R)e^{-2\mu R}]. \quad (78)$$

Also, it is worth noting that the properties

$$PK_1(P)=1 \quad \text{and} \quad d[PK_1(P)]/dP=0$$

when $P=0$ are preserved by the function $(1+P)e^{-P}$.

In the spirit of the preceding approximations for $PK_1(P)$, a reasonable ground-state trial function is [compare with Eq. (73)],

$$\mathfrak{X}_T(x)=\sum_{\alpha l}\sum_{z_n} a(z_n,\alpha_l)[\sqrt{2}\alpha_l^2Q(z_n)]^{-1}\times\{e^{z_n(x\cdot\eta)}+e^{-z_n(x\cdot\eta)}\}\{(1+\omega)e^{-\omega}\}, \quad (79)$$

where

$$\omega=\alpha_lQ^{\frac{1}{2}}(z_n)R.$$

The advantage of this trial function is that the potential matrix elements can be easily evaluated for any potential containing exponentials such as (78). The summations over z_n and α_l are the same as (76) except for the α_l which is to be treated as a variational parameter. It should be noted that the trial function (79) is even in x_4 since it is a ground state function. Also, the radial function $(1+\omega)e^{-\omega}$ satisfies the general condition derived in Sec. III that the first derivative with respect to R vanish at the origin for potentials with the R^{-2} singularity. The corresponding momentum space wave function follows from the Fourier transform of (79)

$$\psi_T(p)=\sum_{\alpha l}\sum_{z_n} 15a(z_n,\alpha_l)\alpha_lQ^{\frac{1}{2}}(z_n)2^{-\frac{1}{2}}\times\{[p,-z_n,\alpha_l]^{-7/2}+[p,z_n,\alpha_l]^{-7/2}\}. \quad (80)$$

It is very similar to (76) except the exponent is $(-\frac{7}{2})$ instead of (-3). The evaluation of matrix elements with these trial functions are straightforward although rather complicated algebraic expressions result.[10]

Potentials with repulsive cores or phenomenological potentials can be expressed in terms of exponentials, and their effects could be readily investigated with trial functions similar to (79).

The scalar massless field can be used to estimate the accuracy of a variational calculation with this trial function. If only $\alpha_l=1$ and $z_n\eta$ of (70) are used the calculation is similar to the previous variational calculation.[10] The result is $\lambda=0.08768$, which is fair considering the simplicity of the trial function although not nearly as good as using the exact type [i.e., $RK_1(R)$] of function. The effect of including a variable α_l is to lower λ to 0.08574, a 2% improvement. In this case $\alpha_1=1$ and α_2 is varied to produce a minimum. For each value of α the tenth-order secular equation is solved for λ and $a(z_n,\alpha_l)$. This ensures that the value obtained for λ is a minimum with respect to the ten independent variable parameters. The minimum occurs for $\alpha_2=1.5$.

For the finite range potential $U(R)$ with $\mu=0.148$ which corresponds to the pion mass, the same set of calculations was performed. With a single $\alpha_l=1$, $\lambda=0.302$. Introducing a second α_l as above, the upper bound on λ decreased to 0.210 which illustrates the importance of a second exponential. This is similar to the Yukawa interaction in the Schrödinger equation where the addition of a second exponential also makes a large improvement. The minimum occurs at $\alpha_2=1.7$ and is quite sensitive to α_2. Using the error in the scalar massless field as a guide, one might expect that the value 0.210 is approximately 2% too large. The interaction strength for a finite range potential could be determined more accurately if the trial function (68) was used instead of (80).

Table IV illustrates the relativistic effects by comparing the just-mentioned results with the values obtained from the Schrödinger equation with the same

TABLE IV.

	Scalar massless field	Finite range potential, $\mu=0.148$
Schrödinger equation	0.0637	0.148
Variation solution of B-S equation	0.0857	0.210

binding energy and the equivalent nonrelativistic potentials. Recall that these calculations correspond to $\eta^2=0.99$. Assuming that the results for the B-S equation are too large by 2%, then the true relativistic result for the finite range potential would be 40% larger than the nonrelativistic result. Although the form of the B-S equation used here does not include the effects of spin, which are quite important in the deuteron, the results show that relativistic effects are very significant. If the true potential for the deuteron has a repulsive core, the relativistic effects are likely to be even larger than the preceding calculations indicate.

ACKNOWLEDGMENTS

The author is indebted to Dr. Gian-Carlo Wick for suggesting this problem and directing it throughout its course. The author wishes also to thank Professor R. E. Cutkosky for many valuable discussions and for reading the manuscript. The numerical calculations were done on an IBM 650 electronic computer at the Carnegie Institute of Technology Computation Center. The author is indebted to Professor Alan J. Perlis, Joe Smith, and Hal Van Zoren of the Computation Center for helpful advice. Part of this work was done while the author held a Theoretical Physics Fellowship from Carnegie Institute of Technology.

APPENDIX

The functions $\mathfrak{F}(z,\zeta)$ and $\mathfrak{V}(z,\zeta)$ defined by Eqs. (60) and (61) may be evaluated in a straightforward manner. Use of Feynman formulas to combine denominators are useful. The results are

$$\mathfrak{F}(z,\zeta)=\left[\frac{\pi^2}{4(1-\eta^2)^3}\right](x+y)^{-4}\times\left\{\frac{1-xy}{1+x^2}+\frac{1-xy}{1+y^2}+4-6(1-\eta^2)(1-xy)-(x+y)^{-1}[\tan^{-1}x+\tan^{-1}y][(1+x^2)+(1+y^2)+4(1-xy)-(1-\eta^2)\times\{2(1+x^2)(1+y^2)+4(1-xy)^2\}]\right\},\quad\text{(A1)}$$

where x and y are defined by

$$x(1-\eta^2)^{\frac{1}{2}}=z\eta\quad\text{and}\quad y(1-\eta^2)^{\frac{1}{2}}=\zeta\eta.$$

The function (A1) appears to have a singularity when $z=-\zeta$; however, it can be shown that the function is defined and has the form

$$\mathfrak{F}(z,-z)=\left[\frac{\pi^2}{4(1-\eta^2)^3}\right]\frac{2}{15(1+x^2)^3}\times\{6(1+x^2)^{-1}+4(1-\eta^2)-5\}.\quad\text{(A2)}$$

For the scalar massless field the potential kernel is

$$\mathfrak{V}(z,\zeta)=\left[\frac{\pi^2}{4(1-\eta^2)^3}\right]\frac{1}{2(x+y)^3}\times\left\{\tan^{-1}x+\tan^{-1}y-\frac{(x+y)(1-xy)}{(1+x^2)(1+y^2)}\right\}.\quad\text{(A3)}$$

Again the singularity for $z=-\zeta$ is only apparent, in fact (A3) simplifies to

$$\mathfrak{V}(z,-z)=\left[\frac{\pi^2}{4(1-\eta^2)^3}\right]3^{-1}(1+x^2)^{-3}.\quad\text{(A4)}$$

Note that in Tables I and II the common factor $\pi^2/4(1-\eta^2)^3$ has been omitted.

Relationship between Systems of Impenetrable Bosons and Fermions in One Dimension*†

M. GIRARDEAU‡
Brandeis University, Waltham, Massachusetts
(Received March 3, 1960)

A rigorous one-one correspondence is established between one-dimensional systems of bosons and of spinless fermions. This correspondence holds irrespective of the nature of the interparticle interactions, subject only to the restriction that the interaction have an impenetrable core. It is shown that the Bose and Fermi eigenfunctions are related by $\psi^B=\psi^F A$, where $A(x_1\cdots x_n)$ is $+1$ or -1 according as the order $pq\cdots r$, when the particle coordinates x_j are arranged in the order $x_p<x_q<\cdots<x_r$, is an even or an odd permutation of $1\cdots n$. The energy spectra of the two systems are identical, as are all configurational probability distributions, but the momentum distributions are quite different. The general theory is illustrated by application to the special case of impenetrable point particles; the one-one correspondence between bosons with this particular interaction and completely noninteracting fermions leads to a rigorous solution of this many-boson problem.

1. INTRODUCTION

IN the following section a very simple and general relationship will be established between one-dimensional systems of impenetrable bosons and fermions. We shall find that the restrictions both to one dimension and to interactions with a completely impenetrable core are essential. Nevertheless, there are at least two motivations for studying such a relationship. First, one is enabled to obtain a rigorous solution of the many-boson problem for the case of impenetrable point particles in a one-dimensional periodic box, and this solution may serve as a useful testing ground for various approximation methods. Second, the relationship for the case of more general interactions may permit comparison of approximation methods designed for Fermi systems with those designed for Bose systems.

The general theory of the Bose-Fermi correspondence is developed in Sec. 2, and is illustrated in Sec. 3 by application to impenetrable point particles in a periodic box, for which the correspondence permits one to obtain a rigorous solution of the many-boson problem by relating it to the trivial problem of a one-dimensional free Fermi gas.

2. PROOF OF THE CORRESPONDENCE

The condition that the interparticle interaction have an "impenetrable core" is most conveniently represented by the following subsidiary condition on the allowable wave functions ψ:

$$\psi(x_1\cdots x_n)=0 \text{ if } |x_j-x_l|\leqslant a, \quad 1\leqslant j<l\leqslant n, \quad (1)$$

where $x_1\cdots x_n$ are the coordinates of the n particles comprising the system, and a is the hard-core diameter. Then the Schrödinger equation is

$$(T+V)\psi=E\psi, \quad (2)$$

where V includes all interactions except the hard cores' and is otherwise completely unrestricted. Consider first any Fermi wave function ψ^F satisfying (2); ψ^F is antisymmetric in the particle coordinates. We define a "unit antisymmetric function" A as follows:

$$A(x_1\cdots n_n)=\prod_{j>l}\operatorname{sgn}(x_j-x_l), \quad (3)$$

where $\operatorname{sgn}(x)$ is the algebraic sign of x; an equivalent definition is that A is $+1$ or -1 according as the order $pq\cdots r$, when the x_j are arranged in the order $x_p<x_q<\cdots<x_r$, is an even or an odd permutation of $1\cdots n$. Then the product

$$\psi^B=\psi^F A \quad (4)$$

is symmetric in the particle coordinates, and hence describes a Bose system provided that the necessary regularity conditions are satisfied. To see that they are, we note that A has discontinuities only at the surfaces $x_j=x_l$, where two particles come together. But ψ^B is continuous even at these surfaces, since ψ^F vanishes there as a result of the Fermi statistics; indeed, it vanishes throughout the region of the hard cores as a result of the subsidiary condition (1). The surfaces $x_j=x_l$ divide the n-dimensional configuration space into $n!$ disjoint regions, in each of which A is constant, equal to either $+1$ or -1. As a result, ψ^B satisfies the Schrödinger equation (2) throughout the allowed portion of configuration space [all $|x_j-x_l|>a(j\neq l)$], by virtue of the fact that ψ^F does; for the same reason, $\psi^B\to 0$ as $|x_j-x_l|\to a$ from above. In the remainder of configuration space (where hard cores overlap), ψ^F and ψ^B are defined by (1). Finally, ψ^B will satisfy boundary conditions of enclosure in a box if ψ^F does, and in the case of odd[1] total number of particles

* Supported in part by U. S. Air Force Office of Scientific Research.

† An abbreviated account of this work was given by M. Girardeau, Bull. Am. Phys. Soc. Ser. II 5, 8 (1960).

‡ Now at Boeing Scientific Research Laboratories, Seattle, Washington.

[1] For the case of odd n, the function $A(x_1\cdots x_n)$ defined by (3) remains well defined if the x_j are interpreted modulo L, in which case A satisfies periodic boundary conditions with periodicity length L. On the other hand, for the case of even n the substitution $x_j\to x_j\pm L$ changes the sign of A. Hence our general theorem on the one-one correspondence is only valid for a system with periodic boundary conditions if n is odd. There are no restrictions on n for boundary conditions of enclosure in a box.

n, it will satisfy periodic boundary conditions if ψ^F does. Upon putting together all the pieces of this rather lengthy verbal proof, we conclude that ψ^B is a solution of the Schrödinger equation (1) satisfying Bose statistics and belonging to the same energy as ψ^F, and satisfying the same boundary and regularity conditions; for the case of periodic boundary conditions, we must add the proviso that the total number of particles be odd. The above proof cannot be generalized to systems of particles moving in three dimensions (or any number of dimensions greater than one) because there does not exist any generalization of the unit antisymmetric function A [Eq. (3)] to the case where the particle coordinates x_j are vectors rather than scalars; this is because the "surfaces" $x_j=x_l$ then fail to divide the configuration space into disjoint regions (they are "lines" rather than "surfaces"); in two or more dimensions one can hold all particles but one fixed and move the remaining particle about throughout the box containing the system without encountering any of the fixed particles, but in one dimension the motion of one particle is blocked by the others. Even for one-dimensional systems, our proof is limited to systems for which the interparticle interaction has a completely impenetrable core, since only then does the Bose wave function defined by (4) have a continuous gradient at the surfaces $x_j=x_l$ [both the wave function and its gradient vanish there because of (1)]. In the degenerate case where the impenetrable core has shrunk to a point ($a=0$), ψ^B has discontinuous gradient at the surfaces $x_j=x_l$, but this is merely a reflection of the singular point interaction; such a discontinuous gradient at $x_j=x_l$ is permitted *only* if there is an impenetrable point core.

The relationship (4) establishes a one-one correspondence between all the Fermi and all the Bose energy eigenfunctions; since $A^2=1$, this correspondence preserves all scalar products. Not only are the energy spectra of the Bose and Fermi systems identical,[2] but also all configurational probability distributions, since these involve only the square of the wave function. On the other hand, the momentum distributions will in general be quite different, since they involve the momentum wave function, and the result of taking the Fourier transform depends on the relative sign of the wave function in various regions of configuration space. One expects the single-particle momentum distribution of the Bose system to reflect a tendency toward Bose-Einstein condensation, while that of the Fermi system should be dominated by the effects of the exclusion principle. We shall find that this is indeed the case for the simple example discussed in Sec. 2.

We conclude our discussion of the general correspondence by noting that the relationship (4) simplifies for the case of the ground state. We note first that for a one-dimensional system the Fermi ground state ψ_0^F vanishes *only* for those configurations $x_1\cdots x_n$ where the hard cores overlap [Eq. (1)]; this follows from (4) and the well-known fact that the Bose ground state ψ_0^B is positive (aside from an arbitrary constant phase factor) except where it is required to vanish by boundary conditions or infinite repulsive interactions.[3] It follows that ψ_0^F has constant sign[4] throughout each of the $n!$ regions into which the configuration space is divided by the surfaces $x_j=x_l$. Thus, since ψ_0^F is antisymmetric, it has the same sign (aside from a constant phase factor) as the unit antisymmetric function $A(x_1\cdots x_n)$ throughout the whole configuration space. As a result, the product $\psi_0^F A$ reduces merely to the absolute value of ψ_0^F, and (4) reduces to

$$\psi_0^B=|\psi_0^F|. \tag{5}$$

3. IMPENETRABLE POINT PARTICLES[5a]

We consider in this section the simplest possible interaction with an impenetrable core, namely the case when the core has shrunk to a point, and there is no other interaction. Then the subsidiary condition (1) reduces to

$$\psi(x_1\cdots x_n)=0 \quad \text{if } x_j=x_l, \quad 1\leqslant j<l\leqslant n. \tag{6}$$

We seek the Bose eigenfunctions ψ^B and energy eigenvalues E of the Schrödinger equation

$$H\psi^B=T\psi^B=\sum_{j=1}^{n}-\frac{\hbar^2}{2m}\frac{\partial^2\psi^B}{\partial x_j^2}=E\psi^B \tag{7}$$

subject to the subsidiary condition (6)[5b] and to periodic boundary conditions with period L.

It is well known that for a system of spinless fermions, point interactions are equivalent to no interaction at all, since the Fermi wave functions automatically vanish when two particles come together. Thus the Fermi energy eigenfunctions ψ^F satisfying the subsidiary condition (6) are just the eigenfunctions of a *free* Fermi gas; the Bose eigenfunctions ψ^B satisfying (6) and (7) are then given in terms of the free Fermi gas eigenfunctions ψ^F by our general theorem (4).

The ground state ψ_0^F of the free Fermi gas is a Slater determinant of n free-particle states e^{ikx}; the allowed values of k determined by the periodic boundary conditions are $k_p=2\pi p/L$ with p any integer. We shall assume that n, the total number of particles, is odd; then the ground state is nondegenerate and is obtained by choosing the values of p lying within the

[2] It is an obvious corollary that also all equilibrium thermodynamic properties of the two systems are identical.

[3] For a proof see O. Penrose and L. Onsager, Phys. Rev. **104**, 576 (1956), Sec. 5.

[4] We assume that the ground state is nondegenerate, so that ψ_0^F can be chosen to be real.

[5] (a) *Note added in proof.* This model was treated some time ago in an unpublished work by J. K. Percus and G. J. Yevick (private communication from Professor Yevick); (b) ψ^B is not required to satisfy the Schrödinger equation on the surfaces $x_j=x_l$, where it vanishes and suffers discontinuities in gradient (but not in value) as a result of the infinite repulsive forces.

Fermi "sphere" $-\frac{1}{2}(n-1)\leqslant p\leqslant\frac{1}{2}(n-1)$. By factoring $e^{-i(n-1)\pi x_j/L}$ out of the jth row of the Slater determinant, one can write this ground state in the form

$$\psi_0^F(x_1\cdots x_n)=C\exp[-i(n-1)\pi L^{-1}\textstyle\sum_j x_j]\begin{vmatrix}1 & z_1 & z_1^2\cdots & z_1^{n-1}\\ \vdots & & & \vdots\\ 1 & z_r & z_n^2\cdots & z_n^{n-1}\end{vmatrix}, \quad (8)$$

where

$$z_j=\exp i2\pi L^{-1}x_j \quad (9)$$

and the normalization factor C is

$$C=(n!)^{-\frac{1}{2}}L^{-\frac{1}{2}n}j^{\frac{1}{2}(n-1)}. \quad (10)$$

The phase factor in (10) is chosen to make ψ_0^F real. The determinant in (8) is of a type familiar to mathematicians; its value[6] is just the product of the $\frac{1}{2}n(n-1)$ differences (z_j-z_l), $j>l$. Thus

$$\psi_0^F=C\exp[-i(n-1)\pi L^{-1}\textstyle\sum_j x_j]\times\prod_{j>l}[\exp(i2\pi L^{-1}x_j)-\exp(i2\pi L^{-1}x_l)]. \quad (11)$$

According to (5), the ground state ψ_0^B of the system of impenetrable point bosons is given in terms of the Fermi ground state (11) by

$$\psi_0^B=|\psi_0^F|=(n!)^{-\frac{1}{2}}L^{-\frac{1}{2}n}\times\prod_{j>l}|\exp(i2\pi L^{-1}x_j)-\exp(i2\pi L^{-1}x_l)|$$
$$=(n!)^{-\frac{1}{2}}L^{-\frac{1}{2}n}2^{\frac{1}{2}n(n-1)}\prod_{j>l}|\sin[\pi L^{-1}(x_j-x_l)]|. \quad (12)$$

The structure of this wave function is very simple. If we vary the position of one particle, keeping all the others fixed, then the wave function is positive and smoothly varying everywhere except at the position of each of the other particles, where it vanishes and has a cusp as a result of the singular repulsive interaction. According to our general results in Sec. 2, the ground state energy is the same as that of ψ_0^F, i.e., that of the free Fermi gas

$$E_0=\frac{\hbar^2}{m}\sum_{p=1}^{\frac{1}{2}(n-1)}\left(\frac{2\pi p}{L}\right)^2=\tfrac{1}{6}(n-n^{-1})\frac{(\pi\hbar\rho)^2}{m}, \quad (13)$$

where $\rho=n/L$, the particle number density.[7a] The expressions (12) and (13) are exact for all odd values of $n\geqslant 3$, but are not valid for even n[1,7b]; however, one expects all extensive properties of the system to be independent of whether n is odd or even in the limit $n\to\infty$.

According to Sec. 2, all configurational probability distributions of our system of impenetrable point bosons are the same as those of the free Fermi gas,[8] which are well known. In particular, the pair correlation function,

$$D(x)=L^2\int_0^L\cdots\int_0^L dx_3\cdots dx_n|\psi_0^B(0,x,x_3\cdots x_n)|^2, \quad (14)$$

is given by

$$D(x)=1-\left[\frac{\sin(\pi\rho x)}{n\sin(\pi L^{-1}x)}\right]^2; \quad (15)$$

for particle separation $x\ll L$ one can use the asymptotic expression

$$D(x)\approx 1-\left[\frac{\sin(\pi\rho x)}{\pi\rho x}\right]^2, \quad x\ll L. \quad (16)$$

$D(x)$ rises from zero for zero particle separation to unity for particle separation much larger than ρ^{-1}. What is more interesting is the single-particle momentum distribution, which is quite different from that of the Fermi gas and should show how the Bose-Einstein condensation is affected by the "impenetrable point" interparticle interaction. However, we have not succeeded in calculating the exact momentum distribution; the difficulties one encounters in such an evaluation are similar to those encountered in the evaluation of the configurational integral in the classical statistical mechanics of interacting particles. An approximate expression for the momentum distribution, obtained by considering only the part of ψ_0^B representing excitation of pairs of particles to equal and opposite nonzero momenta, leads to the conclusion that the interaction "smears" the condensation; the number condensed at the origin of k space is not proportional to n, but rather to $n/\ln n$, and a large number of other allowed momentum sites near the origin have an occupation of the same order of magnitude. The derivation and details are given in the Appendix; it is hoped that this approximate result is at least qualitatively correct.

In conclusion, we consider the low-lying excited states of our Bose system. For simplicity, we limit ourselves

[6] A. C. Aitken, *Determinants and Matrices* (Oliver and Boyd, Ltd., Edinburgh, 1951), 7th ed., p. 112.

[7] (a) *Note added in proof.* It is possible, by a simple change of variables, to extend the results for the ground state of the impenetrable point Bose gas so as to obtain the exact solution for the ground state of a gas of hard-sphere bosons with diameter $a>0$ enclosed in a one-dimensional periodic box. The resultant expression for the ground-state energy,

$$E_0=\tfrac{1}{6}(n-n^{-1})m^{-1}((\pi\hbar\rho)/(1-\rho a))^2,$$

differs only by a "surface" term from a result obtained previously for a Boltzmann system of one-dimensional hard spheres enclosed in a box [see R. J. Rubin, J. Chem. Phys. **23**, 1183 (1955) for other references]. This agreement is to be expected, since the ground state of a Boltzmann system is identical with that of the corresponding Bose system. (b) Our previous statement [Bull. Am. Phys. Soc. Ser. II, 5, 8 (1960)] that the ground state is "slightly more complicated" for even n was somewhat overoptimistic; the fact of the matter is that we have not been able to find it at all for the case of even n.

[8] It is amusing that there exists a soluble one-dimensional problem in the *classical* statistical mechanics of interacting particles in which all configurational probability distributions are the same as those of the free Fermi gas, hence the same as those of our interacting Bose system; the interparticle interaction in this classical problem is, of course, different from our point interaction. See C. W. Ufford and E. P. Wigner, Phys. Rev. **61**, 524 (1942).

to the one-phonon states, i.e., the states related by (4) to those states of the free Fermi gas in which only one particle is excited above the Fermi sea. The allowed momenta of these states are[9] $\hbar k_j = 2\pi\hbar j L^{-1}$, where j is any integer except zero. We shall restrict ourselves to those states $\psi_j{}^B$ with $j>0$, since those with $j<0$ can be obtained from the relationship $\psi_{-j}{}^B = (\psi_j{}^B)^*$. It is easily shown with the aid of (4) and the well-known expressions for the eigenstates of the free Fermi gas that

$$\psi_j{}^B = CA(x_1\cdots x_n)\exp[-i(n-1)\pi L^{-1}\textstyle\sum_j x_j] \times \begin{vmatrix} 1 & z_1 & z_1{}^2 & \cdots & z_1{}^{n-2} & z_1{}^{n-1+j} \\ \vdots & & & & \vdots & \\ 1 & z_n & z_n{}^2 & \cdots & z_n{}^{n-2} & z_n{}^{n-1+j} \end{vmatrix}, \quad (17)$$

where C and the z_j are given by (9) and (10). The preceding determinant differs from the determinant in (8) only by the factor[10] h_j defined as the sum of the

$$\binom{n+j-1}{j}$$

products of the z_l taken j at a time and with repetition permitted, i.e.,

$$h_j = \sum_{\substack{s_1\cdots s_j=1 \\ (s_1\leqslant s_2\leqslant\cdots\leqslant s_j)}}^{n} z_{s_1}\cdots z_{s_j}. \quad (18)$$

On taking account of (4) and (8), one finds

$$\psi_j{}^B = \psi_0{}^B h_j. \quad (19)$$

For the smallest possible phonon momentum, namely, $k_1 = 2\pi/L$, this reduces to the Feynman form

$$\psi_1{}^B = \psi_0{}^B h_1 = \psi_0{}^B \textstyle\sum_j e^{ik_1 x_j}, \quad (20)$$

but the excited states of higher momentum have a more complicated structure.

The energy of the state $\psi_k{}^B$ is[11] clearly $E_0+\epsilon_k$, where E_0 is the ground-state energy (13) and

$$\epsilon_k = (\hbar^2/2m)[(k+k_F)^2 - k_F{}^2] = (\hbar^2/m)k(k_F+\tfrac{1}{2}k); \quad (21)$$

$\hbar k_F$ is the Fermi momentum, so $k_F = (n-1)\pi L^{-1}$. Hence, apart from a term of order n^{-1},

$$\epsilon_k = (\hbar^2/m)k(\pi\rho+\tfrac{1}{2}k). \quad (22)$$

This spectrum has a phonon character at low k,

$$\epsilon_k \approx \hbar c k, \quad k \ll 2\pi\rho \quad (23)$$

with

$$c = \pi\hbar\rho/m; \quad (24)$$

the one-dimensional Fermi gas is anomalous in having such a phonon spectrum, rather than an effective-mass type spectrum as is the case in two or three dimensions. The physical interpretation of this phonon spectrum is, however, quite different in the Bose and Fermi cases. The phonon character of the one-dimensional Fermi-gas spectrum results solely from statistics, whereas that of the Bose gas arises from the repulsive interparticle interactions. The interpretation of the low excitations as phonons is verified by the fact that the sound velocity (24) obtained from the low-k behavior of the excitation spectrum agrees with that obtained on thermodynamic grounds from the expression for the compressibility of the ground state. The pressure p when the system is in the ground state is

$$p = -(\partial E_0/\partial L) = (\rho^2/n)(\partial E_0/\partial\rho). \quad (25)$$

On substituting from (13), one finds

$$p = \pi^2\hbar^2\rho^3/3m \quad (26)$$

aside from a term of order n^{-1}. Then, since $m\rho$ is the mass density, the sound velocity c is given by

$$c = (m^{-1}\partial p/\partial\rho)^{\frac{1}{2}} = \pi\hbar\rho/m, \quad (27)$$

which agrees with (24).

4. DISCUSSION

It has been shown that for one-dimensional systems of interacting particles for which the interaction has an impenetrable core, there exists a very simple and hitherto unsuspected one-one correspondence between the energy eigenfunctions satisfying Bose-Einstein statistics and those satisfying Fermi-Dirac statistics. The form of this correspondence is such that not only are the energy spectra of the Bose and Fermi systems identical, but also all configurational probability distributions; the salient differences between the Bose and Fermi systems are revealed in their qualitatively different momentum distributions. The one-one Bose-Fermi correspondence was used to obtain a rigorous solution of the many-boson problem of impenetrable point particles in a one-dimensional periodic box; aside from its intrinsic interest in view of the rarity of exact solutions of many-body problems, this solution may also be useful as a testing ground for approximation methods in many-body theory.

APPENDIX: PAIR APPROXIMATION TO THE MOMENTUM DISTRIBUTION

We are interested in calculating the single-particle momentum distribution function n_k of the ground state $\psi_0{}^B$ [Eq. (12)] of the system of impenetrable point

[9] It follows from (4) that, for the case of periodic boundary conditions, the momentum of the state ψ^B is the same as that of ψ^F. Since $P \equiv \sum_j (\hbar/i)\partial/\partial x_j$ is the total linear momentum operator, one has

$$P\psi^B = (P\psi^F)A + \psi^F PA.$$

But it follows from the definition of A [Eq. (3) ff.] that the function PA vanishes except on the planes $x_j = x_l$, where ψ^F vanishes. Hence if $P\psi^F = \hbar k\psi^F$, then

$$P\psi^B = (P\psi^F)A = \hbar k\psi^F A = \hbar k\psi^B,$$

Q.E.D.

[10] A. C. Aitken, footnote reference 6, p. 116, prob. 2.

[11] Here we are labeling the state by its phonon wave number k rather than by the integer j, which is related to k by $k = 2\pi j/L$.

bosons. n_k is defined as the mean number of particles with momentum $\hbar k$ and is therefore given by

$$n_k = n \sum_{k_2 \cdots k_n} |\phi_0{}^B(k k_2 \cdots k_n)|^2, \tag{A1}$$

where $\phi_0{}^B$ is the momentum wave function and $k_2 \cdots k_n$ are summed over all wave numbers consistent with the periodic boundary conditions. $\phi_0{}^B$ is the Fourier transform of $\psi_0{}^B$:

$$\phi_0{}^B(k_1 \cdots k_n) = L^{-\frac{1}{2}n} \int_0^L \cdots \int_0^L dx_1 \cdots dx_n \psi_0{}^B(x_1 \cdots x_n) \times e^{-ik_1 x_1} \cdots e^{-ik_n x_n}. \tag{A2}$$

All attempts at a direct evaluation of n_k with the aid of (12), (A1), and (A2) have failed. Using (4) and the faltung theorem for Fourier transforms, one can relate $\phi_0{}^B$ to the momentum wave function of the free Fermi gas and the Fourier transform of $A(x_1 \cdots x_n)$, but this approach does not appear to be fruitful.

Since it is well known that many calculations in many-body theory are more simply and elegantly performed if one works in terms of a quantized-field representation rather than the Schrödinger representation, it is suggested that we employ such a representation in calculating the momentum distribution function n_k. If a_k and $a_k{}^\dagger$ are, respectively, the annihilation and creation operators for Bose particles in the single-particle momentum eigenstate $L^{-\frac{1}{2}} e^{ikx}$, then $N_k = a_k{}^\dagger a_k$ is the occupation-number operator for this single-particle state, and n_k is just the expectation value

$$n_k = \langle \psi_0{}^B | N_k | \psi_0{}^B \rangle, \tag{A3}$$

where $|\psi_0{}^B\rangle$ is the state vector whose Schrödinger wave function is $\psi_0{}^B(x_1 \cdots x_n)$.[12] In order to obtain an expression for the ground state vector $|\psi_0{}^B\rangle$ in terms of some operator (a function of the a_k and $a_k{}^\dagger$) acting on the unperturbed ground state vector,[13] we first rewrite $\psi_0{}^B$ [Eq. (12)] in an exponential form:

$$\psi_0{}^B(x_1 \cdots x_n) = C' \prod_{j>l} |\sin[\pi L^{-1}(x_j - x_l)]| = C' \exp\{\tfrac{1}{2} \sum_{jl}{}' \ln |\sin[\pi L^{-1}(x_j - x_l)]|\}, \tag{A4}$$

where C' is a normalization constant and $\sum_{jl}{}'$ is a sum with $j \neq l$, rather than with $j > l$. On introducing the Fourier decomposition

$$\tfrac{1}{2} \ln |\sin[\pi L^{-1}(x_j - x_l)]| = -\sum_k \lambda_k e^{ik(x_j - x_l)}, \tag{A5}$$

where

$$\lambda_k = -\tfrac{1}{2} L^{-1} \int_0^L \ln \sin(\pi L^{-1} x) e^{-ikx} dx, \tag{A6}$$

one finds

$$\psi_0{}^B(x_1 \cdots x_n) = C' \exp[-\sum_{jl}{}' \sum_k \lambda_k e^{ik(x_j - x_l)}]. \tag{A7}$$

In the Schrödinger representation, the particle density operator $\rho(x)$ is given by

$$\rho(x) = \sum_j \delta(x - x_j), \tag{A8}$$

where δ is a periodic Dirac delta function, i.e., its arguments are to be interpreted modulo L. The Fourier components of the density are

$$\rho_k = \int_0^L \rho(x) e^{-ikx} dx = \sum_j e^{-ikx_j}. \tag{A9}$$

Equation (A7) can be rewritten in terms of the ρ_k:

$$\psi_0{}^B(x_1 \cdots x_n) = C' e^{n \sum_k \lambda_k} e^{-\sum_k \lambda_k \rho_k \rho_{-k}}. \tag{A10}$$

To transform (A10) to the quantized-field representation, we first rewrite it in the form

$$\psi_0{}^B(x_1 \cdots x_n) = C'' e^{-\sum_k \lambda_k \rho_k \rho_{-k}} L^{-\frac{1}{2}n}, \tag{A11}$$

where $C'' = C' \exp(n \sum_k \lambda_k) L^{\frac{1}{2}n}$ and the ρ_k are to be interpreted as operators which reduce merely to multiplicative functions in the Schrödinger representation. The factor $L^{-\frac{1}{2}n}$ inserted on the right side of (A11) is the Schrödinger wave function of the unperturbed system, i.e., the system in which the bosons are completely free so that the subsidiary condition (6) is not imposed. Upon transforming to the quantized-field representation, one therefore finds

$$|\psi_0{}^B\rangle = C'' e^{-\sum_k \lambda_k \rho_k \rho_{-k}} |n\rangle, \tag{A12}$$

where $|n\rangle$ is the state vector representing the unperturbed ground state

$$|n\rangle = (n!)^{-\frac{1}{2}} (a_0{}^\dagger)^n |0\rangle \tag{A13}$$

with $|0\rangle$ the vacuum state. In the quantized-field repre-

[12] Those familiar with Siegert's work on field operators for bosons with impenetrable cores [A. J. F. Siegert, Phys. Rev. 116, 1057 (1959)] may raise the objection that the free-particle annihilation and creation operators a_k and $a_k{}^\dagger$ cannot be used in treating particles with hard cores. Such an objection is based on a misinterpretation of the significance of Eq. (A4) of Siegert's paper:

$$\psi(x)\psi(x') = \psi^\dagger(x)\psi^\dagger(x') = 0, \quad |x - x'| \leqslant a,$$

where $\psi(x)$ and $\psi^\dagger(x)$, respectively, annihilate and create a boson at point x. If this equation were a *necessary* property of the Bose field operators for particles with hard cores of diameter a, then the objection would be justified. However, Siegert's Eq. (A4) is a *sufficient*, but not a necessary, condition for the vanishing of the Schrödinger wave function when hard cores overlap. All that is necessary is the much weaker condition

$$\psi(x)\psi(x')|\rangle = 0, \quad |x - x'| \leqslant a,$$

where $|\rangle$ is any state vector describing particles with impenetrable cores of diameter a (for the case discussed in this Appendix, $a = 0$). It is more convenient for our purposes to retain field operators $\psi(x)$, $\psi^\dagger(x)$ satisfying the usual Bose commutation rules, and hence to interpret the above equation as a subsidiary condition on allowable state vectors $|\rangle$; this subsidiary condition is merely a transcription of (1) into the language of quantized fields.

[13] I am indebted to Professor E. P. Gross for suggesting the following method of finding the ground state vector $|\psi_0{}^B\rangle$.

sentation, the operators ρ_k are given by

$$\rho_k=\sum_{k'k''} a_{k'}{}^\dagger a_{k''}\int_0^L (L^{-\frac{1}{2}}e^{-ik'x})e^{-ikx}(L^{-\frac{1}{2}}e^{ik''x})dx$$

$$=\sum_{k'} a_{k'}{}^\dagger a_{k'+k}. \tag{A14}$$

Equations (A3) and (A12)–(A14) furnish an expression for n_k which is alternative to (A1), but, unfortunately, no more amenable to exact evaluation; the difficulty can be traced to the fact that the argument of the exponential operator in (A12) is quartic in the a_k and $a_k{}^\dagger$, whereas one only knows how to evaluate expectation values in states of the form $e^S|n\rangle$ if S is quadratic (corresponding to a *linear* canonical transformation of the annihilation and creation operators). We are therefore forced to an approximate evaluation of (A3) based on an approximation which renders S quadratic.

In his pioneering work on the many-boson problem,[14] Bogolubov linearized the Heisenberg equations of motion by treating a_0 and $a_0{}^\dagger$ as c numbers. The basic physical reasoning involved is that the expectation value of $N_0=a_0{}^\dagger a_0$ will be very large if any tendency to Bose-Einstein condensation survives the effects of the interparticle interaction; on the other hand, $[a_0,a_0{}^\dagger]=1$. Hence the commutator of a_0 and $a_0{}^\dagger$ is much smaller than their product, so that their replacement by c numbers seems heuristically justified. Bogolubov made the additional approximation of dropping those cubic and quartic terms that remained in the Hamiltonian after replacement of a_0 and $a_0{}^\dagger$ by c numbers. We shall employ similar approximations in evaluating (A3).

It follows from (A14) that

$$\rho_0=\sum_k a_k{}^\dagger a_k\equiv\sum_k N_k\equiv N. \tag{A15}$$

But the ρ_k all commute (as can be proved from the Bose commutation relations for the a_k, $a_k{}^\dagger$) and $|n\rangle$ is an eigenstate of N belonging to eigenvalue n (the total number of particles). Upon separating out the term with $k=0$ in (A12), one accordingly finds

$$|\psi_0{}^B\rangle=C'''e^S|n\rangle, \tag{A16}$$

where C''' is a normalization constant and

$$S=-\sum_{k\neq 0}\sum_{k'k''}\lambda_k a_{k'+k}{}^\dagger a_{k''-k}{}^\dagger a_{k'}a_{k''}; \tag{A17}$$

in obtaining (A16) and (A17) use has also been made of the Bose commutation relations in order to throw the product of annihilation and creation operators into normal order $a^\dagger a^\dagger aa$. Replacing a_0 and $a_0{}^\dagger$ in (A17) by the c number $n_0{}^{\frac{1}{2}}$ and dropping terms cubic and quartic in the a_k, $a_k{}^\dagger$ with $k\neq 0$, one obtains the approximate expression

$$S_p=-n_0\sum_{k\neq 0}\lambda_k(a_ka_{-k}+a_k{}^\dagger a_{-k}{}^\dagger+2N_k), \tag{A18}$$

where the subscript "p" implies "pair approximation." The c number n_0 is to be interpreted as the mean number of particles with momentum zero. To be consistent, we must also make the replacement $a_0{}^\dagger\rightarrow n_0{}^{\frac{1}{2}}$ in the unperturbed state $|n\rangle$ [Eq. (A13)]; the pair approximation to the ground state vector (A16) is therefore

$$|\psi_0{}^B{}_p\rangle=De^{S_p}|0\rangle, \tag{A19}$$

where D is a normalization constant and $|0\rangle$ is the vacuum state. Because of the replacements $a_0\rightarrow n_0{}^{\frac{1}{2}}$, $a_0{}^\dagger\rightarrow n_0{}^{\frac{1}{2}}$, the total number of particles is no longer conserved [$|\psi_0{}^B{}_p\rangle$ is not an eigenstate of N, Eq. (A15)], but we can impose the requirement that the *mean* total number of particles be n:

$$n_0+\sum_{k\neq 0}\langle\psi_0{}^B{}_p|N_k|\psi_0{}^B{}_p\rangle=n. \tag{A20}$$

The state (A19) is not yet in a form in which expectation values can be readily evaluated, since the operator e^{S_p} is not unitary; indeed, it is hermitian. We shall therefore find an equivalent unitary operator by investigating the relationship between $a_k|\psi_0{}^B{}_p\rangle$ and $a_{-k}{}^\dagger|\psi_0{}^B{}_p\rangle$ (it is obvious from the pair structure of $|\psi_0{}^B{}_p\rangle$ that these two states are closely related). Since S_p is hermitian, the operator e^{-S_p} is well-defined, and so

$$\begin{aligned} a_ke^{S_p}|0\rangle&=e^{S_p}(e^{-S_p}a_ke^{S_p})|0\rangle,\\ a_{-k}{}^\dagger e^{S_p}|0\rangle&=e^{S_p}(e^{-S_p}a_{-k}{}^\dagger e^{S_p})|0\rangle. \end{aligned} \tag{A21}$$

We define the similarity (not unitary) transforms[15]

$$a_k(\epsilon)\equiv e^{-\epsilon S_p}a_ke^{\epsilon S_p},\quad a_{-k}{}^\dagger(\epsilon)\equiv e^{-\epsilon S_p}a_{-k}{}^\dagger e^{\epsilon S_p}. \tag{A22}$$

These can be evaluated by a differential-equation technique. The "Heisenberg equations of motion" are

$$\begin{aligned} da_k(\epsilon)/d\epsilon&=[a_k(\epsilon),S_p]=-2n_0\lambda_k[a_k(\epsilon)+a_{-k}{}^\dagger(\epsilon)],\\ da_{-k}{}^\dagger(\epsilon)/d\epsilon&=[a_{-k}{}^\dagger(\epsilon),S_p]\\ &=2n_0\lambda_k[a_k(\epsilon)+a_{-k}{}^\dagger(\epsilon)]. \end{aligned} \tag{A23}$$

Adding these two equations, one finds that the derivative of $[a_k(\epsilon)+a_{-k}{}^\dagger(\epsilon)]$ vanishes, and hence that

$$a_k(\epsilon)+a_{-k}{}^\dagger(\epsilon)=a_k+a_{-k}{}^\dagger \tag{A24}$$

since $a_k(0)=a_k$ and $a_{-k}{}^\dagger(0)=a_{-k}{}^\dagger$. Differentiating the first Eq. (A23) once more and using (A24), one finds

$$d^2a_k(\epsilon)/d\epsilon^2=0, \tag{A25}$$

and hence

$$a_k(\epsilon)=A_k+B_k\epsilon. \tag{A26}$$

The coefficients A_k and B_k can be evaluated from the initial conditions

$$\begin{aligned} a_k(0)&=A_k=a_k,\\ (da_k/d\epsilon)_{\epsilon=0}&=B_k=-2n_0\lambda_k(a_k+a_{-k}{}^\dagger); \end{aligned} \tag{A27}$$

the second Eq. (A23) has been used in obtaining the second Eq. (A27). It follows from (A22), (A26), and

[14] N. N. Bogolubov, J. Phys. (U.S.S.R.) **11**, 23 (1947).

[15] Note that $a_{-k}{}^\dagger(\epsilon)\neq[a_{-k}(\epsilon)]^\dagger$.

(A27) that

$$e^{-S_p}a_k e^{S_p}=a_k(1)=(1-2n_0\lambda_k)a_k-2n_0\lambda_k a_{-k}^\dagger, \quad \text{(A28)}$$

and then from (A24) that

$$e^{-S_p}a_{-k}^\dagger e^{S_p}=a_{-k}^\dagger(1)=2n_0\lambda_k a_k+(1+2n_0\lambda_k)a_{-k}^\dagger. \quad \text{(A29)}$$

By (A21), (A28), and (A29) one then finds, using the fact that $a_k|0\rangle=0$,

$$\begin{aligned} a_k e^{S_p}|0\rangle&=-2n_0\lambda_k e^{S_p}a_{-k}^\dagger|0\rangle,\\ a_{-k}^\dagger e^{S_p}|0\rangle&=(1+2n_0\lambda_k)e^{S_p}a_{-k}^\dagger|0\rangle. \end{aligned} \quad \text{(A30)}$$

Let us define

$$\xi_k\equiv(1-\phi_k^2)^{-\frac{1}{2}}(a_k+\phi_k a_{-k}^\dagger), \quad \text{(A31)}$$

where

$$\phi_k\equiv 2n_0\lambda_k/(1+2n_0\lambda_k). \quad \text{(A32)}$$

Then it follows from (A19) and (A30) that

$$\xi_k|\psi_0{}^B{}_p\rangle=0. \quad \text{(A33)}$$

Furthermore, it follows from (A31) and the Bose commutation relations for the a_k and $a_k^\dagger$ that

$$[\xi_k,\xi_{k'}^\dagger]=\delta_{kk'}, \quad [\xi_k,\xi_{k'}]=0, \quad \text{(A34)}$$

i.e., that the operators ξ_k and $\xi_k^\dagger$ are also Bose annihilation and creation operators for certain "elementary excitations" which can be thought of as phonons; (A33) states that the ground state is the state of no phonons. The canonical transformation (A31) is of the Bogolubov[14] type, although the function ϕ_k is of quite a different form from his.

We can now use (A31) and (A33) to find a *unitary* operator U such that

$$|\psi_0{}^B{}_p\rangle=U|0\rangle. \quad \text{(A35)}$$

Indeed, since the transformation (A31) is canonical, one can define a unitary operator U by the requirement

$$\xi_k=Ua_kU^{-1}=(1-\phi_k^2)^{-\frac{1}{2}}(a_k+\phi_k a_{-k}^\dagger). \quad \text{(A36)}$$

Then by (A35) and (A36),

$$\xi_k|\psi_0{}^B{}_p\rangle=(Ua_kU^{-1})U|0\rangle=Ua_k|0\rangle=0, \quad \text{(A37)}$$

so that (A33) is satisfied. Since $|\psi_0{}^B{}_p\rangle$ is uniquely determined up to normalization by the requirement (A33) for all $k(\neq 0)$, and the state (A35) is normalized, we conclude that the expression (A35), with U determined by (A36), is indeed correct. The explicit form of U is not difficult to determine[16]; it is given by

$$U=\exp[\tfrac{1}{2}\textstyle\sum_{k\neq 0}(a_k a_{-k}-a_k^\dagger a_{-k}^\dagger)\tanh^{-1}\phi_k]. \quad \text{(A38)}$$

It is now straightforward to evaluate the approximate momentum distribution function

$$\begin{aligned} n_k&\approx\langle\psi_0{}^B{}_p|N_k|\psi_0{}^B{}_p\rangle\\ &\approx\langle 0|U^{-1}N_kU|0\rangle, \quad k\neq 0. \end{aligned} \quad \text{(A39)}$$

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

One sees from (A38) that U^{-1} can be obtained from U by replacing ϕ_k by $-\phi_k$; hence by (A36)

$$U^{-1}a_kU=(1-\phi_k^2)^{-\frac{1}{2}}(a_k-\phi_k a_{-k}^\dagger). \quad \text{(A40)}$$

Thus

$$\begin{aligned} U^{-1}N_kU&=(U^{-1}a_k^\dagger U)(U^{-1}a_kU)\\ &=(1-\phi_k^2)^{-1}(\phi_k^2+N_k+\phi_k^2N_{-k}\\ &\qquad-\phi_k a_k a_{-k}-\phi_k a_k^\dagger a_{-k}^\dagger), \end{aligned} \quad \text{(A41)}$$

and by (A39) and (A32)

$$n_k\approx\phi_k^2/(1-\phi_k^2)=4n_0^2\lambda_k^2/(1+2n_0\lambda_k), \quad k\neq 0. \quad \text{(A42)}$$

The integrals (A6) for the λ_k are readily evaluated by contour integration when one takes into account the fact that k is an integral multiple of $2\pi/L$; the result is

$$\lambda_0=\tfrac{1}{2}\ln 2; \quad \lambda_k=\pi/2L|k|, \quad k\neq 0. \quad \text{(A43)}$$

Thus

$$n_k\approx\pi^2n_0^2/L|k|(L|k|+\pi n_0), \quad k\neq 0. \quad \text{(A44)}$$

To determine n_0, the number of particles condensed at the origin of momentum space, we use (A20):

$$n_0+\textstyle\sum_{k\neq 0}n_k=n. \quad \text{(A45)}$$

Since we are only interested in results valid asymptotically as $n\to\infty$, we can replace the summation by an integration

$$\textstyle\sum_{k\neq 0}\to\frac{L}{2\pi}\left(\int_{2\pi/L}^{\infty}dk+\int_{-\infty}^{-2\pi/L}dk\right), \quad \text{(A46)}$$

the interval $(-2\pi/L, 2\pi/L)$ being excluded because the smallest allowed value of $|k|\neq 0$ is $2\pi/L$. Thus, since $n_{-k}=n_k$, (A45) becomes

$$n_0+\frac{L}{\pi}\int_{2\pi/L}^{\infty}n_k dk\approx n. \quad \text{(A47)}$$

On inserting (A44) and performing the resultant elementray integration, one finds

$$n_0[1+\ln(1+\tfrac{1}{2}n_0)]\approx n. \quad \text{(A48)}$$

For large n_0, this reduces to

$$n_0\approx n/\ln n_0. \quad \text{(A49)}$$

This transcendental equation can be solved for n_0 by iteration. The first approximation, obtained by replacing n_0 by n on the right-hand side of (A49), is

$$n_0\approx n/\ln n. \quad \text{(A50)}$$

A better approximation, obtained by inserting (A50) in the right side of (A49), is

$$n_0\approx n/(\ln n-\ln\ln n). \quad \text{(A51)}$$

Equation (A51) shows that (A50) is adequate for very large n; thus (A44) becomes

$$n_k\approx(\pi\rho/\ln n)^2[|k|(|k|+\pi\rho/\ln n)]^{-1}, \quad k\neq 0 \quad \text{(A52)}$$

where $\rho=n/L$.

The expressions (A50) and (A51) show that the interparticle interaction "smears" the Bose-Einstein condensation[17]: the number of particles condensed at $k=0$ is not proportional to n, but rather to $n/\ln n$, and a large number of other allowed momentum sites near the origin have occupations of the same order of magnitude, since for small k (A52) reduces to

$$n_k \approx n/(2|j|\ln n), \quad 0<|k|\equiv(2\pi|j|/L)\ll\pi\rho/\ln n, \tag{A53}$$

where $j=\pm1, \pm2, \cdots$. In spite of this "smearing" effect of the interaction, the condensation is still complete in a certain generalized sense. We can define the condensed fraction f as the fraction of the total number of particles having momenta which are infinitesimal compared to any macroscopic momentum:

$$f\equiv\lim_{\epsilon\to0}\lim_{n\to\infty}(n^{-1}\sum_{|k|<\epsilon\rho}n_k); \tag{A54}$$

in cases where the condensation takes place only into $k=0$, this reduces to the usual definition

$$f=\lim_{n\to\infty}(n_0/n).$$

Then by (A52)

$$f=(\pi\rho)^2\lim_{\epsilon\to0}\lim_{n\to\infty}\left[\frac{1}{n(\ln n)^2}\sum_{|k|<\epsilon\rho}\frac{1}{|k|(|k|+\pi\rho/\ln n)}\right]$$

$$=\pi\rho\lim_{\epsilon\to0}\lim_{n\to\infty}\left[\frac{1}{(\ln n)^2}\int_{2\pi/L}^{\epsilon\rho}\frac{dk}{k(k+\pi\rho/\ln n)}\right]$$

$$=\lim_{\epsilon\to0}\lim_{n\to\infty}\left\{\frac{1}{\ln n}\left[\ln\left(\frac{\epsilon n}{2\pi}\right)-\ln\left(\frac{\epsilon n}{2\pi}+\frac{n}{2\ln n}\right)\right.\right.$$

$$\left.\left.+\ln\left(1+\frac{n}{2\ln n}\right)\right]\right\}=\lim_{\epsilon\to0}1=1. \tag{A55}$$

Thus the Bose-Einstein condensation is *complete* in the generalized sense (A54). This is because n_k [Eq. (A52)] falls off quite rapidly with increasing k, being $\sim(\pi/\epsilon\ln n)^2$ for $|k|\sim\epsilon\rho$; n_k falls essentially to zero in a distance of order $\rho/\ln n$. This behavior is in marked contrast to that of the momentum distribution function of the free Fermi gas, which is equal to unity for $|k|<\pi\rho$ and vanishes for $|k|>\pi\rho$.

[17] The fact that n_0 is not proportional to n is connected with the fact that the ground-state wave function (12) possesses long-range order in view of the very slow rate of change of the factors $\sin[\pi L^{-1}(x_j-x_l)]$; see O. Penrose and L. Onsager, footnote reference 3. This long-range order does not show up in the pair correlation function (15), or indeed those of any finite order; one has to go to the many-body correlation functions to see it.

Generalization of the "Edge-of-the-Wedge" Theorem

H. EPSTEIN*
Palmer Physical Laboratory, Princeton University, Princeton, New Jersey
(Received May 31, 1960)

It is proved that two analytic functions of several complex variables, having the same boundary values when the imaginary parts of the variables tend to zero inside two arbitrary, but fixed, open cones, possess a common analytic continuation in a certain open set. This is a generalization of the "edge-of-the-wedge" theorem, a proof of which is obtained in passing.

IN the study of the Wightman functions and Green functions, or in dispersion theory, the problem frequently arises of deciding whether certain conditions imposed on the boundary values of various analytic functions are equivalent to requirements involving only domains of analyticity. A very useful tool for dealing with such situations is the "edge-of-the-wedge" theorem[1] which may be stated as follows: Let Ω be an open set in C^n, containing a nonempty real environment D.[2] Let $f_1(z)$ and $f_2(z)$ be two functions of $z=(z_1,\cdots,z_n)$ analytic in $(D+iC_1)\bigcap\Omega$ and $(D+iC_2)\bigcap\Omega$, respectively, where C_1 and C_2 are two real open cones in R^n. Furthermore, assume that f_1 and f_2 have equal boundary values (in the sense of distributions) at the real points of D. Then the "edge-of-the-wedge" theorem asserts that, if $C_1=-C_2$, or, equivalently, if $C_1\bigcap(-C_2)$ is a nonempty open cone in R^n, f_1 and f_2 possess a common analytic continuation in some complex neighborhood of D.

It is natural to ask what happens if we abandon the hypothesis that $C_1\bigcap(-C_2)\neq\varnothing$.

Typical examples of this situation are provided by the requirements imposed by local commutativity on the Wightman functions, or the "two-term identities" satisfied by the various boundary values of the p-space analytic function (Green function).[3] The answer to this question is suggested by the convex-tube theorem, and by the edge-of-the-wedge theorem itself. The purpose of this paper is to prove that if C_1 and C_2 are two arbitrary real (nonempty) open cones, f_1 and f_2 possess a common analytic continuation in an open set of the form

$$\{x+iy: x\in D,\ y\in \text{convex closure of } \Sigma_x\bigcap(C_1\bigcup C_2)\},$$

where Σ_x is a nonempty open sphere in R^n defined by: $0\leqslant y_1^2+\cdots+y_n^2<\rho_x$.

The proof of this theorem will be first carried out in the case when the common boundary value is a continuous function. The main tool is a straightforward generalization of the continuity theorem as it was applied by Jost to the proof of the convex tube theorem (the latter proof is reproduced in D. Ruelle's thesis).[3] The extension to distribution-boundary values is based on the method of regularization (used by Gårding and Beurling in the case of the edge-of-the-wedge theorem).[3]

I. SPECIAL CASE FOR TWO COMPLEX VARIABLES

Let C_1 and C_2 be the two real open cones in R^n defined by

$$C_1=\{(y_1,y_2): 0<t_0y_1<y_2<(t_0+\tau)y_1\},$$
$$C_2=\{(y_1,y_2): 0<-t_0y_1<y_2<-(t_0+\tau)y_1\},$$

where $0<t_0$, $0<\tau$. Let $\bar{C}_1$ and $\bar{C}_2$ be the closures of C_1 and C_2.

Let $f_1(z_1,z_2)$ and $f_2(z_1,z_2)$ be two functions of

$$z=(z_1,z_2)=(x_1+iy_1,\ x_2+iy_2),$$

possessing the following properties:

(1) f_α $(\alpha=1,2)$ is defined and continuous in the region

$$\bar{R}_\alpha(\rho)=\{z: |z_1|\leqslant\rho\sqrt{2},\ |z_2|\leqslant\rho\sqrt{2},\ (y_1,y_2)\in\bar{C}_\alpha\}.$$

(2) f_α is analytic and has bounded first derivatives in the region $R_\alpha(\rho)$

$$R_\alpha(\rho)=\{z: |z_1|<\rho\sqrt{2},\ |z_2|<\rho\sqrt{2},\ (y_1,y_2)\in C_\alpha\}.$$

(3) The boundary values of f_1, f_2 for $y_1=y_2=0$ and $|z_1|\leqslant\rho\sqrt{2}$, $|z_2|\leqslant\rho\sqrt{2}$, are the same.

Then there exists a function $f(z)$ analytic in the region $R(\rho)$:

$$R(\rho)=\{(z_1,z_2): |x_1|<\rho/4,\ |x_2|<\rho/4,\ |y_1|<\rho/4,\ |y_2|<\rho/4, t_0|y_1|<y_2<(t_0+\tau)\rho/4\},$$

* Visiting Fellow of the National Academy of Sciences of the United States, on leave from C.N.R.S., France.

[1] The "edge-of-the-wedge" theorem was discovered by H. Bremmermann, R. Oehme, and J. G. Taylor [Phys. Rev. **109**, 2178(1958)]. See also J. G. Taylor, Ann. Phys. **5**, No. 4, 391 (1958). A new and elegant method for the proof of this theorem was given by F. J. Dyson [Phys. Rev. **110**, 579 (1958)]. This method was perfected and made valid for distribution-boundary values by L. Gårding and A. Beurling (to be published). In the case of functions of one complex variable, continuous on the boundary, the theorem was proved by P. Painlevé [Ann. Fac. Toulouse, **2**, 26 (1888)].

[2] By a real environment, we mean any set of the form $\{z=x+iy: y=0 \text{ and } x\in\Omega$, where Ω is an open set in R^n. The usual notations R^n and C^n are used in this paper to denote the n-dimensional vector spaces on the real and on the complex numbers, respectively.

[3] For general considerations on Wightman functions, Green functions, etc., see, in particular A. S. Wightman, contribution to the *Colloque sur les Problèmes Mathématiques de la Théorie Quantique des Champs* (Lille, 1957), and also Nuovo cimento Suppl. 14, 192 (1959); O. Steinmann, thesis, Zurich, 1959; D. Ruelle, thesis, Brussels, 1959; and references given there. The various boundary values of the p-space analytic function have been systematically studied by O. Steinmann, D. Ruelle, N. Burgoyne, H. Araki, etc.; see for instance, H. Araki and N. Burgoyne, Nuovo cimento (to be published).

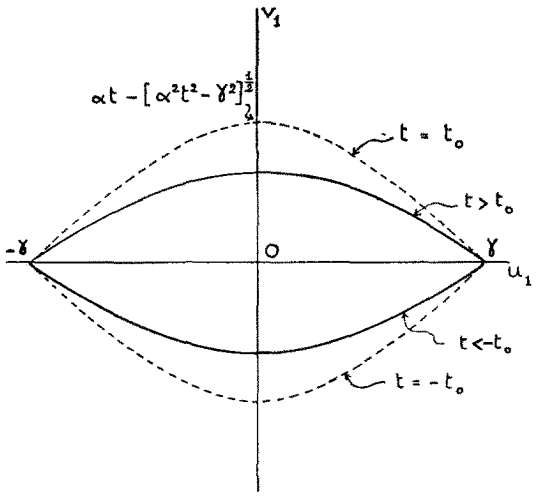

FIG. 1. Points in the w_1 plane satisfying $u_2=\gamma^2$, $y_2=ty_1$.

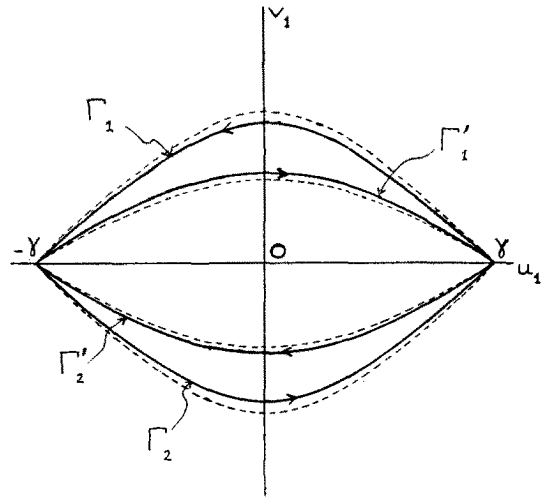

FIG. 2. The contours Γ_1, Γ_2, Γ_1', Γ_2'.

continuous in $\bar{R}(\rho)$, and coinciding with f_1 and f_2 in $\bar{R}_1(\rho)\bigcap\bar{R}(\rho)$ and $\bar{R}_2(\rho)\bigcap\bar{R}(\rho)$, respectively. Furthermore, f satisfies the inequality

$$\max_{z\in R(\rho)} |f(z)| \leqslant \max_{\alpha=1,2} \max_{z'\in\bar{R}_\alpha(\rho)} |f_\alpha(z')|. \qquad (1)$$

Proof

Consider the following analytic change of variables:

$$\begin{cases} w_1=u_1+iv_1=z_1 \\ w_2=u_2+iv_2=-2i\alpha z_2+w_1^2 \end{cases} \qquad \begin{cases} z_1=w_1 \\ z_2=-(i/2\alpha)(w_1^2-w_2) \end{cases}$$

where α is a real, strictly positive constant.

If we hold w_2 fixed, with $u_2=\gamma^2>0$; and look for the points satisfying

$$y_2>0, \quad y_1=sy_2 \quad (s \text{ real}),$$

which, for $s\neq 0$, may be rewritten $y_2=ty_1$, $y_2>0$, we find

$$(v_1-\alpha t)^2-u_1^2-(\alpha^2t^2-\gamma^2)=0.$$

If we require that $|w_1|$ and $|w_2|$ be sufficiently small, and that $\alpha^2t^2-\gamma^2>0$, the corresponding points in the w_1 plane are on an arc of hyperbola ending at the points

$$u_1=\pm\gamma, \quad v_1=0,$$

and passing through the point

$$u_1=0, \quad v_1=\alpha t-[\alpha^2t^2-\gamma^2]^{\frac{1}{2}} \quad \text{for } t>0,$$

or

$$u_1=0, \quad v_1=\alpha t+[\alpha^2t^2-\gamma^2]^{\frac{1}{2}} \quad \text{for } t<0$$

(see Fig. 1).

For $t>t_0$, $0<\gamma<\alpha t_0$, the function $\alpha t-[\alpha^2t^2-\gamma^2]^{\frac{1}{2}}$ is a monotonically decreasing function of t. Therefore, for sufficiently small values of $|w_1|$ and $|w_2|$, with $0<u_2=\gamma^2<\alpha^2t_0^2$, the points satisfying $y_2>t_0|y_1|$ fill the region limited by the two areas of hyperbola corresponding to $t=\pm t_0$. The intersection of the region $(y_1,y_2)\in C_1$ with the analytic manifold $w_2=$constant is represented (in the w_1 plane) by the crescent-shaped area between the arcs corresponding to $t=t_0$ and $t=t_0+\tau$. Similarly the region $y\in C_2$ is represented by the points between the two arcs corresponding to $t=-t_0$ and $t=-(t_0+\tau)$. Let τ' and τ'' be positive real numbers with $0<\tau'<\tau''<\tau$. We denote Γ_1, Γ_1', Γ_2', Γ_2, the arcs corresponding to $t=t_0+\tau'$, $t=t_0+\tau''$, $t=-(t_0+\tau'')$, $t=-(t_0+\tau')$, respectively, with the following orientation: Γ_1' and Γ_2 in the direction of increasing u_1; Γ_1 and Γ_2' in the direction of decreasing u_1. (See Fig. 2.)

The contours Γ_1, Γ_1', Γ_2, Γ_2' depend on $\mathrm{Re}w_2=u_2$. We call $B(u_2)$ [resp. $B'(u_2)$] the open set between Γ_1 and Γ_2 (resp. between Γ_1' and Γ_2'), and $\bar{B}(u_2)$, $\bar{B}'(u_2)$ the closures of these sets.

Clearly $\Gamma_1'(u_2)$ and $\Gamma_2'(u_2)$ can be parametrized in the form

$$\Gamma_1'(u_2): \quad w_1=\zeta(e^{i\theta};u_2), \quad 0\leqslant\theta\leqslant\pi,$$
$$\Gamma_2'(u_2): \quad w_1=\zeta(e^{i\theta};u_2), \quad \pi\leqslant\theta\leqslant 2\pi,$$

where $\zeta(e^{i\theta};u_2)$ has the following properties:

(i) It is a continuous function in $e^{i\theta}$ and w_2, boundedly differentiable in θ for $\theta\neq K\pi$ ($K=0,\pm1,\pm2,\cdots$), and

(ii) For all θ, it has a derivative in u_2 possessing the property (i).

Consider the function defined, *for* w_1 *outside of* $\bar{B}'(u_2)$ by

$$I(w_1,w_2)=\frac{1}{2\pi i}\int_{\Gamma_1'(u_2)}\frac{\tilde{f}_1(\zeta,w_2)}{\zeta-w_1}d\zeta + \frac{1}{2\pi i}\int_{\Gamma_2'(u_2)}\frac{\tilde{f}_2(\zeta,w_2)}{\zeta-w_1}d\zeta,$$

where $\tilde{f}_1(w_1,w_2)=f_1(z_1,z_2)$, $\tilde{f}_2(w_1,w_2)=f_2(z_1,z_2)$. If we consider that, along Γ_1' and Γ_2', ζ is equal to the function

$\zeta(e^{i\theta}; w_2)$ defined earlier, this can be rewritten

$$I(w_1,w_2)=\frac{1}{2\pi i}\int_{\theta=0}^{\theta=\pi}\frac{\tilde f_1(\zeta,w_2)}{\zeta-w_1}d\zeta+\frac{1}{2\pi i}\int_{\theta=\pi}^{\theta=2\pi}\frac{\tilde f_2(\zeta,w_2)d\zeta}{\zeta-w_1}.$$

$I(w_1,w_2)$ is a continuous function of w_1, w_2, when $|w_2|$ is sufficiently small, $w_1\in\bar B'(u_2)$, and $u_2\geqslant 0$, and vanishes for $u_2=0$, because f_1 and f_2 are continuous and, for $u_2\to 0$ the length of the contour of integration tends to zero while the integrand stays bounded.

We shall now prove that, for $|w_2|$ sufficiently small, $u_2>0$, $w_1\notin\bar B'(u_2)$, the function $I(w_1,w_2)$ is analytic. In that region, I is continuous in w_1 and w_2, and obviously analytic in w_1, so that it is sufficient to prove the analyticity in w_2.[4]

Let us denote

$$I_\epsilon(w_1,w_2)=\frac{1}{2\pi i}\int_{\theta=\epsilon}^{\theta=\pi-\epsilon}\frac{\tilde f_1(\zeta,w_2)}{\zeta-w_1}d\zeta+\frac{1}{2\pi i}\int_{\theta=\pi+\epsilon}^{\theta=2\pi-\epsilon}\frac{\tilde f_2(\zeta,w_2)}{\zeta-w_1}d\zeta.$$

Since the contours lie inside the analyticity domain of f_1 and f_2, respectively, $I_\epsilon(w_1,w_2)$ has partial derivatives

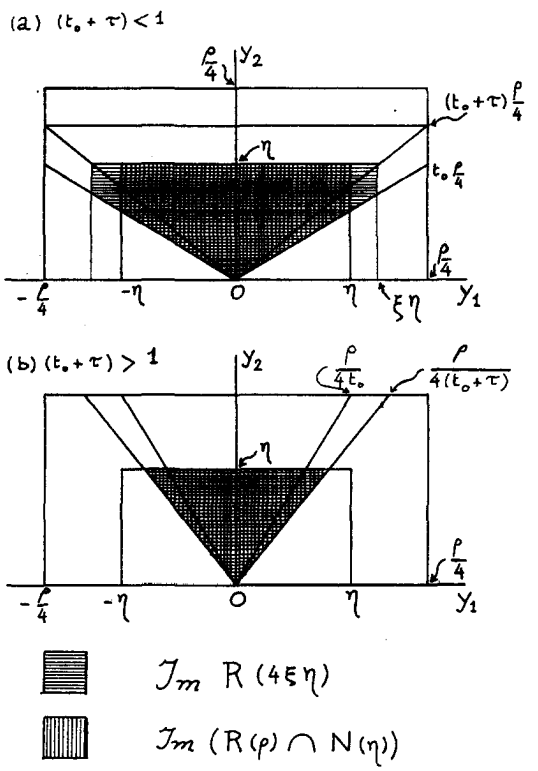

FIG. 3. Imaginary parts of $R(\rho)\cap N(\eta)$ and $R(4\xi\eta)$ for $4\xi\eta<\rho$.

[4] See, for instance, S. Bochner and T. Martin, *Several Complex Variables* (Princeton University Press, Princeton, New Jersey, 1948), p. 32.

in u_2 and v_2, given by

$$\frac{\partial I_\epsilon}{\partial u_2}=\frac{1}{2\pi i}\int_{\theta=\epsilon}^{\theta=\pi-\epsilon}\left[\frac{\partial\psi_1}{\partial u_2}d\zeta+\frac{\partial\psi_1}{\partial\zeta}\frac{\partial\zeta}{\partial u_2}d\zeta+\psi_1 d\left(\frac{\partial\zeta}{\partial u_2}\right)\right]+\frac{1}{2\pi i}\int_{\theta=\pi+\epsilon}^{\theta=2\pi-\epsilon}\left[\frac{\partial\psi_2}{\partial u_2}d\zeta+\frac{\partial\psi_2}{\partial\zeta}\frac{\partial\zeta}{\partial u_2}d\zeta+\psi_2 d\left(\frac{\partial\zeta}{\partial u_2}\right)\right],$$

$$\frac{\partial I_\epsilon}{\partial v_2}=\frac{1}{2\pi i}\int_{\theta=\epsilon}^{\theta=\pi-\epsilon}\left[\frac{\partial\psi_1}{\partial v_2}\right]d\zeta+\frac{1}{2\pi i}\int_{\theta=\pi+\epsilon}^{\theta=2\pi-\epsilon}\frac{\partial\psi_2}{\partial v_2}d\zeta.$$

Here we have put $\psi_\alpha(w_1,\zeta,w_2)=[\tilde f_\alpha(\zeta,w_2)]/(\zeta-w_1)$ $(\alpha=1,2)$, and used the fact that the contours lie inside analyticity regions of ψ_1 and ψ_2. The expressions obtained have limits to which they tend uniformly (and absolutely) when $\epsilon\to 0$ and $|w_2|$ is sufficiently small, $u_2>0$, and $w_1\notin\bar B'(u_2)$. Therefore these limits are the derivatives of $I(w_1,w_2)$ with respect to u_2 and v_2. Moreover, using the analyticity of f_1, f_2, we find

$$\frac{\partial I_\epsilon}{\partial u_2}+i\frac{\partial I_\epsilon}{\partial v_2}=\frac{1}{2\pi i}\int_{\theta=\epsilon}^{\theta=\pi-\epsilon}\left[\frac{\partial\psi_1}{\partial\zeta}\frac{\partial\zeta}{\partial u_2}d\zeta+\psi_1 d\left(\frac{\partial\zeta}{\partial u_2}\right)\right]+\frac{1}{2\pi i}\int_{\theta=\pi+\epsilon}^{\theta=2\pi-\epsilon}\left[\frac{\partial\psi_2}{\partial\zeta}\frac{\partial\zeta}{\partial u_2}d\zeta+\psi_2 d\left(\frac{\partial\zeta}{\partial u_2}\right)\right]$$

$$=\frac{1}{2\pi i}\left[\int_{\theta=\epsilon}^{\theta=\pi-\epsilon}d\left(\psi_1\frac{\partial\zeta}{\partial u_2}\right)+\int_{\theta=\pi+\epsilon}^{\theta=2\pi-\epsilon}d\left(\psi_2\frac{\partial\zeta}{\partial u_2}\right)\right],$$

$$\frac{\partial I_\epsilon}{\partial u_2}+i\frac{\partial I_\epsilon}{\partial v_2}=\frac{1}{2\pi i}\left[\psi_1\frac{\partial\zeta}{\partial u_2}\Big|_{\epsilon=\pi-\epsilon}-\psi_2\frac{\partial\zeta}{\partial u_2}\Big|_{\theta=\pi+\epsilon}-\psi_1\frac{\partial\zeta}{\partial u_2}\Big|_{\theta=\epsilon}+\psi_2\frac{\partial\zeta}{\partial u_2}\Big|_{\theta=2\pi-\epsilon}\right].$$

When $\epsilon\to 0$, this expression tends to zero by virtue of the continuity of the functions involved, and because $f_1(z_1,z_2)$ and $f_2(z_1,z_2)$ are equal at real points. Therefore $I(w_1,w_2)$ is analytic in w_2 and, consequently, in both w_1 and w_2. Since it tends continuously to 0 when $u_2\to 0$ [while w_1 is only restricted by the condition $w_1\notin\bar B'(u_2)$], $I(w_1,w_2)$ vanishes for all values of w_1 and w_2 such that $|w_2|$ is sufficiently small, $u_2>0$, and $w_1\notin\bar B'(u_2)$.[5]

By the same method one can prove that the function

$$J(w_1,w_2)=\frac{1}{2\pi i}\int_{\Gamma_1(u_2)}\frac{\tilde f_1(\zeta',w_2)}{\zeta'-w_1}d\zeta'+\frac{1}{2\pi i}\int_{\Gamma_2(u_2)}\frac{\tilde f_2(\zeta',w_2)}{\zeta'-w_1}d\zeta'$$

[5] If $f(z)$ is a function of $z=(z_1,z_2)=(x_1+iy_1, x_2+iy_2)$, continuous for $|z_1-a_1|<\eta$, $|z_2|<\eta$, $y_2\geqslant 0$, analytic for $y_2>0$ in that region, and vanishes for $y_2=0$, than it vanishes identically. This can be seen by applying Schwarz's reflection principle to $f(z_1,z_2)$ as an analytic function of z_2 for fixed z_1.

is analytic whenever $|w_2|$ is sufficiently small, and w_1 is not on the contour $\Gamma_1 \cup \Gamma_2$ [but, for w_1 inside $B(u_2)$, the argument used to show that I vanishes does not apply to J since, in this case, when we let the contour of integration shrink to a point while keeping w_1 inside $B(u_2)$, the integrand does not stay bounded]. Now, if w_1 is between $\Gamma_1(u_2)$ and $\Gamma_1'(u_2)$, we have, by applying Cauchy's formula,

$$\tilde{f}_1(w_1,w_2)=I(w_1,w_2)+J(w_1,w_2)=J(w_1,w_2),$$

and if w_1 is between $\Gamma_2(u_2)$ and $\Gamma_2'(u_2)$ we have again

$$\tilde{f}_2(w_1,w_2)=I(w_1,w_2)+J(w_1,w_2)=J(w_1,w_2).$$

Thus $J(w_1,w_2)$ provides an analytic continuation of f_1 and f_2 to the points such that $|w_2|$ is sufficiently small and $w_1 \in B(u_2)$.

To complete the proof of our statement, we examine more closely the new points of analyticity just obtained. They include all points such that

(1) $\mathrm{Re} w_2=\gamma^2$, $0<\gamma<\alpha(t_0+\tau)$,

(2) $w_1 \in B(u_2)$, and all the points (w_1',w_2) such that $w_1' \in \bar{B}(u_2)$, satisfy $|z_1|<\rho\sqrt{2}$, $|z_2|<\rho\sqrt{2}$.

By choosing $\alpha=\rho/(t_0+\tau)$ if $t_0+\tau \leqslant 1$, or $\alpha=\rho$ if $t_0+\tau \geqslant 1$, it is not difficult to verify that the new points of analyticity include all those for which

$$|x_1|<\frac{\rho}{4-\epsilon}, \quad |x_2|<\frac{\rho}{4-\epsilon}, \quad |y_1|<\frac{\rho}{4-\epsilon},$$

$$(t_0+\tau)|y_1|<y_2<\min\left\{\frac{\rho}{4-\epsilon}, \frac{\rho(t_0+\tau)}{4-\epsilon}\right\}$$

for some $\epsilon>0$. By adding points of $R_1(\rho)$ and $R_2(\rho)$, one obtains, in particular, the points of $R(\rho)$. The inequality (1) follows immediately from the maximum principle and from the method of construction of f. The continuity of $f(z)=J(w_1,w_2)$ in $\bar{R}(\rho)$ needs be proved only for real points of $\bar{R}(\rho)$ [at other points it is implied by the continuity of f_1 and f_2 in $\bar{R}_1(\rho)$ or $\bar{R}_2(\rho)$ or by the analyticity of f]. For this purpose, consider the intersection of $R(\rho)$ with the open set

$$N(\eta)=\{z: |x_1|<\eta, |z_2|<\eta, |y_1|<\eta, |y_2|<\eta\}.$$

If we denote $\xi=\max[1, (1/t_0+\tau)]$ and choose $\eta<\rho/4\xi$, we have

$$R(\rho) \cap N(\eta) \subset R(4\xi\eta),$$

(see Fig. 3).

If we assume that $\eta<\rho/4\xi$, f_1 and f_2 still satisfy the hypothesis of (I) if we replace ρ by $4\xi\eta$, and so do $f_1-f_1(0)$ and $f_2-f_2(0)$. On applying the inequality (1) we find

$$\max_{z \in R(\rho) \cap N(\eta)} |f(z)-f_1(0)| < \max_{\alpha=1,2} \max_{z' \in \bar{R}_\alpha(4\xi\eta)} |f_\alpha(z)-f_1(0)|.$$

Since $f_1(z)$ and $f_2(z)$ are continuous in $\bar{R}_1(\rho)$ and $\bar{R}_2(\rho)$ and equal to $f_1(0)$ for $z=0$, and since

$$\bar{R}_\alpha(4\xi\eta)=\bar{R}_\alpha(\rho) \cap \{z: |z_1|<4\xi\eta\sqrt{2}, |z_2|<4\xi\eta\sqrt{2}\},$$

this implies that $f(z)$ is continuous and equal to $f_1(0)$ for $z=0$. The continuity of f at other real points of $\bar{R}(\rho)$ can be proved by translating the origin to any such point.

II. GENERALIZATION TO MORE THAN TWO VARIABLES IN A SPECIAL CASE

We shall denote $z=(z_1,\cdots,z_n)$ a vector in C^n, $n>2$, and $x=(x_1,\cdots,x_n)$, $y=(y_1,\cdots,y_n)$ its real and imaginary parts.

Lemma 1

Let C_1 be the real open cone in R^n defined by

$$\begin{cases} 0<t_0 y_1<y_2<(t_0+\tau)y_1 \\ 0 \leqslant |y_k|<a y_2 \quad \text{for } k=3,\cdots,n, \end{cases}$$

where $t_0>0$, $\tau>0$, and $0<a<\frac{1}{2}$.

Let C_2 be the real open cone defined by

$$\begin{cases} 0<-t_0 y_1<y_2<-(t_0+\tau)y_1 \\ 0 \leqslant |y_k|<a y_2 \quad \text{for } k=3,\cdots,n. \end{cases}$$

Let $f_1(z)$ and $f_2(z)$ be two functions satisfying the following conditions:

(1) $f_\alpha(z)$ is defined and continuous in the region

$$\bar{R}_\alpha(\rho)=\{z: |z_j| \leqslant \rho\sqrt{2}\,(j=1,\cdots,n) \text{ and } y \in \bar{C}_\alpha\} \quad (\alpha=1,2).$$

(2) $f_\alpha(z)$ is analytic and has bounded first derivatives in the open region

$$R_\alpha(\rho)=\{z: |z_j|<\rho\sqrt{2}\,(j=1,\cdots,n) \text{ and } y \in C_\alpha\} \quad (\alpha=1,2).$$

(3) The boundary values of f_1 and f_2 for $y=0$, $|z_j|<\rho\sqrt{2}$, are equal.

Then, there exists a function $f(z)$, analytic in the region $R(\rho)$

$$R(\rho)=\Big\{ |x_j|<\frac{\rho}{4}(j=1,\cdots,n); \quad t_0|y_1|<y_2<(t_0+\tau)\frac{\rho}{4}; \quad |y_1|<\frac{\rho}{4}, y_2<\frac{\rho}{4}; \quad |y_k|<a y_2 \quad \text{for } k=3,\cdots,n \Big\},$$

which coincides with f_1 and f_2 in $R_1(\rho)$ and $R_2(\rho)$, respectively. f is continuous at the real points of $\bar{R}(\rho)$ and satisfies the inequality

$$\max_{z \in \bar{R}(\rho)} |f(z)| < \max_{\alpha=1,2} \max_{z' \in \bar{R}_\alpha(\rho)} |f_\alpha(z')|. \qquad (1')$$

This lemma is an intuitively obvious generalization of the result of Sec. I. A straightforward, but tedious proof is given in the Appendix. An alternative approach, kindly suggested to me by Professor H. Rossi, can be briefly sketched as follows: let $y' \in C_1$ and $y'' \in C_2$ be sufficiently close to the origin. By a suitable linear real change of coordinates we can bring the points iy' and iy'' to have coordinates $(iy_1', iy_2', 0 \cdots 0)$ and $(-iy_1', iy_2', 0 \cdots 0)$ and, using the result obtained for two complex variables, find an analytic continuation of f_1 and f_2 in the variables z_1, z_2 along the straight line joining these two points. Moreover it is possible to find a complex open neighborhood N of this line in the variables z_1, z_2, and a polycylinder P in the variables $z_3, \cdots z_n$ such that, for any fixed values of $z_2, \cdots z_n$ in P, there exists a function $\varphi(z_1, z_2, z_3 \cdots z_n)$ continuing f_1 and f_2 analytically in z_1, z_2 in N.

Given the analyticity of $\varphi = f_1$ or $\varphi = f_2$ in the neighborhood of the "ends" of the "cylinder" $N \times P$, it is then possible to deduce the analyticity of φ in $N \times P$ by a slight generalization of a theorem of Rossi.[6]

We remark that $\mathrm{Im}R(\rho)$ is the convex closure of the intersection of $C_1 \bigcup C_2$ with the cube $|y_j| < (\rho/4)$ $(j = 1, \cdots, n)$. Therefore if we denote $\Sigma(r)$ the open sphere defined in R^n by

$$\sum_{j=1}^{n} x_j^2 < r^2,$$

$R(\rho)$ contains in particular the set $S(\rho)$

$$S(\rho) = \left\{ z : x \in \Sigma\left(\frac{\rho}{4}\right),\ y \in \Sigma\left(\frac{\rho}{4}\right), \text{ and } y \in \text{convex closure of } \Sigma\left(\frac{\rho}{4}\right) \bigcap (C_1 \bigcup C_2) \right\}.$$

This permits the following generalization.

Lemma 2

Let C_1 and C_2 be two arbitrary open cones in R^n, $n \geqslant 2$, and $\bar{C}_1$, $\bar{C}_2$ their closures. Let f_1 and f_2 be two functions of $z = (z_1, \cdots, z_n)$ with the following properties.

(1) $f_\alpha(z)$ is defined and continous for

$$\|z\|^2 = \sum_{j=1}^{n} |z_j|^2 \leqslant 2n\rho^2 \quad \text{and} \quad y \in \bar{C}_\alpha \quad (\alpha = 1, 2).$$

(2) $f_\alpha(z)$ is analytic and has bounded first derivatives in the region $R_\alpha(\rho)$ $(\alpha = 1, 2)$:

$$R_\alpha(\rho) = \{ z : \|z\|^2 < 2n\rho^2,\ y \in C_\alpha \}.$$

(3) The values of $f_1(z)$ and $f_2(z)$ for z real and $\|z\|^2 \leqslant 2n\rho^2$ are equal. Then, there exists a function $f(z)$ analytic in the region $S(\rho)$,

$$S(\rho) = \{ z : x \in \Sigma(\rho/4),\ y \in K(\rho) \},$$

and coinciding with f_1 and f_2 in their domains of definition. Here $K(\rho)$ is the convex closure of

$$\Sigma(\rho/4) \bigcap (C_1 \bigcup C_2).$$

Moreover, f satisfies the following inequality:

$$\max_{z \in S} |f(z)| < \max_{\alpha = 1, 2} \max_{z' \in \bar{R}_\alpha(\rho)} |f_\alpha(z')|. \qquad (1'')$$

Remark. C_1 and C_2 may overlap. Each of them may be disconnected.

Proof

This lemma is proved by decomposing C_1 and C_2 into pairs of sufficiently thin open cones, to which the analysis of I and II may be applied after a suitable *real* linear orthogonal change of coordinates has been performed. Such a transformation leaves invariant $\Sigma(\rho/4)$ and the operation of taking the convex closure. It is easy to see that one thus obtains all the points of S except the real points in the case where $0 \in K(\rho)$. At all points z obtained,

$$|f(z)| < \max_{\alpha = 1, 2} \max_{z' \in \bar{R}_\alpha(\rho)} |f_\alpha(z')| \qquad (2)$$

by virtue of 1 and 1'. The necessary and sufficient condition for $K(\rho)$ to contain 0 is that $C_1 \bigcup C_2$ contain simultaneously an open cone Γ and the opposite cone, $-\Gamma$. In that case $\Sigma(\rho/4) \bigcap \Gamma$ is an open set containing an open sphere ω with radius $\lambda\rho$, $(0 < \lambda < \frac{1}{4})$, and $\Sigma(\rho/4) \bigcap (-\Gamma)$ contains $-\omega$, so that $K(\rho)$ contains the open sphere $\Sigma(\lambda\rho)$; we know that f is defined and analytic in the set $S'(\rho)$:

$$S'(\rho) = \{ z : x \in \Sigma(\rho/4),\ y \in K(\rho),\ y \neq 0 \},$$

where it satisfies (2). If $\epsilon < \lambda\rho$ and if $N(\epsilon)$ is the open set: $N(\epsilon) = \{ z : x \in \Sigma(\epsilon),\ y \in \Sigma(\epsilon) \}$, we have

$$N(\epsilon) \bigcap S'(\rho) = \{ z : x \in \Sigma(\epsilon),\ y \in \Sigma(\epsilon),\ y \neq 0 \} \subset S'(\epsilon/\lambda).$$

Now the functions f_1 and f_2 still satisfy the hypotheses of the lemma if we replace ρ by ϵ/λ. As a consequence we have

$$\max_{z \in N(\epsilon) \cap S'(\rho)} |f(z)| \leqslant \max_{\alpha = 1, 2} \max_{z' \in \bar{R}_\alpha(\epsilon/\lambda)} |f_\alpha(z')|.$$

The same theory can be applied to $f_1 - f_1(0)$, $f_2 - f_1(0)$, and $f - f_1(0)$ so that

$$\max_{z \in N(\epsilon) \cup S'(\rho)} |f(z) - f_1(0)| < \max_{\alpha = 1, 2} \max_{z' \in \bar{R}_\alpha(\epsilon/\lambda)} \times |f_\alpha(z') - f_1(0)|,$$

and since

$$\bar{R}_\alpha(\epsilon/\lambda) = \bar{R}_\alpha(\rho) \bigcap \{ \|z\|^2 \leqslant 2n(\epsilon^2/\lambda^2) \}$$

and f_α is continuous in $\bar{R}_\alpha(\rho)$, this shows that f is continuous at the origin. The continuity at the other real

[6] H. Rossi, thesis, MIT, 1959.

points of $S(\rho)$ can be proved by carrying the origin at any of these points. It follows that f is analytic at the real points of $S(\rho)$. This can be seen by applying the Painlevé theorem in each variable when all the others have fixed real values, and by using the fact that a function, continuous in n complex variables, and analytic in each of them, is analytic in all n variables. [For $n>2$ the real points in $S(\rho)$ form a set of dimension $n\leqslant 2n-3$ and the continuity theorem is sufficient to prove the required result.[7]] The inequality (1″) holds at all points of $S(\rho)$ by virtue of (2) when $y\neq 0$, and, for $y=0$, because $z\in\bar{R}_1\bigcap\bar{R}_2$ and $f(z)=f_1(z)=f_2(z)$ at such a real point. This achieves the proof of the lemma which, as has just been seen, contains a special case of the "edge-of-the-wedge" theorem. The latter will be obtained in a more general case in the next section where the proof of the theorem stated in the introduction will be completed.

III. EXTENSION TO THE CASE WHEN THE BOUNDARY VALUE IS A DISTRIBUTION

Notations: $z=(z_1,\cdots,z_n)$ is a vector in C^n, $z=x+iy$, $x=(x_1,\cdots,x_n)$, $y=(y_1,\cdots,y_n)$. We denote

$$\|z\|=[\sum_{j=1}^{n}|z_j|^2]^{\frac{1}{2}},$$

$$\|x\|=[\sum_{j=1}^{n}|x_j|^2]^{\frac{1}{2}},$$

$$\|y\|=[\sum_{j=1}^{n}|y_j|^2]^{\frac{1}{2}}.$$

$\Sigma(r)$ denotes the real open sphere defined in R^n by $\|x\|<r>0$, and $\bar{\Sigma}(r)$ the closure of $\Sigma(r)$.

Theorem

Let C_1 and C_2 be two real open cones in R^n, $\bar{C}_1$ and $\bar{C}_2$ their closures. Let $f_1(z)$ and $f_2(z)$ be two functions satisfying the following conditions.

(1) $f_\alpha(z)$ is defined and analytic in the region R_α:

$$R_\alpha=\{z:\|z\|^2<4n\rho^2,\ y\in C_\alpha\},\quad(\alpha=1,2),$$

and is continuous at the points of $\bar{R}_\alpha$ where $y\neq 0$.

(2) When $\varphi\in\mathfrak{D}(\bar{\Sigma}[2(n)^{\frac{1}{2}}\rho])$, for $\beta=1,2$,

$$\lim_{\substack{y\to 0\\ y\in C_\beta}}\int f_\beta(x+iy)\varphi(x)dx=T(\varphi)$$

where $T\in\mathfrak{D}'(\bar{\Sigma}[2(n)^{\frac{1}{2}}\rho])$ is a distribution independent of β.

[7] See H. Behnke and P. Thullen, *Theorie der Funktioneu Mehrerer Komplexen Veränderlichen* (Verlag Julius Springer, Berlin, Germany, 1933), p. 50.

Then there exists a function $f(z)$ analytic in the region

$$S=\{z:\|x\|<[(n)^{\frac{1}{2}}-1]\rho,\ y\in K\}.$$

Here K is the convex closure of $\Sigma(\rho/4\sqrt{2})\bigcap(C_1\bigcup C_2)$. $\mathfrak{D}(\bar{\Sigma}[2(n)^{\frac{1}{2}}\rho])$ is the space of C^∞ functions with compact support contained in the compact set $\bar{\Sigma}[2(n)^{\frac{1}{2}}\rho]$, as defined by Schwartz in *Théorie des Distributions* (tome 1, p. 64), and $\mathfrak{D}'(\bar{\Sigma}[2(n)^{\frac{1}{2}}\rho])$ is the dual of $\mathfrak{D}(\bar{\Sigma}[2(n)^{\frac{1}{2}}\rho])$.

Proof of the Theorem

Let F_β be the set of distributions $T_\beta{}^y\in\mathfrak{D}'(\bar{\Sigma}[(2n)^{\frac{1}{2}}\rho])$ defined, for $y\in\bar{C}_\beta$, $\|y\|\leqslant(2n)^{\frac{1}{2}}\rho$, $\beta=1$ or 2, by

$$\begin{cases}T_\beta{}^y(\varphi)=\int\varphi(x)f_\beta(x+iy)dx & \text{if } y\neq 0\\ T_\beta{}^0(\varphi)=T(\varphi).\end{cases}$$

Here $\varphi\in\mathfrak{D}(\bar{\Sigma}[(2n)^{\frac{1}{2}}\rho])$. Since $\bar{C}_\beta\bigcap(\bar{\Sigma}[(2n)^{\frac{1}{2}}\rho])$ is a compact set and since $y\to T_\beta{}^y(\varphi)$ is a continuous function of y, the set F_β is compact in the weak topology of $\mathfrak{D}'(\bar{\Sigma}[(2n)^{\frac{1}{2}}\rho])$. Hence,[8] F_β is also compact in the strong topology and is bounded. As a consequence, for any bounded subset X of $\mathfrak{D}(\bar{\Sigma}[(2n)^{\frac{1}{2}}\rho])$ the set of numbers

$$\left\{\left(\frac{\partial T_\beta{}^y}{\partial x_j}\right)(\varphi)\right\},\quad y\in\bar{C}_\beta\bigcap\bar{\Sigma}[(2n)^{\frac{1}{2}}\rho],\quad\varphi\in X$$

is bounded.

Let $\varphi\in\mathfrak{D}(\bar{\Sigma}[(n)^{\frac{1}{2}}\rho])$, $\dot{\varphi}(x)=\varphi(-x)$, and define, for $z=(x+iy)$, $\|x\|^2\leqslant n\rho^2$, $\|y\|^2\leqslant n\rho^2$, $y\in\bar{C}_\beta$,

$$\left.\begin{aligned}G_\beta{}^\varphi(z)&=(T_\beta{}^y*\dot{\varphi})(x)\\&=\int f_\beta(z+\xi)\varphi(\xi)d\xi\\&=\int f_\beta(iy+\xi')\varphi(\xi'-x)d\xi'\end{aligned}\right\}\ \text{if } y\neq 0.$$

We have

$$T_\beta{}^y(\varphi)=G_\beta{}^\varphi(iy);\quad G_\beta{}^\varphi(x_0+x+iy)=G_\beta{}^\psi(x+iy),\quad(3)$$

where $\psi(\xi)=\varphi(\xi-x_0)$.

[8] See L. Schwartz, *Théorie des Distributions* (Hermann & Cie, Paris, France, 1957), tome 1, p. 74 (hereafter referred to as T. D. 1); N. Bourbaki, *Eléments de Mathématique* (Hermann & Cie, Paris, France, 1955), V, Chap. IV, p. 89, def. 5, p. 90, prop. 6 and 7. $\mathfrak{D}(\bar{\Sigma}[(n)^{\frac{1}{2}}\rho])$ is a Montel space by the argument of T. D. 1, p. 70. Its dual is therefore a Montel space. According to Bourbaki, V, Chap. IV, prop. 6, compact (resp. bounded) sets coincide in the strong and weak topologies. Hence Theorem XIII (T. D. 1, p. 74) and the subsequent statements hold for $\mathfrak{D}'(\bar{\Sigma}[(n)^{\frac{1}{2}}\rho])$. See also I. Gelfand and G. Shilov, *Obobshchennye funktsii* (Moscow, 1958), Vol. 2, p. 66 ff.

For $\beta=1, 2$, $G_\beta^\varphi(z)$ is a continuous function of z when $\|x\|^2<n\rho^2$, $\|y\|^2<n\rho^2$, $y\in\bar{C}_\beta$.[9] Moreover, it is analytic in z when $z\in R_\beta'$,

$$R_\beta'=\{\|x\|^2<n\rho^2,\ \|y\|^2<n\rho^2,\ y\in C_\beta\}.$$

By virtue of the remarks made earlier about the boundedness of the set

$$\left\{\frac{\partial T_\beta^y}{\partial x_j}(\varphi)\right\},$$

the set

$$\left\{\frac{\partial}{\partial x_j}G_\beta^\varphi(z)\right\}$$

is bounded for fixed φ when $z\in R_\beta'$, because

$$\frac{\partial}{\partial x_j}G_\beta^\varphi(x+iy)=\left(\frac{\partial T_\beta^y}{\partial x_j}\right)(\psi),\quad \psi(\xi)=\varphi(\xi-x)$$

and because the set of the functions ψ is bounded when φ is fixed and $\|x\|^2<n\rho^2$. But, in R_β', we have

$$\frac{\partial}{\partial x_j}G_\beta^\varphi(x+iy)=i\frac{\partial}{\partial y_j}G_\beta(x+iy).$$

Thus, all the first derivatives of $G_\beta^\varphi(z)$ are bounded for $z\in R_\beta'$. Since $G_1^\varphi(x)=G_2^\varphi(x)=(T*\dot{\varphi})(x)$ for real x, $\|x\|^2<n\rho^2$, we can apply the lemma 2 of Sec. II. We find that, for $\varphi\in\mathfrak{D}(\bar{\Sigma}[(n)^{\frac{1}{2}}\rho])$, there exists a function $G^\varphi(z)$ continuing G_1^φ and G_2^φ, and analytic in the region

$$S'=\{z:\ \|x\|^2<(\rho^2/32),\ y\in K\},$$

where K is the convex closure of $\Sigma(\rho/4\sqrt{2})\bigcap(C_1\bigcup C_2)$. Furthermore,

$$\max_{z\in S'}|G^\varphi(z)|<\max_{\beta=1,2}\ \max_{z'\in\bar{R}_\beta'}|G_\beta^\varphi(z')|.$$

When $\varphi\to 0$ in $\mathfrak{D}(\bar{\Sigma}[(n)^{\frac{1}{2}}\rho])$, $G_\beta^\varphi(z')$ tends to zero uniformly in z', $z'\in\bar{R}_\beta'$. Hence $G^\varphi(z)$ tends to zero uniformly in z, $z\in S'$. This implies that, for $z\in S'$, $\varphi\to G^\varphi(z)$ defines a distribution. Let T^y be the distribution given by $T^y(\varphi)=G^\varphi(iy)$, for $y\in K$,

$$\varphi\in\mathfrak{D}(\bar{\Sigma}[(n)^{\frac{1}{2}}\rho]).$$

[9] It is fairly obvious that $G_\beta^\varphi(z)$ is a continuous function of z for $y\in\bar{C}_\beta$, $y\neq 0$, $\|x\|^2\leqslant n\rho^2$, $\|y\|^2\leqslant n\rho^2$. For $y=0$, $G_\beta^\varphi(x)$ is a uniformly continuous function of real x. Moreover, $G_\beta^\varphi(x+iy)=T_\beta^y(\psi)$, where $\psi(\xi)=\varphi(\xi-x)$. When φ is fixed in $\mathfrak{D}(\bar{\Sigma}[(n)^{\frac{1}{2}}\rho])$, ψ depends continuously on x and describes a compact, therefore bounded set when x describes the compact set $\bar{\Sigma}(n^{\frac{1}{2}}\rho)$. Since $T_\beta^y\to T_\beta^0$ weakly when $y\to 0$, T_β^y also converges strongly to T_β^0, i.e., uniformly on any bounded set. Therefore $G_\beta^\varphi(x+iy)=T_\beta^y(\psi)\to G_\beta^\varphi(x)$ uniformly in x when $y\to 0$. The continuity of $G_\beta^\varphi(x+iy)$ in both x and y follows.

By analytic continuation of (3), we have, for

$$\|x\|<\rho/4\sqrt{2},\quad y\in K,\quad \varphi\in\mathfrak{D}(\bar{\Sigma}[(n^{\frac{1}{2}}-1/4\sqrt{2})\rho]),$$
$$G^\varphi(x+iy)=G^\psi(iy),\quad \psi(\xi)=\varphi(\xi-x),$$

i.e.,

$$G^\varphi(x+iy)=(T^y*\dot{\varphi})(x).$$

Because $G^\varphi(z)$ is analytic in S', T^y is an infinitely differentiable function of $y\in K$ in the weak topology of distributions and, therefore, also in the strong topology. T^y can be considered as a distribution f in $2n$ variables as can be seen by the nuclear theorem, or by defining directly

$$f(\Phi)=\int dy[T^y,\Phi(x,y)],$$

where $\Phi=\Phi(x,y)\in\mathfrak{D}\{\Sigma[((n)^{\frac{1}{2}}-1)\rho]\times K\}$. By considering functions Φ of the form $\varphi(x)\chi(y)$, we find that

$$\partial f/\partial x_j+i\partial f/\partial y_j=0.$$

By applying the regularity theorem[10] to the operator

$$\Delta=\sum_{j=1}^{n}\left(\frac{\partial^2}{\partial x_j^2}+\frac{\partial^2}{\partial y_j^2}\right),$$

we see that f is a C^∞ function, therefore an analytic function in S. It coincides with f_1 and f_2 in their domains of definition because they coincide there in the sense of distributions, while they are C^∞ functions. This completes the proof of the theorem.

ACKNOWLEDGMENTS

I am deeply indebted to Professor H. Rossi, and Professor A. S. Wightman for reading the manuscript of this paper and suggesting several improvements. I also thank Professor V. Bargmann and W. S. Brown for helpful discussions. This work was done under an appointment supported by the International Cooperation Administration under the Visiting Scientists Program administrated by the National Academy of Sciences of the United States of America.

APPENDIX

Proof of Lemma 1

Let E be the open set of all points satisfying

$$|z_1|<\rho\sqrt{2},\quad |z_2|<\rho\sqrt{2}$$
$$0\leqslant t_0|y_1|<y_2$$
$$z_k=\xi_k+\epsilon_k z_2,\quad k=3,\cdots,n$$
$$|\xi_k|<\rho/2,\quad |\epsilon_k|<a;\quad \xi_k,\ \epsilon_k \text{ real}.$$

[10] For proofs of the regularity theorem, see for instance, P. Lax, Communs. Pure and Appl. Math. 8, 615 (1955); L. Schwartz, *Lectures on Elliptic Partial Differential Equations* (National University of Colombia, Bogota, 1956); E. Nelson, lecture notes, Princeton University, 1960.

If we fix ξ_k and ϵ_k $(k=3, \cdots, n)$, $f_1(z)$ and $f_2(z)$ become two functions of z_1 and z_2 satisfying the hypotheses of I. Therefore by virtue of the preceding argument in two variables, there exists, for each admissible value of $\xi=(0,0,\xi_3,\cdots,\xi_n)$ and $\epsilon=(0,0,\epsilon_3,\cdots,\epsilon_n)$, a function $\varphi(z_1,z_2;\xi,\epsilon)$, analytic in z_1 and z_2 for

$$|x_1|<\frac{\rho}{4},\quad |y_1|<\frac{\rho}{4},\quad |x_2|<\frac{\rho}{4},\quad |y_2|<\frac{\rho}{4},$$

$$t_0|y_1|<y_2<\frac{\rho(t_0+\tau)}{4},$$

and coinciding with $f_\alpha(z_1, z_2, \xi_k+\epsilon_k z_2)$ in their domains of definition.

The region E contains all points satisfying:

$$t_0|y_1|<y_2<(t_0+\tau)\frac{\rho}{4},\quad |x_1|<\frac{\rho}{4},\quad |x_2|<\frac{\rho}{4},$$

$$|y_1|<\frac{\rho}{4},\quad y_2<\frac{\rho}{4},$$

$$z_k=\xi_k+\epsilon_k z_2,\quad |\xi_k|<\rho/2,\quad |\epsilon_k|<a\quad (k=3,\cdots,n).$$

The last conditions can be rewritten,

$$|y_k|<ay_2,\quad \left|x_k-\frac{y_k}{y_2}x_2\right|<\frac{\rho}{2}.$$

This is satisfied, in particular, by all points for which

$$t_0|y_1|<y_2<(t_0+\tau)\frac{\rho}{4},\quad |x_1|<\frac{\rho}{4},\quad |x_2|<\frac{\rho}{4},$$

$$|y_1|<\frac{\rho}{4},\quad y_2<\frac{\rho}{4}$$

$$|y_k|<ay_2,\quad |x_k|<\frac{\rho}{4}\quad (k=3,\cdots,n),$$

namely, in $R(\rho)$. We now prove that, if $z\in R(\rho)$, then $F(z)=\varphi(z_1,z_2;\xi,\epsilon)$ is analytic at z. Here it is understood that

$$\xi_k=x_k-y_kx_2/y_2,\quad \epsilon_k=y_k/y_2.$$

Let $B(\rho)$ denote the projection of $R(\rho)$ onto the space of the variables $z_2, \cdots, z_n$. Let $(z_2^0, \cdots, z_n^0)$ be a point in $B(\rho)$. If $z=(z_1, z_2^0, \cdots, z_n^0)$, then $F(z)$ is analytic in z_1 for

$$z_1\in\Delta'(z_2^0)=\{z_1:\ |x_1|<\tfrac{1}{4}\rho,\ |y_1|<y_2/t_0\}.$$

Moreover, $F(z)$ is analytic in all n variables for

$$z_1\in\Delta(z_2^0)=\{z_1:\ |x_1|<\tfrac{1}{4}\rho,\ y_2/(t_0+\tau)<y_1<y_2/t_0\}.$$

Applying the generalized Hartogs theorem (see footnote 4, p. 141), we conclude that $F(z)$ is analytic in $R(\rho)$. Our construction of f, using the result of the two-variable case, makes it obvious that the inequality (1′) holds. The continuity of f at the real points of $\bar{R}(\rho)$ follows from an argument similar to that made in the case of two variables.

"Front" Description in Relativistic Quantum Mechanics

R. ACHARYA AND E. C. G. SUDARSHAN
Department of Physics and Astronomy, University of Rochester, Rochester, New York
(Received May 16, 1960)

The problem of introducing a Cartesian position operator canonically conjugate to the momentum operator into relativistic one-particle theories is investigated independent of any particular relativistic wave equation. The known result that such a description is possible for particles with nonvanishing mass is rederived. The general problem of introduction of canonical variables into relativistic theories is formulated and solved. The configurational indices so obtained correspond to directed plane wavefronts rather than point particles.

1. INTRODUCTION

SPINNING particles are associated with covariant differential equations within the framework of relativistic quantum mechanics.[1] Thus for spin 1/2 particles one has the Dirac equation, for spin 0 particles one has the Klein-Gordon or Kemmer equation; and for the photon one has the Maxwell equation. To demonstrate that these relativistic "wave equations" do describe relativistic quantum-mechanical particles it is necessary to carry out a particle interpretation of these equations.[2] But in most of these particle interpretations there are puzzling features associated with the configurational space descriptions; the appearance of such puzzling features is perhaps best known in the case of the Dirac equation, but they are nevertheless present in the case of the other equations as well.

These paradoxes always arise from an incorrect identification of the covariant amplitude entering the relativistic equation with the Schrödinger amplitude describing the particle. While the covariant form of the equations exhibit the Lorentz invariance of the theory, it is necessary to be able to reduce the covariant equation to the canonical form involving the Schrödinger amplitude, since it is this amplitude which undergoes the unitary transformations under the various operations belonging to the proper inhomogeneous Lorentz group[3]; and this unitary transformation property is basic[4] to a quantum theory with an underlying Hilbert space (the scalar products in which are to be relativistic invariants). In the case of the Dirac equation this reduction from the covariant amplitude to the Schrödinger amplitude is accomplished by the Foldy-Wouthuysen-Tani transformation.[5] In a subsequent paper Foldy has carried through this reduction in a simple manner for the spin 0 and spin 1 equations also. In all these "reduced" forms solutions with both positive and negative frequencies appear on a symmetrical footing and this is a consequence of the covariance of the relativistic equations with respect to the complex Lorentz group.[6]

Two points are to be noted. First of all, the starting point here is a covariant wave equation and one might ask the question as to whether the final results depend on the particular type of wave equation one started with; this is especially important for higher spin equations. We shall show that the position operators are independent of the choice of the relativistic equation and are very simply related to the structure of the unitary irreducible representations of the Lorentz group. Secondly, the standard methods fail for particles with vanishing mass and finite nonzero spin; it is known, for example, that the photon cannot have a localized state in the sense of being an eigenfunction of the three components of the vector position operator.[8] It is the purpose of this paper to investigate what is the maximal configuration specification that one can provide in such a case; it will turn out that the maximal specification corresponds to a "front" form, i.e., the basic elements correspond to directed planes rather than to points. Here again, one starts with the irreducible representations to which the configuration indices are directly related.

2. CANONICAL VARIABLES FOR A RELATIVISTIC PARTICLE

In classical mechanics, with each degree of freedom one associates a pair of dynamical variables p, q which

[1] See, for example, S. S. Schweber, H. A. Bethe, and F. DeHoffman, *Mesons and Fields* (Row, Peterson and Company, New York, 1955) Vol. 1; E. M. Corson, *Tensors, Spinors and Relativistic Wave Equations* (Blackie and Sons, Limited, London, England, 1953).

[2] P. A. M. Dirac, Proc. Roy. Soc. (London) **A155**, 447 (1936); N. Kemmer, *ibid.* **A173**, 91 (1939); R. H. Good, Jr., Phys. Rev. **105**, 1914 (1957).

[3] C. Fronsdal, Phys. Rev. **113**, 1367 (1959).

[4] E. P. Wigner, Ann. Math. **40**, 149 (1939).

[5] L. L. Foldy and S. A. Wouthuysen, Phys. Rev. **78**, 29 (1950); S. Tani, Progr. Theoret. Phys. Kyoto **6**, 267 (1951).

[6] This is an example of the fact that requiring a quantum-mechanical system be describable by "Euclidean" differential equations (i.e., covariant equations not involving the sign of the time components of timelike four-vectors) imposes further restrictions on the system going beyond relativistic invariance alone. All the relativistic wave equations that are usually studied are Euclidean equations and these equations hence remain invariant under the complex Lorentz group. On the other hand, the physical requirement of relativistic invariance does not demand Euclidean invariance. In fact, Euclidean equations cannot be written down for a system with only positive energies; and the appearance of an arbitrary but fixed time direction in every realization of the "particle type" representations of the inhomogeneous Lorentz group is no accident. Similar ideas have been discussed previously[7] by Schwinger and by Nakano.

[7] J. Schwinger, Proc. Natl. Acad. Sci. U. S. 44, 956 (1958); T. Nakano, Progr. Theoret. Phys. Kyoto **21**, 241 (1959).

[8] T. D. Newton and E. P. Wigner: Revs. Modern Phys. **21**, 400 (1949).

obey the fundamental Poisson bracket relations

$$\{p_i,q_j\}=\delta_{ij}, \quad \{p_i,p_j\}=\{q_i,q_j\}=0 \tag{1}$$

at any instant of time, the indices referring to the various degrees of freedom. A triplet of such pairs of canonical variables is associated with a "particle" (without spin), provided the three momenta p_i and the three coordinates q_i transform as the three components of a vector under rotations. Under such transformations the Poisson brackets are preserved and therefore rotations are canonical transformations. For the simplest kind of particles the dynamical variables J_i which constitute the three components of the angular momentum pseudovector are algebraically related to the momenta and coordinates in the form

$$J_i=\epsilon_{ijk}q_jp_k=L_i. \tag{2}$$

It may be, however, that the angular momentum is not equal to this expression but is of the form

$$J_i=L_i+S_i. \tag{3}$$

Then we say that the particle is spinning and the three components S_i constitute the pseudovector spin operator. Since the angular momentum variable is also the infinitesimal generator of rotations, using the expression (1), (2) we obtain

$$\{S_i,p_j\}=\{S_i,q_j\}=0. \tag{4}$$

On using this result together with the fact that the angular momentum operator J is a pseudovector, one obtains the further result

$$\{S_i,S_j\}=\epsilon_{ijk}S_k. \tag{5}$$

Thus the spin variables do not belong to a canonical set, nor are they expressible in terms of the canonical momenta and coordinates.[9] For a "free" spinning particle, the symmetry with respect to translations and rotations in three-dimensional space, i.e., the inhomogenous Euclidean group, requires the momenta p_i and the total angular momenta J_i to be constants of motion. In addition, if the Hamiltonian is independent of spin variables, the spin variables S_i will also be constants of motion. Notice that invariance under the inhomogeneous Euclidean group does not prevent the pseudosclar variable S_ip_i from entering the Hamiltonian; if it does, then the spin vector is no longer a constant of motion, but the "longitudinal spin" or helicity

$$h=\frac{1}{|p|}S_ip_i=\frac{1}{|p|}J_ip_i \tag{6}$$

is a constant of motion; consequently any supplementary condition involving a specific value of h is preserved in time.

Turning to quantum mechanics, the "particle" is now associated with the same set of dynamical variables which are now represented by noncommuting operators. The Poisson brackets are to be related to commutators and by virtue of the relations (5) and (6) the spin and helicity become quantized; however, the important point to notice is that an elementary quantum-mechanical system is associated with irreducible representations of the inhomogeneous Euclidean group in three dimensions and *not* canonical operators. The Euclidean group consists of the six generators of translations and rotations which obey the commutation relations

$$\begin{aligned}&[T_i,T_j]=0,\\ &[R_i,R_j]=i\epsilon_{ijk}R_k,\\ &[R_i,T_j]=i\epsilon_{ijk}T_k.\end{aligned} \tag{7}$$

The two operators

$$T^2=T_iT_i; \quad T\cdot R=R\cdot T=T_iR_i \tag{8}$$

commute with all the six generators T_i, R_i and are therefore represented by numbers in every irreducible representation, the first number being nonnegative. Coordinate or spin operators cannot be defined over these irreducible representations, since the translation generators are identical with the momentum operators so that q_i, for example, would not commute with either of the quantities T^2 or $T\cdot R$.

Such a mixing of the various irreducible representations is already brought about by the requirement of relativistic invariance. The irreducible unitary representations of the inhomogeneous Lorentz group have been investigated by Wigner,[4,10] and he has found several classes of representation. We shall be particularly interested in three such classes: class I corresponds to particles with finite mass and finite spin; class II to particles with zero mass and finite spin; and class III to particles with imaginary mass and zero spin.[11] In the first case the manifold of states corresponds to the irreducible representations of the Euclidean group with $\infty>T^2\geqslant 0$ and $h=-s, -s+1, \cdots s$ where $h=(T^2)^{-\frac{1}{2}}T\cdot R$ and $2s$ is a nonnegative integer; the second class has those with $\infty>T^2\geqslant 0$ and $h=s$ with $2|s|$ an integer; the third class has $\infty>T^2\geqslant|m^2|>0$ and $h=0$. In all cases the energy is real and nonnegative. The Schrödinger amplitude may hence be written[10] as $\psi(\mathbf{p},\zeta)$ with $\mathbf{p}$ and ζ corresponding to the three components of momentum and the single helicity index. The Schrödinger equation becomes

$$i\frac{\partial}{\partial t}\psi(\mathbf{p},\zeta)=+(p^2+m^2)^{\frac{1}{2}}\psi(\mathbf{p},\zeta), \tag{9}$$

[9] We mean by "canonical variables" what Schwinger calls "canonical variables of the first kind"; compare J. Schwinger, *Handbuch der Physik* (to be published).

[10] V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. U. S. 34, 211 (1948).

[11] E. C. G. Sudarshan and V. K. Deshpande (to be published).

and the unitary scalar product

$$(\psi,\phi)=\sum_{\xi}\int d^3p\psi^*(\mathbf{p},\xi)\phi(\mathbf{p},\xi). \tag{10}$$

From (9) one can see that the energy is positive definite.

3. DIRAC EQUATION

Let us now consider the simplest relativistic equation, namely, the Dirac equation which represents particles of spin $\frac{1}{2}$. The covariant differential equation

$$[\gamma^\mu(\partial/\partial x^\mu)+im]\psi=0 \tag{11}$$

can be reduced to the pseudo-Hamiltonian form

$$i\partial\psi/\partial t=(\boldsymbol{\alpha}\cdot\mathbf{p}+\beta m)\psi \tag{12}$$

by multiplication by $i\gamma^0$. It is well known[12] that if one tries to identify $\psi(\mathbf{x},t)$ with a Schrödinger amplitude, so that $\mathbf{x}$ is the representative of the position, then the representative of the velocity is the matrix $\boldsymbol{\alpha}$. This has the consequence that the components of the "velocity" do not commute with each other; and the eigenvalue of any component of the "velocity" is ±1 (in units of the velocity of light) and, further, the velocity and sign of the energy cannot be simultaneously diagonalized. The identification of the covariant amplitude ψ as a Schrödinger amplitude appeared to be sanctioned by the positive definiteness of the probability density $\psi^*\psi$. Rather than reject the identification of $\psi(x)$ with the Schrödinger amplitude, these unusual features were attributed to be a mysterious feature of relativistic equations. Another such feature was in the lack of time independence of the "orbital angular momentum" $\mathbf{x}\times\mathbf{p}$ for the *free* particle.

The correct position operator and localized amplitudes were worked out by various people[13] but the correct identification of the Schrödinger amplitude was made by Foldy and Wouthuysen[5] who showed that the amplitude

$$\varphi(\mathbf{x},t)=-i\exp\left\{\frac{\beta\boldsymbol{\alpha}\cdot\mathbf{p}}{2p}\tan^{-1}\left(\frac{p}{m}\right)\right\}\psi(\mathbf{x},t) \tag{13}$$

satisfies the Schrödinger-like equation

$$i\frac{\partial\varphi}{\partial t}=\beta(p^2+m^2)^{\frac{1}{2}}\varphi. \tag{14}$$

For $\varphi(\mathbf{x},t)$ the standard identification of position and orbital angular momentum operators[14] leads to no unusual features. Notice that according to (13), φ is unitarily related to ψ so that the probability density is unaltered; but the probability current is altered in going from ψ to φ. This alteration is brought about by dropping all terms which mixed positive and negative energies. We also notice that all the identifications of the dynamical variables of the particle commute with the operator β for the sign of the energy. Hence the proper Schrödinger equation is obtained by restricting φ in (14) to contain only positive energy solutions: we would then get the true Schrödinger equation

$$i\partial\Phi/\partial t=+(p^2+m^2)^{\frac{1}{2}}\Phi, \tag{15}$$

where Φ is a two-component amplitude.

With this formulation of the theory we find that the position operator $\mathbf{q}$ has the representative $\mathbf{x}$. In other words, we have a Cartesian position operator with commuting components.[15] Foldy has shown[16] that the Klein-Gordon and Proca fields also can be reduced to the forms (14) and (15) in an analogous manner, so that we can define Cartesian position operators for these systems also. In passing, we also notice that the position operators are canonically conjugate to the momentum operators

$$[x_r,p_s]=i\delta_{rs}. \tag{16}$$

In contrast to the representation introduced by Foldy-Wouthuysen and by Tani (the C representation), another representation (called the E representation) may be introduced[17] in which the analog of (15) is given by

$$i\partial\varphi_E/\partial t=(\boldsymbol{\alpha}\cdot\mathbf{p}/p)(p^2+m^2)^{\frac{1}{2}}\varphi_E. \tag{17}$$

In connection with this amplitude φ_E a new E-position operator was also introduced [which was defined on positive and negative energy solutions of (17) separately] which had several new features. The components of the E-position operator did not commute, but Eq. (16) was satisfied. The "longitudinal component" of the E-position operator

$$\mathbf{x}_E{}^{\text{long.}}=\tfrac{1}{2}(\mathbf{x}_E\cdot\hat{p}+\hat{p}\cdot\mathbf{x}_E)\hat{p}, \tag{18}$$

where

$$\hat{p}=\mathbf{p}/p \tag{19}$$

was identical with the longitudinal component of the C-position operator; but the transverse parts were not identical. Hence the E-position operator could not be used[18] to specify the "localization indices," i.e., a set

[12] P. A. M. Dirac, *Principles of Quantum Mechanics* (Oxford University Press, New York, 1947), 3rd ed.

[13] See, for instance, M. H. L. Pryce, Proc. Roy. Soc. (London) **A195**, 62 (1948).

[14] Foldy and Wouthuysen distinguish these proper identifications by the prefix "mean"; we prefer to omit this prefix since the Dirac position operator (whose representative is the operation of multiplication of the covariant amplitude by x), for example, is *not* an operator defined over the states of the particle (since it mixes the positive energy solutions with negative energy solutions). Of course both the C-position operator as well as the E-position operator introduced in the following are defined over the (positive energy) states of the particle.

[15] Of course, any unitary transform $q=U(p^2)\mathbf{q}U^{-1}(p^2)$ is also a Cartesian position operator and U can be a function of p^2 only if the polar vector transformation property of $\mathbf{q}$ is to be preserved.

[16] L. L. Foldy, Phys. Rev. **102**, 568 (1956). See also K. M. Case, *ibid.* **95**, 1323 (1954).

[17] M. Cini and B. Touschek, Nuovo cimento **1**, 422 (1958); S. K. Bose, A. Gamba, and E. C. G. Sudarshan, Phys. Rev. **113**, 1661 (1959).

[18] The "remedy" suggested by Y. Pac, Progr. Theoret. Phys. Kyoto **22**, 857 (1959), is incorrect since the "mean E-position

of three numbers which could be used to specify a coordinate system in Hilbert space.[19] But the E-position operator has the nice feature that the E-orbital angular momentum is the transverse part of the total angular momentum; and furthermore the E-position operator was defined for a "two-component" neutrino (i.e., a zero mass, spin $\frac{1}{2}$ particle with only one helicity) for which the C-position operator cannot be defined.

Since $\mathbf{x}_E$ satisfies (16) it follows that

$$[\mathbf{x}_E{}^{\text{long.}},\hat{p}]=0. \tag{20}$$

Hence these quantities constitute configuration indices; these are only three independent indices since the unit vector $\hat{p}$ is completely specified by two angles and together with these $\mathbf{x}_E{}^{\text{long.}}$ provide only a distance

$$\begin{aligned}\rho&=\tfrac{1}{2}(\mathbf{x}_E\cdot\hat{p}+\hat{p}\cdot\mathbf{x}_E),\\&=\tfrac{1}{2}(\mathbf{x}\cdot\hat{p}+\hat{p}\cdot\mathbf{x}).\end{aligned} \tag{21}$$

The configurational description provided by the set $(\rho,\hat{p})$ is that of a directed wavefront for which $\hat{p}$ indicates the unit normal and ρ is the perpendicular distance from an arbitrarily chosen origin. The wavefront for a free particle advances normal to itself and the speed of advance

$$v=\partial\rho/\partial t=p/(p^2+m^2)^{\frac{1}{2}} \tag{22}$$

is the speed of a particle of momentum $\mathbf{p}$. Since the values of $(\rho,\hat{p})$ are continuous, the eigenstates are not normalizable but are the limits of normalizable functions. If by $|\rho',\theta,\varphi\rangle$ we represent such an "ideal" state, a normalizable state is given by

$$|f\rangle=f(\rho',\theta,\varphi)|\rho',\theta,\varphi\rangle, \tag{23}$$

with

$$\begin{aligned}\langle f|f\rangle&\equiv\int f^*(\rho',\theta,\varphi)f(\rho',\theta,\varphi)\rho'^2d\rho'\sin\theta d\theta d\varphi.\\&=1.\end{aligned} \tag{24}$$

Since these considerations hold for finite mass and zero mass particles (including the two-component neutrino) one expects them to be of more general validity than the Cartesian configuration description. By a reduction similar to the one employed by Foldy[16] we can demonstrate this result for the Klein-Gordon, Proca, and Maxwell equations, but nothing essentially new is obtained in this fashion. Instead we demonstrate the generality of the "front" form of configurational description by relating the configuration indices $(\rho,\hat{p})$ to the Lorentz group.

It is perhaps appropriate to point out that the choice from among a set of unitarily equivalent position operators[15] is an arbitrary one and is equivalent to the assignment of a specific law for the interaction of the system with prescribed external fields which are "known" to be "localized" in a suitable manner.

4. CONFIGURATIONAL INDICES AND THE LORENTZ GROUP

We have remarked (in Sec. 2) that the proper Lorentz transformations already mix the irreducible representations of the three-dimensional inhomogeneous Euclidean group. For class I particles the manifold[4,10] responds to all possible real values for the three momentum components. Hence in this case the operator $Q_r=i(\partial/\partial P_r)$ which differentiates the momentum amplitude can be defined. For class II particles there is the helicity restriction so that one cannot differentiate freely with respect to the three components but the operator

$$Q^l=\tfrac{1}{2}[(P_rP_r)^{-\frac{1}{2}}P_sQ_s+Q_s(P_rP_r)^{-\frac{1}{2}}P_s] \tag{25}$$

is defined. Finally for class III particles, since the momentum spectrum excludes a sphere of radius $(-m^2)^{\frac{1}{2}}$, not even Q^l is defined.

Let us now construct the generalized E-position operator $\mathbf{q}_E$. For this purpose introduce the operator

$$\mathbf{Q}^{\text{tr}}=(P_rP_r)^{-1}\mathbf{P}\times\mathbf{J} \tag{26}$$

and write

$$\mathbf{q}_E=\mathbf{Q}^{\text{tr}}+Q^l(P_rP_r)^{-\frac{1}{2}}\mathbf{P}. \tag{27}$$

By direct computation one can verify that in the case of the Dirac equation $\mathbf{q}_E$ so defined coincides with the E-position operator. We also notice that for a nonrelativistic spinning particle, for which $\mathbf{J}=\mathbf{L}+\mathbf{S}$,

$$\mathbf{q}_E=\mathbf{x}+(P_rP_r)^{-1}\mathbf{P}\times\mathbf{S}, \tag{28}$$

which shows that this operator differs from the standard position operator by "spin" contributions. In the general case the two terms on the right-hand side may not be separately defined[19] but $\mathbf{q}_E$ is defined in all cases. In complete analogy with the Dirac particle case, the components of $\mathbf{q}_E$ do not commute and one has

$$[q_r{}^E,q_s{}^E]=i\epsilon_{rst}\frac{\mathbf{J}\cdot\mathbf{P}}{(P_rP_r)^2}P_i=i\epsilon_{rst}\frac{\mathbf{S}\cdot\mathbf{P}}{(P_rP_r)^2}P_i \tag{29}$$

so that the lack of commutation is a "spin" effect. Finally the generalized E-orbital angular momentum

$$\mathbf{q}_E\times\mathbf{P}=\mathbf{J}-\mathbf{J}\cdot\hat{P}\hat{P} \tag{30}$$

is equal to the transverse part of the total angular momentum. All these operators are defined for both class I and class II particles. Thus the E-position operator is a more "natural" position operator.[11]

However, for spinning particles $\mathbf{p}_E$ cannot be used as a set of configuration indices, but we can proceed in exactly the same fashion as in the previous section and

operator" introduced there, namely, $\mathbf{x}-(1/2p^2)(\boldsymbol{\sigma}\times\mathbf{p})$ does not commute with the sign of the energy of the E representation and hence is not defined on positive and negative energy states separately.

[19] It is well known that the transverse part of the "spin" is not gauge-invariant in the case of the photon.

introduce the "front" form[20] with

$$\rho = Q^l = \tfrac{1}{2}(\hat{P}\cdot\mathbf{Q} + \mathbf{Q}\cdot\hat{P}). \tag{31}$$

Thus a uniform description of all observed particles (belonging to class I and class II) is made possible in the "front" form.

The question now arises as to when one can introduce a "point" form in terms of a Cartesian position operator in analogy to the C-position operator. From the foregoing demonstration it follows that this is equivalent to the possibility of separating out the spin part [compare Eq. (28)]. For this purpose it suffices to be able to define the pure orbital angular momentum $\mathbf{L}=\mathbf{J}-\mathbf{S}$; this can always be done if Q_r can be defined since $\mathbf{L}=\mathbf{Q}\times\mathbf{P}$. Hence the "spin" can be separated out and the Cartesian configuration indices introduced in the case of class I particles. Thus for class I particles we can use either the "front" form or the "point" form of configurational description but for class II particles only the "front" form is possible. For either class of particles, in the "front" form the helicity may be used along with the configurational indices for a complete specification of the state.

5. DISCUSSION

The development in the previous sections shows that if we are willing to identify the simplest relativistic quantum-mechanical entities with relativistic particles, at least for spinning "particles" of zero mass, the notion of position does exhibit unusual features. Notice that this unfamiliar nature has nothing to do directly with the vanishing mass (and consequent infinite Compton wavelength) since for zero mass particles without spin, one can define localized states; and one could do this for a Dirac particle with vanishing mass. It is curious that such particles do not seem to exist and that the two known particles of zero mass, namely, the neutrino and the photon, do not permit configurational description in the "point" form.

It is of particular significance to note that the photon does not admit a "point" description and hence the statement that a photon is at any specific point at a definite instant is meaningless. And the notion of "signal" propagation and signal velocity in relativistic quantum mechanics is much more subtle than has been generally acknowledged. We also notice that while special relativity requires the invariance of dynamical laws under change of the Lorentz frame, physical interpretation requires the choice of a definite (arbitrary, but fixed) time direction. The specification of the proper quantum-mechanical state and more generally the description of a sequence of quantum-mechanical phenomena is then dependent upon this chosen time direction. This chosen time direction enters in a natural fashion into the realization of the representations of the Lorentz group by (infinite-dimensional) unitary matrices. In the physically interesting cases of free particles considerable ingenuity has gone into the construction of covariant relativistic equations to represent free particles of class I and class II. The elegance of formal manifest covariance is thus an irrelevant feature. The question naturally arises whether one can demand that interacting particles be represented by local manifestly covariant differential equations. If there is any *fundamental* reason for such a requirement, the present authors are unaware of it; and the analysis of the quantum-mechanical description initiated in the foregoing sections casts doubts on the existence of any such reason.

From our point of view the choice of the time direction is a necessary prerequisite to any attempt at physical description; and thus the generalization of the notion of localized states to class II particles given by Fronsdal[3] is unacceptable within this framework.

In this paper we have confined our attention to free particles only. The study of interacting relativistic particles is best done in relation to specific relativistic equations. The systematic analysis along these lines is to be presented in another paper in collaboration with K. Bardakci.

ACKNOWLEDGMENTS

The authors wish to thank I. Bialynicki-Birula for a critical reading of the manuscript and for valuable comments. We are pleased to acknowledge interesting discussions with K. Bardakci.

[20] The definitions and notions introduced here are different in principle from those of Dirac. See P. A. M. Dirac, Revs. Modern Phys. 21, 392 (1949).

On the Nonexistence of a Class of Static Einstein Spaces Asymptotic at Infinity to a Space of Constant Curvature

H. A. BUCHDAHL*
Institute for Advanced Study, Princeton, New Jersey
(Received July 18, 1960)

It is known that there exist no nontrivial static regular solutions of the Einstein vacuum equations $R_{kl}=0$ which are asymptotically Galilean at infinity. One may ask correspondingly whether there exist static solutions of the equations $R_{kl}=\lambda g_{kl}(\lambda<0)$ which are regular at all finite points and asymptotic (in a sense to be defined) to a space of constant curvature at infinity. The answer to this question is here shown to be in the negative. The proof rests upon the possibility of writing a certain quadratic invariant density of the Riemann tensor in the form of an ordinary divergence.

1. INTRODUCTION

IT is well known [1-5] that there exist no static solutions of the Einstein vacuum equations[6]

$$R_{kl}=0 \tag{1.1}$$

which are regular everywhere and asymptotically Galilean at infinity, other than those representing everywhere flat spaces. The static and asymptotically Galilean character of the solution is understood to mean that in a properly adapted coordinate system the metric tensor satisfies the conditions

$$g_{a4}=0, \quad g_{kl,4}=0; \tag{1.2}$$

and

$$\lim_{d\to\infty} g_{kl}=\eta_{kl}, \quad (\eta_{kl}=\mathrm{diag}(-1,-1,-1,+1), \quad d^2=-\eta_{ab}x^a x^b), \tag{1.3}$$

respectively. In the stated result the assumption is implicit that the entire V_4 can be covered with a system of topologically Euclidean coordinates.

The question naturally arises as to whether it may be possible to arrive at any analogous results when one contemplates the Einstein vacuum equations including the cosmological term

$$R_{kl}=\lambda g_{kl}, \tag{1.4}$$

in place of (1.1), that is to say when one considers Einstein spaces ($\lambda<0$) instead of special Einstein spaces ($\lambda=0$). In that case the imposition of conditions regarding "flatness at infinity" plainly become meaningless, and need to be replaced by some other corresponding conditions. This paper therefore considers the following question, made more precise in Sec. 8: do there exist static regular Einstein spaces of a certain type which are asymptotic at infinity to a space of constant (negative) curvature[7] but not of over-all constant curvature? This problem, though somewhat specialized, corresponds to that described earlier in a natural way, and the answer to it turns out to be in the negative.

Now the general methods based on global theorems concerning elliptic linear differential operators[5] on the one hand, and the method of surface integrals in terms of the integrands used previously[1-3] on the other do not seem to be readily adaptable to the case in hand. For the purpose of using the method of surface integrals it will therefore first be shown explicitly how a certain quadratic invariant density $\mathfrak{K}$ of the Riemann tensor may always be written as an *ordinary* divergence $\mathfrak{W}^t{}_{,t}$, whether the metric be static or otherwise. In the non-static case I have been able to obtain $\mathfrak{W}^t$ explicitly only in the form involving the components of the linear spinor connection.[8] Whether this is an essential feature of $\mathfrak{W}^t$, if the latter be required not to involve third or higher derivatives of the g_{kl}, is an open question. In the static case $\mathfrak{W}^t$ can be written in a variety of forms, all of which are remarkable for their simplicity, especially when the V_4 is an Einstein space; $\mathfrak{W}^t$ is then a vector density in V_3, i.e., the three-dimensional subspace $x^4=\mathrm{const}$ [the coordinate system satisfying (1.2)]. At any rate, the present development is intended not so much to show how a particular problem may be solved as to indicate a method whereby certain questions concerning the existence of various types of solutions of the field equations might possibly be answered.

2. QUADRATIC INVARIANTS

(a) In a general V_4 the Riemann tensor possesses four independent quadratic invariants, which may be

* On leave from the Physics Department, University of Tasmania, Hobart, Tasmania, Australia.

[1] R. Serini, Atti accad. nazl. Lincei (5) 27[1], 235 (1918). [Presented in outline in footnote reference 3.]

[2] A. Einstein, Rev. univ. nac. Tucumán Ser. A, 2, 11 (1941).

[3] A. Einstein and W. Pauli, Ann. Math. 44, 131 (1943).

[4] A. Lichnerowicz, Compt. rend. 222, 432 (1946).

[5] A. Lichnerowicz, *Théories relativistes de la gravitation et de l'électromagnétisme* (Masson, Paris, 1955), Chap. 8.

[6] As regards roman indices, those denoted by the first eight letters of the alphabet run from 1 to 3, and the remaining letters from 1 to 4; x^4 is the timelike coordinate.

[7] "Constant curvature" is always intended to mean constant Riemannian curvature.

[8] The spinor analysis involved herein, and the notation used is that of L. Infeld and B. L. van der Waerden, Sitzber. preuss. Akad. Wiss. Physik. math. Kl. (1933), 380.

chosen to be

$$K_1=R^2,\quad K_2=R_{st}R^{st},\quad K_3=R_{klst}R^{klst},$$
$$K_4=e^{klpq}R_{klst}R_{pq}{}^{st},\quad (e^{klst}=(-g)^{-\frac{1}{2}}\epsilon^{klst}). \tag{2.1}$$

It is known[9,10] that (in a V_4) the Hamiltonian derivative of K_4 and of the invariant

$$K=K_1-4K_2+K_3 \tag{2.2}$$

both vanish *identically*. The fact that there are just two such invariants suggests that they may appear in a natural way as the real and imaginary parts, respectively, of a complex invariant S formed of the *spin* curvature tensor.[11] This is, indeed, the case; and the identical vanishing of the Hamiltonian derivative of S may be shown in a particularly simple way, without the introduction of special coordinate nets, etc. Thus consider the invariant

$$S=e^{klst}P_{\beta kl}{}^{\alpha}P_{\alpha st}{}^{\beta}, \tag{2.3}$$

where S is the spin curvature tensor of Infeld and van der Waerden.[8] Expressing it in terms of the Riemann tensor one has

$$S=e^{klst}\sigma^{m\dot{\lambda}}{}_{\alpha}\sigma^{n}{}_{\dot{\lambda}\beta}\sigma^{p\dot{\beta}}{}_{\dot{\rho}\alpha}\sigma^{q}{}_{\dot{\rho}\alpha}R_{mnkl}R_{pqst},$$

or, using the properties of the σ symbols,[12]

$$S=\tfrac{1}{8}e^{klst}(g^{mn}g^{pq}+g^{mq}g^{np}-g^{mp}g^{nq}+ie^{mnpq})R_{mnkl}R_{pqst}$$
$$=-K_4+\tfrac{1}{8}ie^{klst}e^{mnpq}R_{mnkl}R_{pqst}.$$

The second term on the right may be transformed as follows:

$$e^{klst}e^{mnpq}R_{mnkl}R_{pqst}=e^{klst}e_{mnpq}R^{mn}{}_{kl}R^{pq}{}_{st}$$
$$=-\delta_{mnpq}{}^{klst}R_{kl}{}^{mn}R_{st}{}^{pq}=-4K,$$

the generalized Kronecker delta having been written as a determinant of simple Kronecker deltas. Accordingly, one has the required result

$$S=-K_4-\tfrac{1}{2}iK. \tag{2.4}$$

(b) Consider now the integral

$$J=\int(-g)^{\frac{1}{2}}Sd^{(4)}x=\int\epsilon^{klst}P^{\alpha}{}_{\beta kl}P^{\beta}{}_{\alpha st}d^{(4)}x.$$

Then for infinitesimal variations of the spinor connection which vanish on the boundary of the region of integration

$$\delta J=2\int\epsilon^{klst}P^{\alpha}{}_{\beta kl}\delta P^{\beta}{}_{\alpha st}d^{(4)}x$$
$$=4\int\epsilon^{klst}P^{\alpha}{}_{\beta kl}(\delta\Gamma^{\beta}{}_{\alpha t});_{s}d^{(4)}x$$
$$=-4\int\epsilon^{klst}P^{\alpha}{}_{\beta kl;s}\delta\Gamma^{\beta}{}_{\alpha t}d^{(4)}x\equiv 0.$$

In the first step the spinor version of "Palatini's trick" has been used, the second step is the result of an integration by parts, and the last represents an application of the Bianchi identity. These few lines therefore suffice to demonstrate the property of K_4 and K in question.

3. IDENTITY $\mathfrak{R}=\mathfrak{W},_{t}{}^{t}$

From the foregoing results, one will conclude that $\mathfrak{R}[=(-g)^{\frac{1}{2}}K]$ in particular should be expressible as an ordinary divergence. The question therefore arises as to the explicit form of a quantity $\mathfrak{W}^t$ such that $\mathfrak{R}=\mathfrak{W},_{t}{}^{t}$ identically. It may be obtained by inspection as follows. From (2.3) and the definition of $P_{\beta kl}{}^{\alpha}$ one has

$$\mathfrak{S}=4\epsilon^{klst}(\Gamma^{\alpha}{}_{\beta k,l}+\Gamma^{\lambda}{}_{\beta k}\Gamma^{\alpha}{}_{\lambda l})(\Gamma^{\beta}{}_{\alpha s,t}+\Gamma^{\rho}{}_{\alpha s}\Gamma^{\beta}{}_{\rho t})$$
$$=4\epsilon^{klst}\Gamma^{\beta}{}_{\alpha s,t}(\Gamma^{\alpha}{}_{\beta k,l}+2\Gamma^{\lambda}{}_{\beta k}\Gamma^{\alpha}{}_{\lambda l}),$$

the remaining terms vanishing identically, since the antisymmetrized product of an *even* number of Γ's all of whose *spinor* indices are paired so vanishes. With this remark in mind one then sees immediately that

$$\mathfrak{S}=\mathfrak{U},_{t}{}^{t},$$

where

$$\mathfrak{U}^t=4\epsilon^{klst}\Gamma^{\beta}{}_{\alpha s}(\Gamma^{\alpha}{}_{\beta k,l}+\tfrac{2}{3}\Gamma^{\lambda}{}_{\beta k}\Gamma^{\alpha}{}_{\lambda l}). \tag{3.1}$$

Hence in any V_4 (which has the signature -2)

$$\mathfrak{R}=\mathfrak{W},_{t}{}^{t},\quad \mathfrak{W}^t=-2\,\mathrm{Im}\mathfrak{U}^t. \tag{3.2}$$

$\mathfrak{W}^t$ is, of course, determined only to within an arbitrary additive term of the form $f^{[tm]},_{m}$, but as far as integrals over closed surfaces are concerned this is of no consequence.

4. STATIC CASE

(a) When the V_4 is static $\mathfrak{W}^t$ may be exhibited in a very simple tensorial form, the only allowed transformations of coordinates now being of the type $'x^a='x^a(x^b)$. Instead of using (3.1) it is more convenient to start *ab initio*. On taking the metric in the form

$$ds^2=g_{ab}dx^adx^b+g_{44}(dx^4)^2,\quad (g_{kl,4}=0), \tag{4.1}$$

and writing

$$g_{44}=e^{2q}=f^2, \tag{4.2}$$

whichever is the most convenient, the only essentially distinct surviving components of the Riemann tensor

[9] R. Bach, Math. Z. **9**, 110 (1921).
[10] C. Lanczos, Ann. Math. **39**, 842 (1938).
[11] Now and hereafter the signature of V_4 is always understood to be -2.
[12] Harish-Chandra, Proc. Indian Acad. Sci. **23**, 152 (1946).

are easily found to be

$$R^a{}_{bcd} = {}_0R^a{}_{bcd}, \quad R^4{}_{bc4} = q_{;bc} + q_{;b}q_{;c} = t_{bc}, \text{ say.} \quad (4.3)$$

Here tensors having a subscript 0 on the left, covariant differentiation, and juggling of indices all refer to the space whose metric tensor is g_{ab}, i.e., to the subspace V_3 of V_4 defined by $x^4 = \text{constant}$. Then

$$K_3 = ({}_0R_{abcd})({}_0R^{abcd}) + 4R^4{}_{bc4}R_4{}^{bc4}. \quad (4.4)$$

Again,

$$R_{ab} = {}_0R_{ab} + t_{ab}, \quad R_{44} = f^2 t, \quad (t = t_a{}^a); \quad R = {}_0R + 2t. \quad (4.5)$$

Using (4.3–5) one finds thus that

$$K = {}_0K - 8t_{ab}({}_0R^{ab}) + 4{}_0Rt. \quad (4.6)$$

However, in *any*[13] V_3,

$$R_{abcd} = 4g_{[a[d}(R_{c]b]} - \tfrac{1}{4}g_{c]b]}R),$$

which implies ${}_0K = 0$. Now

$$(-g)^{\frac{1}{2}} = f(-\det g_{ab})^{\frac{1}{2}} = fw, \text{ say,} \quad (4.7)$$

and it is convenient also to write $\mathfrak{f}$ for wf, but in so doing one has to remember that $\mathfrak{f}$ does not arise from f through multiplication by $(-g)^{\frac{1}{2}}$. In other words $\mathfrak{f}$ is a scalar density in V_3. Then

$$\mathfrak{R} = 4{}_0R\mathfrak{f}_{;b}{}^b - 8{}_0R^{ab}\mathfrak{f}_{;ab} = 8[(\tfrac{1}{2}g^{ab}{}_0R - {}_0R^{ab})\mathfrak{f}_{;a}]_{;b},$$

in view of the identity of Bianchi. Hence one has finally

$$\mathfrak{W}^a = 8(\tfrac{1}{2}g^{ab}{}_0R - {}_0R^{ab})\mathfrak{f}_{;b}, \quad \mathfrak{W}^4 = 0. \quad (4.8)$$

It may be remarked that $\mathfrak{W}^a$ is itself a divergence (in the covariant sense), viz.,

$$\mathfrak{W}^a = \mathfrak{F}^{ab}{}_{;b}, \quad \mathfrak{F}^{ab} = 8(\tfrac{1}{2}g^{ab}{}_0R - {}_0R^{ab})\mathfrak{f}. \quad (4.9)$$

5. ELEMENTARY APPLICATION

It may be appropriate to interpolate at this stage the derivation of a known result. Consider any static, topologically Euclidean solution of (1.1) which is asymptotically Galilean at infinity, but does not represent a space which is everywhere flat. Let $\mathcal{K}(\rho)$ be the value of the integral of $\mathfrak{R}$ extended over a sufficiently large region W of V_3 bounded by a "spherical" surface Ω: $r = \text{constant} = \rho$, say. Then in virtue of the theorem of Gauss

$$\mathcal{K}(\rho) = \int_\Omega \mathfrak{W}^a n_a d^{(2)}x. \quad (5.1)$$

Using quasi-polar coordinates it is not difficult to confirm on the basis of the linear approximation that $\mathfrak{W}^1$ is at most $0(r^{-3})$ for sufficiently large r. Accordingly the integral on the right of (5.1) tends to zero as $\rho \to \infty$, so that $\mathcal{K}[=\mathcal{K}(\infty)] = 0$. However, if at any *arbitrary* point P one introduces coordinates such that $(g_{kl})_P = \eta_{kl}$ then, at P, K ($= K_3$ here) is the sum of terms of the form $(R_{klmn})^2$, keeping in mind that components of the Riemann tensor an *odd* number of the indices of which have the value 4 here vanish identically. K is therefore positive semidefinite and so must vanish everywhere if $\mathcal{K}$ is to be zero. By the same token the individual summands must vanish separately, i.e., $R_{klmn} = 0$. There exist therefore in fact no solutions of the type here contemplated.

6. STATIC EINSTEIN SPACES

(a) Let it be supposed that the V_4 is a static Einstein space, i.e., the metric (4.1) satisfies the Eqs. (1.4). Then it follows from (4.5) that

$${}_0R^{ab} = \lambda g^{ab} - t^{ab}, \quad (R_4{}^4 =)t = \lambda, \quad (6.1)$$

i.e.,

$$\tfrac{1}{2}g^{ab}{}_0R - {}_0R^{ab} = t^{ab} = f^{-1}f^{;ab}.$$

Equation (4.9) then assumes the remarkably simple form

$$\mathfrak{W}^a = 8\Delta(\mathfrak{f}^{;a}), \quad (6.2)$$

where Δ is the Laplacian operator in V_3. Alternatively one has from (4.8)

$$\mathfrak{W}^a = 4wf^{-1}(f_{;b}f^{;b})^{;a}. \quad (6.3)$$

(b) In the case of a static space of constant (negative) curvature the integral $\mathcal{K}(\rho)$ will obviously diverge as ρ^3 as $\rho \to \infty$, and the same will be true if the Einstein space is in some sense asymptotic to such a space of constant curvature (cf. Sec. 8). It is therefore convenient to introduce the tensor

$$J_{klmn} = R_{klmn} - \tfrac{2}{3}\lambda g_{k[n}g_{m]l}. \quad (6.4)$$

Then

$$K' = J_{klmn}J^{klmn} = K_3 - (8/3)\lambda^2. \quad (6.5)$$

However, because of (1.4), $K_3 = K$, and since the space is static K' is positive semidefinite[14]; thus

$$K' = K - (8/3)\lambda^2 \geq 0. \quad (6.6)$$

If V_4 is of constant curvature K', unlike K, vanishes identically.

(c) Now, by (3.2) and (4.7)

$$\mathfrak{R}'(=(-g)^{\frac{1}{2}}K') = \mathfrak{W}_{,a}{}^a - (8/3)\lambda^2\mathfrak{f}. \quad (6.7)$$

On the other hand, the second member of (6.1) may be written

$$\mathfrak{f}_a{}^{;a} = \lambda\mathfrak{f}. \quad (6.8)$$

Using (5.4) and (6.8) it follows from (6.7) that $\mathfrak{R}'$ also may be written as an ordinary divergence, viz.,

$$\mathfrak{R}' = \mathfrak{H}_{,a}{}^a, \quad \mathfrak{H}^a = 4wf^{-1}(f_{;b}f^{;b} - \tfrac{1}{3}\lambda f^2)^{;a}. \quad (6.9)$$

Incidentally one has here, analogously to (4.9),

$$\mathfrak{H}^a = {}'\mathfrak{F}^{ab}{}_{;b}, \quad {}'\mathfrak{F}^{ab} = 8(\mathfrak{f}^{;ab} - \tfrac{1}{3}\lambda g^{ab}\mathfrak{f}). \quad (6.10)$$

[13] L. P. Eisenhart, *Riemannian Geometry* (Princeton University Press, New Jersey, 1926), Chap. II, p. 91.

[14] See the argument of Sec. 5.

7. STATIC SPACES OF CONSTANT CURVATURE

If a V_n is of constant curvature, then

$$R^k{}_{lst}=\tfrac{2}{3}\lambda g_{l[s}\delta_{t]}{}^k. \tag{7.1}$$

If a V_4 is static it follows from this and the first member of (4.3) that V_3 is also of constant curvature and hence there exist coordinates such that[15]

$$g_{ab}=(1-u)^{-2}\eta_{ab}, \quad 12u=\lambda\eta_{cd}x^c x^d; \tag{7.2}$$

and then

$$f=(1+u)/(1-u). \tag{7.3}$$

If one writes $\lambda=-3k^2$ and introduces new coordinates

$$\begin{gathered} x^1=\bar{r}\sin\theta\cos\phi, \quad x^2=\bar{r}\sin\theta\sin\phi, \\ x^3=\bar{r}\cos\theta, \quad x^4=t, \quad \bar{r}=2k^{-2}r^{-1}[(1+k^2r^2)^{\frac{1}{2}}-1], \end{gathered} \tag{7.4}$$

then the metric of the static space S of constant curvature becomes

$$ds^2=-(1+k^2r^2)^{-1}dr^2-r^2(d\theta^2+\sin^2\theta d\phi^2)+(1+k^2r^2)dt^2. \tag{7.5}$$

If k is real (and this condition will henceforth be imposed), then (7.5) is a static form of the metric of an *open* de Sitter space.

It should be noted that it would be useless to consider closed de Sitter spaces ($\lambda>0$) in the present context since in such a space the spatial distances from the origin of all points of the space have an upper bound.

8. STATIC SPACES S^* ASYMPTOTICALLY OF CONSTANT CURVATURE

This section deals with the definition of those Riemann spaces S^* asymptotic to S which are to be admitted for consideration. An S^* is defined by the following conditions [of which (4) is merely one of convenience]:

(1) there exists a set C of topologically Euclidean coordinates x^s covering the entire space, such that

(2), (3), (4) relative to C the metric tensor of S^* satisfies the conditions (1.2), is regular (of class C^2) at all finite points, and has determinant -1; while

(5) the transformation of coordinates (7.4) gives the metric the form

$$\begin{aligned} ds^2=&-(1+k^2r^2-2mr^{-1}+\omega_1)^{-1}dr^2-r^2(1+\omega_2)d\theta^2 \\ &-r^2(\sin^2\theta+\omega_3)d\phi^2-2r^2\eta d\theta d\phi \\ &+(1+k^2r^2-2mr^{-1}+\omega_4)dt^2, \end{aligned} \tag{8.1}$$

$$m=\text{constant}; \quad \omega_s=0(r^{-2}), \quad \partial_r\omega_s=0(r^{-3}), \quad \partial_\theta\omega_s=0(r^{-2}), \quad \partial_\phi\omega_s=0(r^{-2}),$$

for all $r>r_1$, where r_1 is sufficiently large[16];

(6) the Eqs. (1.4) are satisfied.

[15] Footnote reference 13, Chap. 2, p. 85.

[16] For the purposes of later sections weaker conditions upon the asymptotic behavior of ω_2 and ω_3, and of all the first derivatives of ω_s other than $\partial_r\omega_4$ would actually be adequate.

The function η is fixed by condition (4) since the determinant g of the metric tensor must now be $-r^4\sin^2\theta$. The somewhat strange choice of the generic form of the metric is one of convenience; it is important for g to have the chosen form, but in that case one cannot in general orthogonalize the metric completely.

In the first place it may be remarked that when $k=0$, so that S is flat (though this possibility will hereafter be excluded), the usual linearized field equations show that with a suitable choice of coordinate system the metric tensor must (then) have the form (8.1) to the required order. Further, the same is true ($k\neq 0$ now) when V_3 is spherically symmetric, for then (8.1) with $\omega_s=0$ is in fact an exact solution of the field equations when r is not too small. If one adopts a physical viewpoint[17] one may look upon (8.1) as representing the asymptotic form of the solution of the field equations corresponding to a finite distribution of "particles" about the origin. Then in a sense the foregoing definition implies the assumption that the effects of the "multipole moments" of the distribution asymptotically fall off more rapidly than the effects of the "monopole moment." It is possible that the formal counterpart to this behavior may be deducible from the field equations. However, though it is easy to write down the generic linearized equations, viz.,

$$\Box h_{st}+h_{;st}-2h_{(s;t)}=2k^2(h_{st}-g_{st}h), \quad (h=h_s{}^s,\ h_s=h_s{}^t{}_{;t}), \tag{8.2}$$

where h_{st} are the "infinitesimal" differences between corresponding components of the metric tensors of S^* and S, the explicit form of these equations is so complex that I have hitherto been unable to arrive at any definite conclusions in this respect. At any rate, the definition of S^* given in the foregoing will here be adopted.

9. PROOF THAT m MUST VANISH

Equation (6.8) may be written

$$(wg^{ab}f_{,b})_{,a}=\lambda wf,$$

or, since $wf=r^2\sin\theta$,

$$(r^2\sin\theta g^{ab}f^{-1}f_{,b})_{,a}=\lambda r^2\sin\theta. \tag{9.1}$$

Now integrate both members of (9.1) over W, taking $\rho>r_1$ (cf. Sec. 5), and apply the theorem of Gauss, keeping in mind condition (3) of Sec. 8. Then

$$\int_\Omega (r^2\sin\theta g^{11}f^{-1}f_{,1})_{r=\rho}d\theta d\phi=-4\pi k^2\rho^3. \tag{9.2}$$

Using (8.1), the left-hand member of (9.2) is easily evaluated, the result being $-4\pi[k^2\rho^3+m+0(\rho^{-1})]$. By allowing ρ to tend to infinity it therefore follows at once that m must be zero.

[17] One should then, to be consistent, restrict m to be non-negative.

10. PROOF THAT S^* MUST BE OF OVER-ALL CONSTANT CURVATURE

The theorem of Gauss may now be applied also to the integral of $\mathfrak{R}'$ extended over W. Thus, because of (6.9), using a nomenclature analogous to that occurring in Sec. 4.

$$\mathcal{K}'(\rho)=\int_W \mathfrak{R}' d^{(3)}x=\int_\Omega (\mathfrak{H}^1)_{r=\rho} d\theta d\phi. \tag{10.1}$$

Now from (8.1), when $r>r_1$,

$$g^{bc}f_{,b}f_{,c}=-k^2[k^2r^2+2mr^{-1}+0(r^{-2})],$$
$$-\tfrac{1}{3}\lambda f^2=k^2[k^2r^2+1-2mr^{-1}+0(r^{-2})],$$
$$g^{11}wf^{-1}=-[1+0(r^{-4})]r^2\sin\theta,$$

where despite the result of the previous section m has not been set equal to zero for the time being. It therefore follows that

$$\mathcal{K}'(\rho)=-16\pi mk^2+0(\rho^{-1}),$$

and therefore as $\rho\to\infty$

$$\mathcal{K}'=-16\pi mk^2. \tag{10.2}$$

Setting $m=0$ at this stage in accordance with the result of Sec. 9 one has $\mathcal{K}'=0$. Then the same kind of argument which was used in Sec. 5 to show that there R_{klmn} had to vanish will show here that one has to have

$$J_{klmn}=0, \tag{10.3}$$

i.e., S^* must be of over-all constant curvature.

As a final remark it may be noticed that if one imposes the restriction $m\geq 0$ from the outset (cf. footnote reference 17) then Sec. 9 may be omitted entirely, and at the same time one need no longer bother with the normalization of g. This is at once obvious from (10.2) since certainly $\mathcal{K}'\geq 0$.

ACKNOWLEDGMENTS

I should like to express my sincere thanks to Professor J. R. Oppenheimer for the splendid hospitality, and partial support, of the Institute for Advanced Study. I am also very grateful to Professor V. Bargmann for critically reading the manuscript, and to Dr. R. S. Palais for an informative discussion.

Galvanomagnetic and Thermomagnetic Effects in Isotropic Materials

A. C. PIPKIN AND R. S. RIVLIN
Brown University, Providence, Rhode Island
(Received August 29, 1960)

Invariant-theoretical considerations are employed to obtain constitutive equations for the current density vector, the heat flux vector, and the magnetic intensity field in isotropic materials (both holohedral and hemihedral) when an electric field, a magnetic induction field, and a temperature gradient are simultaneously present in the material. Certain of the interaction effects which are indicated by these constitutive equations are discussed.

INTRODUCTION

IN two previous papers,[1,2] the manner in which invariant-theoretical considerations may be used to derive nonlinear constitutive equations in continuum physics has been discussed. In a subsequent paper,[3] these ideas are applied to the formulation of the constitutive equations for electrical or thermal conduction in an isotropic solid which undergoes deformation. In the present paper, a similar approach is taken to the formulation of constitutuve equations which describe the possible effects on the electrical current density vector $\mathbf{J}$, the heat flux vector $\mathbf{q}$, and magnetic field $\mathbf{H}$ resulting from the simultaneous presence in a material of an electric field $\mathbf{E}$, a temperature gradient τ, and a magnetic induction field $\mathbf{B}$. The materials with which we are concerned in this paper are isotropic and both the cases of holohedral and hemihedral isotropy are considered. However, the approach taken may readily be applied to materials with other types of symmetry. It is assumed that the materials with which we are concerned do not undergo deformation, i.e., thermal expansion, electrostrictive and magnetostrictive effects are neglected. Methods similar to those used in the present paper could be extended to include such effects.

In formulating the constitutive equations, $\mathbf{J}$, $\mathbf{q}$, and $\mathbf{H}$ have been taken as the dependent variables and $\mathbf{E}$, τ, and $\mathbf{B}$ as the independent variables. This is not a unique choice and some other choice of three dependent and three independent variables from these six vectors might prove preferable from certain points of view; however, for any particular choice the resulting constitutive equations can easily be obtained by appropriate substitutions from those derived here, to which they are, of course, algebraically equivalent.

In the present paper, the constitutive equations are first derived without restriction on the magnitude of the variables involved. It can easily be seen that they describe many well-known thermoelectric, galvanomagnetic, and thermomagnetic effects. They also indicate the possibility of the existence of further effects which, as far as the author is aware, have not yet been reported. A complete analysis of all such effects would probably not be a worthwhile task out of the context of an experimental program aimed at observing them, since their nature is fairly evident directly from the constitutive equations.

The general constitutive equations derived in this paper are specialized to the cases when they are linear in the independent variables, of second degree in these, and of third degree.[4] The constitutive equations so obtained may be expected, for a given material, temperature, etc., to apply for increasingly large ranges of values of the independent variables.

The ranges over which each of these sets of constitutive equations provide a close approximation to experiment will depend, of course, on the material, its temperature, and so on. From the experimental results (see, for example, Jan[5]) on such galvanomagnetic and thermomagnetic effects as the Hall effect, the magnetoresistive effects, the Nernst and von Ettinghausen effects, it is apparent that at any rate in solids the magnitude of such interaction effects generally increases as the temperature is lowered. With this increase comes the need for constitutive equations which include higher-degree terms and possibly those of the full generality obtained in this paper.

In previous work, constitutive equations equivalent to the first-, second-, and third-order equations considered here have been derived for the holohedral isotropic case and for the various crystal classes. The constitutive equation is taken (as in the present paper) to be a polynomial relation between a dependent vector and a number of independent vectors. The restrictions imposed by the material symmetry on the coefficients of terms of given partial degrees in the independent vectors are then considered. This is usually an extremely laborious procedure and can only be carried out for terms of relatively low total degree. A brief description of some of this work is given in the review article by Jan,[5] where references to the original papers are given. This approach is also taken in a review article by Smith.[6] In contrast, the more powerful invariant-theoretical method adopted in the present

[1] A. C. Pipkin and R. S. Rivlin, Arch. Rational Mech. Anal. 4, 129 (1959).
[2] R. S. Rivlin, Arch. Rational Mech. Anal. 4, 262 (1960).
[3] A. C. Pipkin and R. S. Rivlin, J. Math. Phys. 1, 127 (1960).
[4] The third-degree constitutive equations are discussed explicitly only in the case of holohedral materials.
[5] J.-P. Jan, in *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic Press Inc., New York, 1957), Vol. 5, p. 1.
[6] C. S. Smith in *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic Press Inc., New York, 1957), Vol. 6, p. 175.

paper yields constitutive equations of closed form which are applicable generally, without limitation on the degree of the polynomial.

In some of the earlier work, the further restrictions imposed on the form of the constitutive equations by the assumption that they must obey Onsager's principle are also discussed. Such restrictions are not discussed here since it is considered that the area over which Onsager's principle may legitimately be employed is in considerable doubt.

2. CONSTITUTIVE EQUATION FOR THE ELECTRICAL CURRENT

We assume that an electric field $\mathbf{E}$, a magnetic induction field $\mathbf{B}$, and a temperature gradient field τ act in an electrical conductor or semiconductor. Let $\mathbf{J}$ be the associated current density vector. Then we may assume that $\mathbf{J}$ is a polynomial[7] function of the vectors $\mathbf{E}$, $\mathbf{B}$, and τ, thus:

$$\mathbf{J}=\mathbf{F}(\mathbf{E},\mathbf{B},\tau). \tag{2.1}$$

We now consider the restrictions imposed on the constitutive equation (2.1) by symmetry of the material. Let $\{\mathbf{S}\}$ denote the group of orthogonal transformations characterizing the symmetry of the material and let $\mathbf{S}$ be a generic transformation of this group. Then

$$\mathbf{S}\cdot\mathbf{S}^t=\mathbf{S}^t\cdot\mathbf{S}=\mathbf{I}, \tag{2.2}$$

where $\mathbf{S}^t$ denotes the transpose of $\mathbf{S}$ and $\mathbf{I}$ is the unit matrix.

The constitutive equation (2.1) must be form-invariant under the group of transformations $\{\mathbf{S}\}$. This implies that

$$\mathbf{J}^*=\mathbf{F}(\mathbf{E}^*,\mathbf{B}^*,\tau^*), \tag{2.3}$$

where

$$\mathbf{J}^*=\mathbf{S}\cdot\mathbf{J},\quad \mathbf{E}^*=\mathbf{S}\cdot\mathbf{E},\quad \mathbf{B}^*=\pm\mathbf{S}\cdot\mathbf{B}$$

and

$$\tau^*=\mathbf{S}\cdot\tau, \tag{2.4}$$

the positive sign in the expression (2.4) for $\mathbf{B}^*$ being taken if $\mathbf{S}$ is a proper orthogonal transformation and the negative sign if it is an improper orthogonal transformation.

From (2.1), (2.3), and (2.4), we obtain

$$\mathbf{F}(\mathbf{E}^*,\mathbf{B}^*,\tau^*)=\mathbf{S}\cdot\mathbf{F}(\mathbf{E},\mathbf{B},\tau), \tag{2.5}$$

where $\mathbf{E}^*$, $\mathbf{B}^*$, τ^* and $\mathbf{E}$, $\mathbf{B}$, τ are related by (2.4). If ψ is an arbitrary absolute vector and

$$\psi^*=\mathbf{S}\cdot\psi, \tag{2.6}$$

we obtain from (2.5), with (2.2),

$$\psi^*\cdot\mathbf{F}(\mathbf{E}^*,\mathbf{B}^*,\tau^*)=\psi\cdot\mathbf{F}(\mathbf{E},\mathbf{B},\tau)=\mathfrak{F}(\text{say}). \tag{2.7}$$

$\mathfrak{F}$ is then an absolute scalar invariant under the group $\{\mathbf{S}\}$ of the three absolute vectors ψ, $\mathbf{E}$, and τ and the axial vector $\mathbf{B}$. It may therefore be expressed as a polynomial in the elements of an integrity basis for these vectors. We note that it is linear in ψ. We note also that

$$\mathbf{F}(\mathbf{E},\mathbf{B},\tau)=\partial\mathfrak{F}/\partial\psi. \tag{2.8}$$

[7] We mean by this that each component of $\mathbf{J}$ in a given rectangular Cartesian coordinate system is a polynomial in the components of $\mathbf{E}$, $\mathbf{B}$, and τ in the same system.

3. HEMIHEDRAL ISOTROPIC MATERIALS

If the material considered is a hemihedral isotropic material, then the group $\{\mathbf{S}\}$ describing its symmetry is the proper orthogonal group. We may, in this case, treat the vector $\mathbf{B}$ as though it were an absolute vector. On omitting elements which are nonlinear in ψ, an integrity basis for the vectors ψ, $\mathbf{E}$, $\mathbf{B}$, and τ under the proper orthogonal group is formed by (see, for example, Weyl[8])

$$\psi\cdot\mathbf{E},\quad \psi\cdot\mathbf{B},\quad \psi\cdot\tau,\quad [\psi,\mathbf{E},\mathbf{B}],\quad [\psi,\tau,\mathbf{B}],\quad [\psi,\mathbf{E},\tau] \tag{3.1}$$

and

$$\mathbf{E}\cdot\mathbf{E},\quad \mathbf{B}\cdot\mathbf{B},\quad \tau\cdot\tau,\quad \mathbf{E}\cdot\mathbf{B},\quad \tau\cdot\mathbf{B},\quad \mathbf{E}\cdot\tau,\quad [\mathbf{E},\tau,\mathbf{B}], \tag{3.2}$$

where square brackets denote the scalar triple product. $\mathfrak{F}$ is therefore expressible in the form

$$\mathfrak{F}=\alpha_1\psi\cdot\mathbf{E}+\alpha_2\psi\cdot\mathbf{B}+\alpha_3\psi\cdot\tau+\alpha_4[\psi,\mathbf{E},\mathbf{B}]+\alpha_5[\psi,\tau,\mathbf{B}]+\alpha_6[\psi,\mathbf{E},\tau], \tag{3.3}$$

where the α's are polynomials in (3.2). From (3.3), (2.8), and (2.1) we obtain

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_2\mathbf{B}+\alpha_3\tau+\alpha_4\mathbf{E}\times\mathbf{B}+\alpha_5\tau\times\mathbf{B}+\alpha_6\mathbf{E}\times\tau. \tag{3.4}$$

4. HOLOHEDRAL ISOTROPIC MATERIALS

The constitutive equation for a holohedral isotropic material may be obtained from the constitutive Eq. (3.4) for a hemihedral isotropic material by introducing the further requirement that it be form-invariant under the central inversion transformation $\mathbf{S}=(-1,-1,-1)$.

We note that for this transformation, we have

$$\begin{aligned}&\mathbf{E}^*=-\mathbf{E},\quad \mathbf{B}^*=\mathbf{B},\quad \tau^*=-\tau,\quad \mathbf{J}^*=-\mathbf{J},\\ &\mathbf{E}^*\times\mathbf{B}^*=-\mathbf{E}\times\mathbf{B},\quad \tau^*\times\mathbf{B}^*=-\tau\times\mathbf{B},\\ &\qquad\qquad \mathbf{E}^*\times\tau^*=\mathbf{E}\times\tau,\\ &\mathbf{E}^*\cdot\mathbf{E}^*=\mathbf{E}\cdot\mathbf{E},\quad \mathbf{B}^*\cdot\mathbf{B}^*=\mathbf{B}\cdot\mathbf{B},\quad \tau^*\cdot\tau^*=\tau\cdot\tau,\\ &\mathbf{E}^*\cdot\mathbf{B}^*=-\mathbf{E}\cdot\mathbf{B},\quad \mathbf{E}^*\cdot\tau^*=\mathbf{E}\cdot\tau,\quad \tau^*\cdot\mathbf{B}^*=-\tau\cdot\mathbf{B}\end{aligned} \tag{4.1}$$

and

$$[\mathbf{E}^*,\tau^*,\mathbf{B}^*]=[\mathbf{E},\tau,\mathbf{B}].$$

We therefore have, from (3.4), that

$$\begin{aligned}&\alpha_1(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{E}+\alpha_2(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{B}+\alpha_3(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\tau\\ &\quad+\alpha_4(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{E}\times\mathbf{B}+\alpha_5(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\tau\times\mathbf{B}\\ &\quad+\alpha_6(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{E}\times\tau\\ &=\alpha_1(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{E}-\alpha_2(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{B}\\ &\quad+\alpha_3(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\tau+\alpha_4(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{E}\times\mathbf{B}\\ &\quad+\alpha_5(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\tau\times\mathbf{B}\\ &\qquad\qquad-\alpha_6(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{E}\times\tau,\end{aligned} \tag{4.2}$$

[8] H. Weyl, *The Classical Groups* (Princeton University Press, Princeton, New Jersey, 1946).

polynomial dependence of the α's on $\mathbf{E}\cdot\mathbf{E}$, $\mathbf{B}\cdot\mathbf{B}$, $\tau\cdot\tau$, $\mathbf{E}\cdot\tau$, and $[\mathbf{E},\tau,\mathbf{B}]$ being understood. It follows from (4.2) that α_2 and α_6 must be polynomials of odd degree in $\mathbf{E}\cdot\mathbf{B}$ and $\tau\cdot\mathbf{B}$, while the remaining α's are of even degree in $\mathbf{E}\cdot\mathbf{B}$ and $\tau\cdot\mathbf{B}$.

The constitutive equation for the electrical current in a holohedral isotropic material may therefore be written as

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_3\tau+\alpha_4\mathbf{E}\times\mathbf{B}+\alpha_5\tau\times\mathbf{B}+[\alpha_2'\mathbf{E}\cdot\mathbf{B}+\alpha_2''\tau\cdot\mathbf{B}]\mathbf{B}+[\alpha_6'\mathbf{E}\cdot\mathbf{B}+\alpha_6''\tau\cdot\mathbf{B}]\mathbf{E}\times\tau, \quad (4.3)$$

where the α's are polynomial in

$$\mathbf{E}\cdot\mathbf{E},\ \mathbf{B}\cdot\mathbf{B},\ \tau\cdot\tau,\ \mathbf{E}\cdot\tau,\ [\mathbf{E},\tau,\mathbf{B}],\ (\mathbf{E}\cdot\mathbf{B})(\tau\cdot\mathbf{B})\ (\mathbf{E}\cdot\mathbf{B})^2$$

and

$$(\tau\cdot\mathbf{B})^2. \quad (4.4)$$

The constitutive Eq. (4.3) may be expressed in an even simpler form by employing the identities

$$(\mathbf{E}\cdot\mathbf{B})\mathbf{E}\times\tau\equiv[\mathbf{E},\tau,\mathbf{B}]\mathbf{E}-(\mathbf{E}\cdot\mathbf{E})\tau\times\mathbf{B}+(\mathbf{E}\cdot\tau)\mathbf{E}\times\mathbf{B}$$

and (4.5)

$$(\tau\cdot\mathbf{B})\mathbf{E}\times\tau\equiv[\mathbf{E},\tau,\mathbf{B}]\tau+(\tau\cdot\tau)\mathbf{E}\times\mathbf{B}-(\mathbf{E}\cdot\tau)\tau\times\mathbf{B}$$

which may be easily derived from the identity (9.3) in the Appendix by making appropriate substitutions. On using these relations, we see that (4.3) may be expressed in the form

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_3\tau+\alpha_4\mathbf{E}\times\mathbf{B}+\alpha_5\tau\times\mathbf{B}+[\alpha_2'\mathbf{E}\cdot\mathbf{B}+\alpha_2''\tau\cdot\mathbf{B}]\mathbf{B}, \quad (4.6)$$

where the α's are polynomials in the quantities (4.4).

5. FIRST- AND SECOND-ORDER CONSTITUTIVE EQUATIONS FOR THE CURRENT

If we neglect terms of higher degree than the first in $\mathbf{E}$, $\mathbf{B}$, and τ in the constitutive equation for the current, then in the hemihedral case it takes the form

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_2\mathbf{B}+\alpha_3\tau, \quad (5.1)$$

and in the holohedral case it takes the form

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_3\tau, \quad (5.2)$$

where the α's are constants. Equations (5.1) and (5.2) are the first-order constitutive equations for hemihedral and holohedral isotropic materials, respectively.

If we neglect terms of higher degree than the second in $\mathbf{E}$, $\mathbf{B}$, and τ, the constitutive equation for the hemihedral isotropic material becomes

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_2\mathbf{B}+\alpha_3\tau+\alpha_4\mathbf{E}\times\mathbf{B}+\alpha_5\tau\times\mathbf{B}+\alpha_6\mathbf{E}\times\tau \quad (5.3)$$

and that for the holohedral isotropic material becomes

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_3\tau+\alpha_4\mathbf{E}\times\mathbf{B}+\alpha_5\tau\times\mathbf{B}, \quad (5.4)$$

where again the α's are constants. Equations (5.3) and (5.4) are the second-order constitutive equations for hemihedral and holohedral isotropic materials, respectively.

We shall now make the assumption[9] that the material is such that $\mathbf{J}=0$ when $\mathbf{E}=0$ and $\tau=0$. Then $\alpha_2=0$ in (5.1) and (5.3). In the first of these cases the constitutive equation takes the form (5.2), i.e., the first-order constitutive equations for hemihedral and holohedral isotropic materials are the same. Taking $\tau=0$ in (5.2), we see that α_1 is the ohmic electrical conductivity of the material. Taking $\mathbf{E}=0$, we see that a temperature gradient τ gives rise to an electrical current in the same direction, the Thomson effect.

Turning now to the second-order constitutive equations, we see that if $\alpha_2=0$ and we take $\tau=0$, the equations for the hemihedral and holohedral cases become the same and take the form[10]

$$\mathbf{J}=\alpha_1\mathbf{E}+\alpha_4\mathbf{E}\times\mathbf{B}. \quad (5.5)$$

α_1 is again the electrical conductivity and the coefficient α_4 determines the magnitude of the Hall effect. If $\tau\neq0$, the second-order constitutive equations differ in the holohedral and hemihedral cases. In the holohedral case, we see that if $\mathbf{B}=0$, it reduces to the first-order constitutive equation. Taking $\mathbf{E}=0$, we see that if a temperature gradient τ and magnetic induction field $\mathbf{B}$ exist in the material, then apart from the current resulting from the Thomson effect an additional current perpendicular to τ and $\mathbf{B}$ will, in general, be produced (the Nernst effect). This effect may also arise in the hemihedral isotropic case, but in this case we observe the possibility of a further effect resulting from the presence in the constitutive Eq. (5.3) of the term $\alpha_6\mathbf{E}\times\tau$. This leads to the possibility[11] that if an electric field and nonparallel temperature gradient exist, a current may be produced at right-angles to both of these even if $\mathbf{B}=0$.

6. CONSTITUTIVE EQUATIONS FOR THE HEAT FLUX

We now assume that the heat flux vector $\mathbf{q}$ is also a polynomial function of the vectors $\mathbf{E}$, $\mathbf{B}$, and τ. Since $\mathbf{q}$ is an absolute vector, we may, by considerations similar to those used in discussing the constitutive equation for the electrical current, derive a constitutive equation for the heat flux vector which is similar in form to (3.4) in the hemihedral case and to (4.6) in the

[9] This assumption need not necessarily be valid for all hemihedral isotropic materials.

[10] To the degree of approximation involved in deriving this equation, it is equivalent to the equation

$$\mathbf{E}=a_1\mathbf{J}+a_4\mathbf{J}\times\mathbf{B},$$

where a_1 and a_4 are constants. The latter equation is usually employed when a constitutive equation which describes the Hall effect is required. However, it is usually derived from a consideration of the mechanism which results in the Hall effect. Here, Eq. (5.5) and the remaining constitutive equations are derived purely from invariant-theoretical considerations.

[11] As far as the author is aware, this effect has so far not been observed.

holohedral case. In both cases, we merely replace $\mathbf{J}$ by $\mathbf{q}$ and bear in mind that the α's are not the same functions of their arguments as in the constitutive equations for the current density vector. The first- and second-order constitutive equations for $\mathbf{q}$ may be obtained in a manner similar to that used in deriving the first- and second-order constitutive equations for $\mathbf{J}$. If we make the assumption that $\mathbf{q}=0$ when $\mathbf{E}=0$ and $\tau=0$, we obtain as the first-order constitutive equation, in both the hemihedral and holohedral cases,

$$\mathbf{q}=\beta_1\mathbf{E}+\beta_3\tau, \tag{6.1}$$

where the β's are constants.

The second-order constitutive equation for hemihedral isotropic materials becomes

$$\mathbf{q}=\beta_1\mathbf{E}+\beta_3\tau+\beta_4\mathbf{E}\times\mathbf{B}+\beta_5\tau\times\mathbf{B}+\beta_6\mathbf{E}\times\tau \tag{6.2}$$

and that for holohedral isotropic materials becomes

$$\mathbf{q}=\beta_1\mathbf{E}+\beta_3\tau+\beta_4\mathbf{E}\times\mathbf{B}+\beta_5\tau\times\mathbf{B}, \tag{6.3}$$

where the β's are again constants.

From Eq. (6.1) we see that β_3 is the thermal conductivity of the material. Provided β_1 is not zero, we see that an electric field $\mathbf{E}$ in the material gives rise to a parallel heat flux. From Eq. (6.3), it is seen that for a holohedral isotropic material in which nonparallel electric and magnetic induction fields exist, a heat flux may be produced perpendicular to both of these (the von Ettinghausen effect). Also, if a temperature gradient and nonparallel magnetic induction field exist in the material a heat flux may be produced perpendicular to both of these (the Righi-Leduc effect).

These effects may also be obtained in a hemihedral isotropic material. It is seen from Eq. (6.2) that, in addition, if an electric field and nonparallel temperature gradient exist in the material, a heat flux perpendicular to both of these may be produced.[12]

7. CONSTITUTIVE EQUATION FOR THE MAGNETIC FIELD INTENSITY

We assume that the magnetic field intensity $\mathbf{H}$ is a polynomial function of the vectors $\mathbf{E}$, $\mathbf{B}$, and τ, thus

$$\mathbf{H}=\mathbf{G}(\mathbf{E},\mathbf{B},\tau). \tag{7.1}$$

Again, this constitutive equation must be form-invariant under the group of transformations $\{\mathbf{S}\}$ describing the symmetry of the material. This implies that

$$\mathbf{G}(\mathbf{E},\mathbf{B},\tau)=\pm\mathbf{S}\cdot\mathbf{G}(\mathbf{E},\mathbf{B},\tau), \tag{7.2}$$

where $\mathbf{S}$ is a generic transformation of the group $\{\mathbf{S}\}$ and the positive or negative sign is taken accordingly as $\mathbf{S}$ is a proper or improper orthogonal transformation. $\mathbf{E}^*$, $\mathbf{B}^*$, τ^* and $\mathbf{E}$, $\mathbf{B}$, τ are related by (2.4).

If the material is hemihedral isotropic, the restriction (7.2) implies, with (7.1), that $\mathbf{H}$ must be expressible in the form

$$\mathbf{H}=\gamma_1\mathbf{B}+\gamma_2\mathbf{E}+\gamma_3\tau+\gamma_4\mathbf{E}\times\mathbf{B}+\gamma_5\tau\times\mathbf{B}+\gamma_6\mathbf{E}\times\tau, \tag{7.3}$$

where the γ's are polynomial functions of the quantities (3.2). This result may be derived in precisely the same manner as the constitutive Eq. (3.4) for the electric current density, from the fact that the positive sign applies in (7.2) for proper orthogonal transformations.

If the material is holohedral isotropic, then (7.3) must be form-invariant under the central inversion transformations $\mathbf{S}=(-1,-1,-1)$. By using the results (4.1) and the relation $\mathbf{H}^*=\mathbf{H}$, we obtain

$$\begin{aligned}&\gamma_1(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{B}+\gamma_2(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{E}+\gamma_3(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\tau\\&\quad+\gamma_4(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{E}\times\mathbf{B}+\gamma_5(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\tau\times\mathbf{B}\\&\quad+\gamma_6(\mathbf{E}\cdot\mathbf{B},\tau\cdot\mathbf{B})\mathbf{E}\times\tau\\&\quad=\gamma_1(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{B}-\gamma_2(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{E}\\&\qquad-\gamma_3(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\tau-\gamma_4(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{E}\times\mathbf{B}\\&\qquad-\gamma_5(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\tau\times\mathbf{B}\\&\qquad+\gamma_6(-\mathbf{E}\cdot\mathbf{B},-\tau\cdot\mathbf{B})\mathbf{E}\times\tau,\end{aligned} \tag{7.4}$$

polynomial dependence of the γ's on $\mathbf{E}\cdot\mathbf{E}$, $\mathbf{B}\cdot\mathbf{B}$, $\tau\cdot\tau$, $\mathbf{E}\cdot\tau$, and $[\mathbf{E},\tau,\mathbf{B}]$ being understood. It follows from (7.4) that γ_1 and γ_6 must be polynomials of even degree in $\mathbf{E}\cdot\mathbf{B}$ and $\tau\cdot\mathbf{B}$, while the remaining γ's are of odd degree in these arguments.

The constitutive equation for the magnetic field intensity $\mathbf{H}$ in a holohedral isotropic material may therefore be written in the form

$$\begin{aligned}\mathbf{H}=&\gamma_1\mathbf{B}+[\gamma_2'\mathbf{E}\cdot\mathbf{B}+\gamma_2''\tau\cdot\mathbf{B}]\mathbf{E}+[\gamma_3'\mathbf{E}\cdot\mathbf{B}+\gamma_3''\tau\cdot\mathbf{B}]\tau\\&+[\gamma_4'\mathbf{E}\cdot\mathbf{B}+\gamma_4''\tau\cdot\mathbf{B}]\mathbf{E}\times\mathbf{B}\\&+[\gamma_5'\mathbf{E}\cdot\mathbf{B}+\gamma_5''\tau\cdot\mathbf{B}]\tau\times\mathbf{B}+\gamma_6\mathbf{E}\times\tau,\end{aligned} \tag{7.5}$$

where the γ's are polynomials in the quantities (4.4).

By using a relation of the type (9.3), we may, without loss of generality, take $\gamma_4''=0$ or $\gamma_5'=0$ in (7.5).

If we neglect terms in the constitutive Eq. (7.3) for $\mathbf{H}$ of higher degree than the first in $\mathbf{E}$, $\mathbf{B}$, and τ, we obtain the first-order constitutive equation for the magnetic field intensity in a hemihedral isotropic material:

$$\mathbf{H}=\gamma_1\mathbf{B}+\gamma_2\mathbf{E}+\gamma_3\tau, \tag{7.6}$$

where the γ's are constants. Similarly, we may obtain from (7.5) the first-order constitutive equation for $\mathbf{H}$ in a holohedral isotropic material:

$$\mathbf{H}=\gamma_1\mathbf{B}. \tag{7.7}$$

If we further assume that $\mathbf{H}=0$ when $\mathbf{B}=0$, then $\gamma_2=\gamma_3=0$ in (7.6) and it is seen that (7.7) represents the first-order constitutive equation for $\mathbf{H}$ for both holohedral and hemihedral isotropic materials.

The second-order constitutive equation for $\mathbf{H}$ in a hemihedral isotropic material is, with $\mathbf{H}=0$ when $\mathbf{B}=0$,

$$\mathbf{H}=\gamma_1\mathbf{B}+\gamma_4\mathbf{E}\times\mathbf{B}+\gamma_5\tau\times\mathbf{B}, \tag{7.8}$$

[12] As far as the author is aware this effect has so far not been observed.

where the γ's are constants. The corresponding equation for a holohedral isotropic material is (7.7).

8. THIRD-ORDER CONSTITUTIVE EQUATIONS FOR HOLOHEDRAL ISOTROPIC MATERIALS

If, in the constitutive equations for a holohedral isotropic material, we neglect terms of higher degree then the third in $\mathbf{E}$, $\mathbf{B}$, and $\boldsymbol{\tau}$, we obtain, with the condition that $\mathbf{H}=0$ when $\mathbf{B}=0$,

$$\mathbf{J}=(\alpha_1+\alpha_{11}\mathbf{E}\cdot\mathbf{E}+\alpha_{12}\mathbf{B}\cdot\mathbf{B}+\alpha_{13}\boldsymbol{\tau}\cdot\boldsymbol{\tau}+\alpha_{14}\mathbf{E}\cdot\boldsymbol{\tau})\mathbf{E}$$
$$+(\alpha_3+\alpha_{31}\mathbf{E}\cdot\mathbf{E}+\alpha_{32}\mathbf{B}\cdot\mathbf{B}+\alpha_{33}\boldsymbol{\tau}\cdot\boldsymbol{\tau}+\alpha_{34}\mathbf{E}\cdot\boldsymbol{\tau})\boldsymbol{\tau}$$
$$+\alpha_4\mathbf{E}\times\mathbf{B}+\alpha_5\boldsymbol{\tau}\times\mathbf{B}+[\alpha_2'\mathbf{E}\cdot\mathbf{B}+\alpha_2''\boldsymbol{\tau}\cdot\mathbf{B}]\mathbf{B}, \quad (8.1)$$

and

$$\mathbf{H}=(\gamma_1+\gamma_{11}\mathbf{B}\cdot\mathbf{B}+\gamma_{12}\mathbf{E}\cdot\mathbf{E}+\gamma_{13}\boldsymbol{\tau}\cdot\boldsymbol{\tau}+\gamma_{14}\mathbf{E}\cdot\boldsymbol{\tau})\mathbf{B}$$
$$+[\gamma_2'\mathbf{E}\cdot\mathbf{B}+\gamma_2''\boldsymbol{\tau}\cdot\mathbf{B}]\mathbf{E}+[\gamma_3'\mathbf{E}\cdot\mathbf{B}+\gamma_3''\boldsymbol{\tau}\cdot\mathbf{B}]\boldsymbol{\tau}, \quad (8.2)$$

where the α's and γ's are constants. The constitutive equation for $\mathbf{q}$ is similar in form to (8.1).

We shall now consider some effects, additional to those discussed for the first- and second-order equations, the possibility of which is allowed by Eq. (8.1). If $\boldsymbol{\tau}=0$, Eq. (8.1) becomes

$$\mathbf{J}=(\alpha_1+\alpha_{11}\mathbf{E}\cdot\mathbf{E}+\alpha_{12}\mathbf{B}\cdot\mathbf{B})\mathbf{E}+\alpha_4\mathbf{E}\times\mathbf{B}+(\alpha_2'\mathbf{E}\cdot\mathbf{B})\mathbf{B}. \quad (8.3)$$

Let us suppose that the vectors $\mathbf{E}$ and $\mathbf{B}$ are inclined at an angle φ. We choose the reference system x in such a way that the x_1 axis is parallel to $\mathbf{E}$ and the x_2 axis is in the plane formed by $\mathbf{E}$ and $\mathbf{B}$, so that $\mathbf{E}=(E,0,0)$ and $\mathbf{B}=(B\cos\varphi,\ B\sin\varphi,\ 0)$. It then follows from (8.3) that

$$J_1=[\alpha_1+\alpha_{11}E^2+(\alpha_{12}+\alpha_2'\cos^2\varphi)B^2]E,$$
$$J_2=\alpha_2'EB^2\sin\varphi\cos\varphi, \quad (8.4)$$
$$J_3=\alpha_4EB\sin\varphi.$$

If $\alpha_{11}\neq 0$, we have non-Ohmic electrical conductivity. The presence of a magnetic induction field $\mathbf{B}$ produces a change of current in the direction of $\mathbf{E}$ proportional to B^2. Also provided that $\varphi\neq 0$ or $\pi/2$, and $\alpha_2'\neq 0$, it also gives rise to a current proportional to B^2 in the x_2 direction. The current density component J_3 is, of course, the Hall current given by the second-order constitutive equation.

We now consider that $\mathbf{E}=0$, but $\boldsymbol{\tau}\neq 0$. Equation (8.1) then becomes

$$\mathbf{J}=(\alpha_3+\alpha_{32}\mathbf{B}\cdot\mathbf{B}+\alpha_{33}\boldsymbol{\tau}\cdot\boldsymbol{\tau})\boldsymbol{\tau}+\alpha_5\boldsymbol{\tau}\times\mathbf{B}+\alpha_2''(\boldsymbol{\tau}\cdot\mathbf{B})\mathbf{B}. \quad (8.5)$$

We suppose that the vectors $\boldsymbol{\tau}$ and $\mathbf{B}$ are inclined at an angle φ. Then, in a manner similar to that employed in discussing the case when $\boldsymbol{\tau}=0$ but $\mathbf{E}\neq 0$, taking $\boldsymbol{\tau}=(\tau,0,0)$ and $\mathbf{B}=(B\cos\varphi,\ B\sin\varphi,\ 0)$, we obtain

$$J_1=[\alpha_3+\alpha_{33}\tau^2+(\alpha_{32}+\alpha_2''\cos^2\varphi)B^2]\tau,$$
$$J_2=\alpha_2''\tau B^2\sin\varphi\cos\varphi, \quad (8.6)$$
$$J_3=\alpha_5\tau B\sin\varphi.$$

If $\alpha_{33}\neq 0$, we have a nonlinear Thomson effect. The presence of a magnetic induction field $\mathbf{B}$ produces a change, proportional to B^2, in the Thomson coefficient relating the temperature gradient and the current density in the direction of this gradient. Also, provided that $\varphi\neq 0$ or $\pi/2$, and $\alpha_2''\neq 0$, we obtain a current proportional to B^2 in the x_2 direction. The current density component J_3 is, of course, obtained from the second-order constitutive equations and is associated with the Nernst effect.

Since the constitutive equation for the heat flux vector is similar to that for the electric current density vector, the effects of the magnetic induction field on the heat flux vector are similar to those on the electric current density.

We shall now discuss the constitutive equation (8.2) for the magnetic field intensity in a similar manner. If $\boldsymbol{\tau}=0$, but $\mathbf{E}\neq 0$, the equation becomes

$$\mathbf{H}=(\gamma_1+\gamma_{11}\mathbf{B}\cdot\mathbf{B}+\gamma_{12}\mathbf{E}\cdot\mathbf{E})\mathbf{B}+(\gamma_2'\mathbf{E}\cdot\mathbf{B})\mathbf{E}. \quad (8.7)$$

Then, choosing the reference system x so that

$$\mathbf{B}=(B,0,0) \quad \text{and} \quad \mathbf{E}=(E\cos\varphi,\ E\sin\varphi,\ 0),$$

we obtain

$$H_1=[\gamma_1+\gamma_{11}B^2+(\gamma_{12}+\gamma_2'\cos^2\varphi)E^2]B,$$
$$H_2=\gamma_2'E^2B\sin\varphi\cos\varphi \quad \text{and} \quad H_3=0. \quad (8.8)$$

Similarly, if $\mathbf{E}=0$, but $\boldsymbol{\tau}\neq 0$, Eq. (8.2) becomes

$$\mathbf{H}=(\gamma_1+\gamma_{11}\mathbf{B}\cdot\mathbf{B}+\gamma_{13}\boldsymbol{\tau}\cdot\boldsymbol{\tau})\mathbf{B}+(\gamma_3''\boldsymbol{\tau}\cdot\mathbf{B})\boldsymbol{\tau}. \quad (8.9)$$

Then, taking $\mathbf{B}=(B,0,0)$ and $\boldsymbol{\tau}=(\tau\cos\varphi,\ \tau\sin\varphi,\ 0)$, we obtain

$$H_1=[\gamma_1+\gamma_{11}B^2+(\gamma_{13}+\gamma_3''\tau^2\cos^2\varphi)]B,$$
$$H_2=\gamma_3''\tau^2B\sin\varphi\cos\varphi \quad \text{and} \quad H_3=0. \quad (8.10)$$

9. APPENDIX

If δ_{ij} denotes the three-dimensional Kronecker delta and e_{ijk} the three-dimensional alternating symbol, then[1]

$$\delta_{ij}e_{klm}-\delta_{ik}e_{jlm}+\delta_{il}e_{jkm}-\delta_{im}e_{jkl}=0. \quad (9.1)$$

If A_i, B_i, C_i, and D_i are the components of four vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$, then multiplying (9.1) throughout by $A_iB_jC_kD_l$, we obtain

$$A_iB_ie_{klm}C_kD_l-A_iC_ie_{jlm}B_jD_l+A_iD_ie_{jkm}B_jC_k$$
$$-A_me_{jkl}B_jC_kD_l=0. \quad (9.2)$$

This relation may be rewritten in vector notation as

$$(\mathbf{A}\cdot\mathbf{B})\mathbf{C}\times\mathbf{D}=(\mathbf{A}\cdot\mathbf{C})\mathbf{B}\times\mathbf{D}-(\mathbf{A}\cdot\mathbf{D})\mathbf{B}\times\mathbf{C}$$
$$+[\mathbf{B},\mathbf{C},\mathbf{D}]\mathbf{A}. \quad (9.3)$$

ACKNOWLEDGMENTS

The results presented in this paper were obtained in the course of research sponsored by the Office of Ordnance Research, U. S. Army. Our thanks are due to Dr. R. A. Toupin for interesting discussions relating to this work.

Computation of Order Parameters in an Ising Lattice by the Monte Carlo Method*

J. R. EHRMAN,† L. D. FOSDICK, AND D. C. HANDSCOMB‡
Digital Computer Laboratory, University of Illinois, Urbana, Illinois
(Received May 20, 1960)

The long-range and short-range order parameters are computed for the Ising lattice using a Monte Carlo sampling scheme. The square lattice, the simple cubic lattice, and the body-centered cubic lattice are considered. In the three-dimensional calculations both the antiferromagnetic and ferromagnetic cases are considered as well as the coupling to an external magnetic field of various strengths. Good agreement is found where the results can be compared with the exact two-dimensional results, and in the three-dimensional case the results agree well with those obtained from series approximations in the regions where the series approximations are valid. The present method appears to give good results for the short-range order even very close to the critical temperature, but in this neighborhood the long-range order estimate is crude. The computations were performed on the high-speed computer ILLIAC, located at the University of Illinois.

I. INTRODUCTION

IN a recent investigation[1] the Monte Carlo method was used to compute parameters describing the short-range and long-range order in a face-centered cubic binary alloy. That investigation was preceded by some computations[2] of order parameters in a two-dimensional Ising lattice[3] as a test of the feasibility of the Monte Carlo method for this kind of calculation. This early work was quite successful and the continuation of this work to a treatment of three-dimensional Ising lattices is the subject of the present paper. Since the early work on the two-dimensional lattice has not been previously reported in detail, it has been included in the present discussion. Two three-dimensional lattices are treated in the present work: the simple cubic and the body-centered cubic. This treatment includes both ferromagnetic and antiferromagnetic coupling, and coupling to an external magnetic field. Parameters describing the short-range order and the long-range order have been computed.

The method used here was first used by Metropolis and others[4] to treat the two-dimensional hard sphere gas, and it has been used subsequently by others[5–8] for further computations on the equation of state of gases. The essence of this method can be described briefly as follows. A sequence of configuration states for the system is developed using transition probabilities p_{ij}, where p_{ij} gives the probability that state i will be followed immediately by state j. These transition probabilities are chosen to make the distribution of states in the sequence tend toward a Boltzmann distribution as the number of states in the sequence increases. At some point the sequence is truncated and, neglecting some of the initial states in the chain, the states of the truncated sequence are used as an ensemble to estimate the average value of certain system parameters; in the present case, estimates of the average value of the order parameters are computed. It seems appropriate to describe this approach as a "mathematical experiment" because it is somewhat analogous to observing the parameters directly in the real physical system, as in a physical experiment. In the latter case nature provides the averaging, whereas in the mathematical experiment this is simulated on a model. It should be quickly pointed out, however, that the kinetics associated with the mathematical experiment do not necessarily represent those of the real system; they may represent the real system kinetics to some degree, but it is not essential to the method. This approach can provide a very good physical picture of the microscopic character of the system and it is therefore capable of providing new insights to the problem which might be very difficult to obtain from a more conventional, analytical approach in which the system is represented in a comparatively abstract fashion.

Without a high-speed computing machine this approach would not be feasible. It is not surprising therefore that interest in this method has increased as these machines have become more available, and it is to be expected that this interest will continue to grow as the capabilities of these machines grow. The ILLIAC, a high-speed computing machine located at the Digital Computer Laboratory of the University of Illinois, was used to perform the computations described in the present work.

* Supported in part by the Office of Naval Research.
† Present address: Dept. of Physics, University of Illinois.
‡ Research Assistant visiting from Oxford University (September, 1958–June, 1959).
[1] L. D. Fosdick, Phys. Rev. **116**, 565 (1959).
[2] L. D. Fosdick, Bull. Am. Phys. Soc. Ser. II, **2**, 239 (1957).
[3] For a review article on the Ising model of ferromagnetism, see G. F. Newell and E. W. Montroll, Revs. Modern Phys. **25**, 353 (1953).
[4] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[5] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **22**, 881 (1954).
[6] W. W. Wood and F. R. Parker, J. Chem. Phys. **27**, 720 (1957).
[7] W. W. Wood and J. D. Jacobson, J. Chem. Phys. **27**, 1207 (1957).
[8] Z. W. Salsburg, J. D. Jacobson, W. Fickett, and W. W. Wood, J. Chem. Phys. **30**, 65 (1959).

II. GENERAL REMARKS ON APPLICATION OF MONTE CARLO METHOD TO ISING LATTICE PROBLEM

The fundamental ideas of the method which was used in these computations have been discussed in the references cited in the introduction, especially footnote references 1, 4, and 6. A familiarity with these ideas will be assumed and here our attention will be directed at their application to the Ising lattice problem.

Each site of the Ising lattice has an associated two-valued spin coordinate $\mu(k)$ for the kth site; $\mu(k)=+1$ or -1, corresponding to the two allowed orientations of the spin on the kth site. The ith configuration state of a lattice of N sites is completely described by the N-component vector μ_i, whose components are the N spin coordinates, $\mu_i(k)$. The energy of this spin array is assumed to be due to nearest-neighbor interactions and the interaction with an external magnetic field. In particular, the energy of the ith configuration is given by

$$E_i=-J\sum_{k',k}^{(1)}\mu_i(k)\mu_i(k')+H\sum_k\mu_i(k), \tag{1}$$

where the first sum extends over all pairs of sites in the lattice such that k and k' are nearest neighbors and the second sum extends over all sites in the lattice. It will be recognized that the change in the coupling energy for a nearest-neighbor pair going from a state of parallelism to a state of antiparallelism is $2J$, i.e., $\epsilon(\uparrow\downarrow)-\epsilon(\uparrow\uparrow)=2J$. The array is ferromagnetic if J is positive and antiferromagnetic if J is negative. On referring again to Eq. (1), it will be observed that the change in the energy arising from the external field coupling is $2H$ when a spin changes from a state of parallelism with the field to a state of antiparallelism with the field, i.e., $\epsilon(\uparrow)-\epsilon(\downarrow)=2H$. Since it is convenient and customary to use parameters which are the ratios of the coupling energies to kT, where k is the Boltzmann constant and T is the absolute temperature, we define $K=J/kT$ and $L=H/kT$.

The order in the system, at equilibrium, is computed as a function of K and L. To describe the short-range order, the parameters $f_i(j)$ are used where $f_i(j)$ is the fraction of jth neighbor sites which are occupied by an antiparallel pair of spins in the ith configuration

$$f_i(j)=\frac{1}{\alpha(j)N}\sum_{k',k}^{(j)}[1-\mu_i(k)\mu_i(k')], \tag{2}$$

where $\alpha(j)$ is the number of jth neighbors of a lattice site, and the summation extends over all pairs of sites k and k' where k and k' are jth neighbors. Using Boltzmann statistics, the average value of this parameter for jth neighbors is given by the usual formula

$$f(j)=\sum_i f_i(j)e^{-E_i/kT}/\sum_i e^{-E_i/kT}, \tag{3}$$

where the summations extend over all configuration states of the system. In the Monte Carlo estimation of this quantity the expression on the right is replaced by the average over a small sample of $f_i(j)$'s drawn from a distribution in which the ith configuration state tends to be proportional to $\exp(-E_i/kT)$. This proportionality holds strictly only in the limit as the sequential process of generating new configurations is continued indefinitely. However, the proportionality usually holds with sufficient accuracy for worthwhile results when the sequence is truncated in order to keep the computing time within reasonable bounds. For the two-dimensional square lattice and the body-centered cubic lattice the average of the first-neighbor order parameter $f(1)$ is estimated. In the simple-cubic lattice both $f(1)$ and $f(2)$ are estimated.

The long-range order of the ith configuration state is described by the parameter S_i, where

$$S_i=\frac{1}{N}\left|\sum_k\mu_i(k)\right|, \tag{4}$$

the summation extending over all sites. The average value of S_i, denoted by S, is given by the formula analogous to Eq. (3) and the Monte Carlo estimate of S is likewise obtained from sampling as indicated before for $f(j)$. Estimates of S have been obtained for the two three-dimensional lattices. It is known that for an infinite lattice and $L=0$, S vanishes at a critical value of K, denoted by K_c, and that for $K\leq K_c$, S remains equal to zero. The finite system used in these computations can only be expected to approximate this behavior, and one expects a rapid decrease in S in the neighborhood of K_c, but S will remain nonzero, though small, even when $K\ll K_c$. When the external magnetic field is zero, S can be identified as the spontaneous magnetization per spin.[9] For a lattice composed of an infinite number of spins a very small positive magnetic field removes all states of the lattice from the ensemble average for which the total spin is negative, but with a finite lattice this is not strictly true. This approximation to the spontaneous magnetization of an infinite array will be worst near the critical temperature where the difference in behavior between the infinite system and the finite system becomes particularly important.

The procedure for generating the ensemble of configurations over which the averaging is to be performed is very similar to the one described in footnote reference 1. We are given a lattice of N sites which is in a particular configuration state; say i; hence $\mu_i(k)$ is known for all k. A site of the lattice is selected and the change in energy ΔE which would accompany a reversal in orientation of the spin on that site is computed. If $\Delta E\leq 0$, then the spin is reversed and if $\Delta E>0$, then the spin orientation is reversed with probability $\exp(-\Delta E/kT)$. The latter process is done by generating a pseudo-random number ξ from a uniform distribution

[9] C. N. Yang, Phys. Rev. 85, 808 (1952).

on the interval (0,1) and if $\xi<\exp(-\Delta E/kT)$, the spin orientation is reversed; otherwise it is not reversed. Next, another site is selected and the process is repeated. The sites were numbered and selected sequentially and when all sites of the lattice had been treated once, as just described, an iteration of the calculation was said to be completed. At the end of an iteration the values of the order parameters $f_i(j)$ and S_i were recorded; the details of this step varied somewhat in the different calculations and they will be described more fully in the following sections. The iterations continued until a condition for stopping was satisfied at which point the last R samples which were generated were used to compute the averages

$$f(1)=\frac{1}{R}\sum_{(i)} f_i(1), \quad f(2)=\frac{1}{R}\sum_{(i)} f_i(2), \quad S=\frac{1}{R}\sum_{(i)} S_i, \tag{5}$$

where in each case the sum extends over the last R samples. In the three-dimensional computations the stopping condition was based on a comparison of results from two independent computations, similar to that described in footnote reference 1; the details will be presented later.

It is not difficult to see that this process satisfies the necessary and sufficient conditions[1] for producing a sequence of configurations which tends toward a Boltzmann distribution as the length of the sequence tends toward infinity. The ergodic condition is satisfied: Any pair of configurations can be linked by at least one finite sequence of configurations in which one spin at a time is reversed, and the probability for this reversal is nonzero. The other condition is also satisfied: The 2^N component probability vector Ψ with components

$$\Psi_i=e^{-E_i/kT}/\textstyle\sum_i e^{-E_i/kT} \quad (i=1, 2, \cdots, 2^N) \tag{6}$$

is an eigenvector, with eigenvalue unity, of the stochastic matrix P, whose elements p_{ij} are the transition probabilities linking the 2^N configuration states of the lattice in a complete iteration. This can be seen as follows. The stochastic matrix P may be regarded as the product of N matrices $P(1)$, $P(2)\cdots P(N)$, where $P(k)$ has components $p_{ij}(k)$ which are the probabilities for a transition from state i to j by a reversal in the orientation of the spin on site k. For given i there is exactly one j, say j', for which $p_{ij}(k)\neq 0$ and $p_{ii}(k)=1-p_{ij'}(k)$. If $E_i>E_{j'}$ we have $p_{ij'}(k)=1$ and if $E_i<E_{j'}$, we have $p_{ij'}(k)=\exp[-(E_{j'}-E_i)/kT]$. Similarly, for given j there is exactly one i, say i' for which $p_{i'j}(k)\neq 0$, where $i'\neq j$. In short, there will be either one or two nonzero elements in every row and column of $P(k)$. Consider the product

$$\Psi P(k)=\Phi,$$

and in particular

$$\textstyle\sum_i \Psi_i p_{ij}(k)=\phi_j. \tag{7}$$

Let $p_{i'j}(k)$ be the nonzero off-diagonal element in the jth column of $P(k)$. If $E_{i'}\geq E_j$, then $p_{i'j}=1$ and $p_{jj}=1-\exp[-(E_{i'}-E_j)/kT]$ and it follows from Eqs. (6) and (7) that

$$\phi_j=Ae^{-E_{i'}/kT}+[1-e^{-(E_{i'}-E_j)/kT}]Ae^{-E_j/kT}$$

hence

$$\phi_j=Ae^{-E_j/kT},$$

where

$$A^{-1}=\textstyle\sum_i e^{-E_i/kT}.$$

If $E_{i'}<E_j$, then $p_{i'j}=e^{-(E_j-E_{i'})/kT}$ and $p_{jj}=0$ and it follows from Eqs. (6) and (7) that

$$\phi_j=[e^{-(E_j-E_{i'})/kT}]Ae^{-E_{i'}/kT},$$

hence

$$\phi_j=Ae^{-E_j/kT}.$$

It follows that $\Phi\equiv\Psi$, hence Ψ is an eigenvector of $P(k)$ with eigenvalue unity and since this is true for all k it is true for the product $P=P(1)\ P(2)\cdots P(N)$. It is to be noted that this result is independent of the order in which the sites are numbered. This result is also true for

$$Q=\frac{1}{N}\sum_{k=1}^{N} P(k),$$

which is the transition matrix when a site is selected at random and the same process performed on it. The details of the individual computations follow.

III. SQUARE LATTICE

The two critical items affecting the practicality of the method, namely the rate of convergence of the generated ensemble to a Boltzmann distribution and the size of the lattice needed for worthwhile results, were investigated using this model. The short-range order parameter $f(1)$ was computed and compared with the exact value for an infinite system.[10] The computations were made at different values of the parameter $x=e^{-2K}$ and with zero external field, $L=0$. Periodic boundary conditions were imposed by linking the left edge of the lattice to the right edge by nearest-neighbor bonds and similarly the top edge was linked to the bottom edge.

The lattice configuration was represented by binary numbers in the computer, with each binary digit corresponding to a lattice site and the value of that digit representing the orientation of the spin on the site. In each iteration of the computation the sites were selected for consideration systematically. The successive sites along a row were considered until the end of the row was reached and then the sites in the next adjacent row were considered until finally all N sites had been considered.

At the completion of one iteration the value of $f_i(1)$ was recorded. To examine the behavior of estimates of $f(1)$ as it is computed at different points along the

[10] B. Kaufman and L. Onsager, Phys. Rev. **76**, 1244 (1949).

TABLE I. Values of $f(1,j)$ obtained for the square lattice. The exact value for $f(1)$ (footnote reference 10) appears at the end of the right-hand column for every x.

		10×10 lattice		20×20 lattice		37×37 lattice	
x	j	Initially ordered	Initially disordered	Initially ordered	Initially disordered	Initially ordered	Initially disordered
0.30	1	0.0209±0.0020	0.0428±0.0054	0.0222±0.0010	0.0760±0.0045	0.0229±0.0006	0.0444±0.0037
	2	0.0238±0.0023	0.0240±0.0023	0.0220±0.0009	0.0220±0.0012	0.0230±0.0005	0.0228±0.0006
	3	0.0246±0.0021	0.0246±0.0021	0.0234±0.0011	0.0228±0.0013	0.0223±0.0006	0.0227±0.0006
	4	0.0199±0.0019	0.0194±0.0019				
	5	0.0204±0.0019	0.0200±0.0018				
	6	0.0208±0.0020	0.0212±0.0021			$f(1)=0.0223$ (exact)	
0.40	1	0.0956±0.0044	0.1338±0.0074	0.1043±0.0026	0.1198±0.0041	0.1026±0.0015	0.1497±0.0037
	2	0.1083±0.0043	0.1194±0.0061	0.0992±0.0021	0.1007±0.0023	0.1155±0.0013	0.1104±0.0017
	3	0.0896±0.0045	0.1052±0.0052	0.1027±0.0026	0.1076±0.0027	0.1112±0.0016	0.1192±0.0013
	4	0.0884±0.0044	0.0947±0.0046				
	5	0.0912±0.0044	0.0940±0.0050				
	6	0.0867±0.0041	0.0886±0.0043			$f(1)=0.1074$ (exact)	
0.43	1	0.1650±0.0050	0.1702±0.0058	0.1656±0.0036	0.1879±0.0041	0.1819±0.0017	0.1859±0.0028
	2	0.1568±0.0053	0.1802±0.0066	0.1748±0.0032	0.1841±0.0030	0.1833±0.0015	0.1968±0.0014
	3	0.1768±0.0070	0.1690±0.0061	0.1708±0.0030	0.1630±0.0032	0.1759±0.0018	0.1958±0.0014
	4	0.1436±0.0056	0.1621±0.0055				
	5	0.1479±0.0055	0.1501±0.0058				
	6	0.1640±0.0061	0.1403±0.0052			$f(1)=0.1900$ (exact)	
0.45	1	0.2085±0.0057	0.2024±0.0058	0.2145±0.0028	0.2091±0.0029	0.2127±0.0018	0.2186±0.0022
	2	0.1858±0.0055	0.2170±0.0059	0.2162±0.0032	0.2236±0.0024	0.2236±0.0011	0.2293±0.0012
	3	0.2061±0.0062	0.2131±0.0056	0.2718±0.0027	0.2305±0.0024	0.2271±0.0011	0.2279±0.0014
	4	0.1892±0.0051	0.1800±0.0057	0.2316±0.0024	0.2256±0.0030		
	5	0.1968±0.0048	0.1986±0.0052				
	6	0.2018±0.0058	0.2110±0.0055			$f(1)=0.2245$ (exact)	
0.50	1	0.2820±0.0053	0.2904±0.0045	0.2833±0.0026	0.2898±0.0021	0.2784±0.0016	0.2838±0.0018
	2	0.2808±0.0052	0.2812±0.0043	0.2770±0.0021	0.2778±0.0021	0.2846±0.0013	0.2863±0.0011
	3	0.2788±0.0043	0.2756±0.0049	0.2827±0.0022	0.2839±0.0022	0.2830±0.0012	0.2838±0.0018
	4	0.2682±0.0054	0.2841±0.0049				
	5	0.2721±0.0046	0.2756±0.0045				
	6	0.2804±0.0039	0.2906±0.0042			$f(1)=0.2834$ (exact)	
0.60	1	0.3616±0.0038	0.3652±0.0036	0.3568±0.0022	0.3583±0.0018	0.3554±0.0014	0.3561±0.0012
	2	0.3606±0.0036	0.3560±0.0038	0.3533±0.0017	0.3582±0.0020	0.3562±0.0009	0.3574±0.0010
	3	0.3620±0.0033	0.3607±0.0037	0.3564±0.0015	0.3543±0.0020	0.3562±0.0009	0.3556±0.0009
	4	0.3500±0.0040	0.3571±0.0036				
	5	0.3570±0.0037	0.3597±0.0037				
	6	0.3506±0.0038	0.3504±0.0041			$f(1)=0.3570$ (exact)	
0.70	1	0.3976±0.0035	0.4176±0.0038	0.3964±0.0035	0.4022±0.0016	0.4001±0.0035	0.4076±0.0011
	2	0.4122±0.0035	0.4118±0.0033	0.4055±0.0020	0.4062±0.0017	0.4068±0.0009	0.4078±0.0009
	3	0.4135±0.0028	0.4020±0.0034	0.4037±0.0015	0.4049±0.0016	0.4027±0.0008	0.4048±0.0010
	4	0.3918±0.0034	0.3924±0.0033				
	5	0.4099±0.0036	0.4172±0.0036				
	6	0.4088±0.0035	0.3998±0.0042			$f(1)=0.4056$ (exact)	
0.80	1	0.4330±0.0041	0.4356±0.0033	0.4301±0.0048	0.4512±0.0016	0.4116±0.0070	0.4413±0.0008
	2	0.4376±0.0031	0.4312±0.0030	0.4424±0.0015	0.4475±0.0016	0.4425±0.0010	0.4421±0·0008
	3	0.4472±0.0030	0.4331±0.0036	0.4426±0.0015	0.4367±0.0018	0.4441±0.0007	0.4371±0.0009
	4	0.4500±0.0037	0.4424±0.0033				
	5	0.4485±0.0028	0.4493±0.0029				
	6	0.4345±0.0029	0.4498±0.0030			$f(1)=0.4432$ (exact)	

chain, the sequence of values of $f_i(1)$ was broken into groups of 128 and the average and standard deviation computed for each of the groups: the sequence of averages thus generated will be denoted by $f(1,1)$, $f(1,2)$, $f(1,3)\cdots$.

The computations were performed for three different lattice sizes: 10 sites on an edge, 20 sites on an edge, and 37 sites on an edge. The computing time to complete one iteration was approximately 2 sec for the 10×10 lattice and this time varies linearly with the number of sites in the lattice.

In order to observe the effect of the choice of the initial configuration on these results, all of the computations were performed twice, using quite different starting conditions for the two cases. In one case the initial configuration was one of complete order with all spins up; that is, $\mu(k)=1$ for all k. In the other case the initial configuration was highly disordered; this configuration was generated by assigning the values of $\mu(k)$ such that there would be equal probability for up and down spins.

The results of the calculation of $f(1,j)$ are shown in

Table I. From the exact treatment of the square Ising lattice it is known that there is a second order phase transition, located at $x=0.4142$, at which the configurational specific heat becomes logarithmically infinite. Since the specific heat is proportional to the variance of the energy

$$d\langle E\rangle/dT \propto \langle[E-\langle E\rangle]^2\rangle,$$

it is not surprising that the strongest fluctuations and largest errors in the results occur in the neighborhood of $x=0.43$. It is interesting to note that standard deviations vary proportionally to $1/N^{\frac{1}{2}}$ as is to be expected for purely statistical reasons. It is satisfying to find that a prohibitively large lattice, and the 37×37 lattice approaches this, is not really needed for a good estimate of $f(1)$. In fact the difference in accuracy between the two larger systems can hardly be called significant. Hence, it appears that beyond the 20×20 lattice a large increase in N, and consequently a large increase in computing time, would be needed for a small increase in accuracy. It also appears that the effects of the initial configuration are lost rather quickly and in fact the average taken over the first sample of 128 configurations is frequently in good agreement with the exact value: this is particularly true for the system which started from a completely ordered configuration. It should be pointed out here that the very first configuration, the completely ordered one or the disordered one, is not included in the averaging.

IV. SIMPLE CUBIC LATTICE

The ILLIAC program which was prepared to do the computations on the simple cubic lattice is more elaborate than the one which was prepared for the square lattice computations. An important part of the present program is a test to determine a point in the sequence of configurations at which the computation of the ensemble averages is to commence. This test, called the convergence test, resembles a similar test used in the work described in footnote reference 1, and its present application is explained in the following.

The problem of obtaining a reasonably accurate result without using enormous amounts of computing time depends partly, as mentioned earlier, on the rate of convergence of the generated ensemble to a Boltzmann distribution. The work on the square lattice has shown that in many cases the convergence is quite rapid. With the square lattice the exact solution provides a guide for checking the ensemble averages, but when the exact solution is not known a rule must be made for selecting the point at which the averaging may commence. The problem of constructing such a rule is tricky. One might consider the successive values of a parameter, such as the short-range order parameter, and commence averaging where this sequence appears to be steady in some sense. This is dangerous for there may be two, or more, sets of states each of which has a relatively high probability of occuring, but which are linked by small transition probabilities. Such a situation would result in fairly steady sequential values for certain parameters except that here and there the apparent equilibrium value of the parameter might change abruptly. It may happen that these abrupt changes occur so infrequently that they would not even occur in a very long (on the computational time scale) computation. In this instance an average taken over the apparently steady sequence might give a very incorrect result. Since the region in phase space which contributes significantly to the ensemble average becomes broader near a phase transition it is to be expected that this phenomenon is likely to occur in such a region. It is true that no such difficulties were ever apparent in the two-dimensional lattice computations but similar difficulties have been encountered in the hard-sphere equation of state computations.

The rule which has been adopted for the present computations is characterized by the fact that two statistically independent sets of results are developed and the point of convergence is established when these results agree within a given margin of error. Two distinct lattices are used to develop two statistically independent sequences of configurations. One of the sequences starts from a configuration of complete order, while the other starts from a disordered configuration, just like the two initial configurations discussed in the square lattice computations. The difference between the present case and the former is that now the two sequences are developed simultaneously and hence may be compared, one with the other, as the two sequences are developed. We denote the sequence starting from a configuration of complete order as the low-temperature (LT) sequence. Correspondingly, the sequence starting from a disordered configuration is called the high-temperature (HT) sequence. The generation of the ensemble is divided into three stages. In the first stage an LT sequence of M_0 configurations and an HT sequence of M_0 configurations are developed. Three parameters associated with each configuration are held in the computer store: $f_i(1)$, $f_i(2)$, and S_i. In the second stage the average of $f_i(1)$ taken over the last M_0 configurations is computed for each sequence and similarly for $f_i(2)$: these are designated $[f(1)]_{LT}$, $[f(1)]_{HT}$, $[f(2)]_{LT}$, and $[f(2)]_{HT}$. The convergence test is passed when the following two inequalities are satisfied for the first time:

$$\frac{|[f(1)]_{LT}-[f(1)]_{HT}|}{[f(1)]_{LT}+[f(1)]_{HT}}<\epsilon_1, \tag{8a}$$

$$\frac{|[f(2)]_{LT}-[f(2)]_{HT}|}{[f(2)]_{LT}+[f(2)]_{HT}}<\epsilon_2, \tag{8b}$$

where ϵ_1 and ϵ_2 are small positive numbers. If both conditions are not satisfied a new configuration is generated for the LT sequence, and for the HT sequence. The oldest information in the sequence, namely, that

TABLE II. Summary of results from the simple cubic lattice. The superscript "a" on the *ID* number indicates the large (16×16×16) lattice.

ID	*K*	*L*	*S*	*f*(1)	*f*(2)	M_0	ΔM	Total
1	0.5000	0	0.9919±0.0004	0.0079±0.0003	0.0080±0.0003	50	50	156
2	0.4500	0	0.9823±0.0009	0.0169±0.0007	0.0174±0.0008	25	25	55
3	0.4000	0	0.9699±0.0014	0.0280±0.0010	0.0293±0.0011	25	25	62
4	0.3380	0	0.9424±0.0011	0.0516±0.0006	0.0549±0.0007	100	300	403
5	0.2857	0	0.8653±0.0060	0.1073±0.0021	0.1181±0.0026	50	50	102
6	0.2500	0	0.7440±0.0025	0.1853±0.0011	0.2096±0.0013	100	300	400
7	0.2381	0	0.6551±0.0060	0.2286±0.0024	0.2623±0.0029	50	50	161
8[a]	0.2381	0	0.6486±0.0024	0.2313±0.0010	0.2652±0.0013	50	50	100
9	0.2273	0	0.4976±0.0090	0.2841±0.0024	0.3315±0.0031	50	50	116
10	0.2174	0	0.3054±0.0123	0.3301±0.0026	0.3891±0.0033	50	50	100
11[a]	0.2174	0	0.1883±0.0091	0.3405±0.0016	0.4013±0.0020	50	50	100
12	0.2000	0	0.1603±0.0040	0.3713±0.0007	0.4344±0.0008	100	300	400
13[a]	0.2000	0	0.0781±0.0062	0.3708±0.0013	0.4341±0.0017	50	50	100
14	0.1667	0	0.0819±0.0022	0.4047±0.0005	0.4639±0.0005	100	300	400
15	0.1430	0	0.0686±0.0018	0.4211±0.0004	0.4755±0.0003	100	300	400
16	0.1250	0	0.0563±0.0030	0.4329±0.0007	0.4818±0.0006	50	50	100
17	0.2500	0.0625	0.8184±0.0032	0.1449±0.0017	0.1599±0.0020	50	50	102
18	0.2273	0.0625	0.7354±0.0044	0.1958±0.0021	0.2192±0.0025	50	50	101
19	0.2000	0.0625	0.5529±0.0062	0.2847±0.0024	0.3237±0.0028	50	50	100
20	0.1667	0.0625	0.3284±0.0071	0.3678±0.0018	0.4179±0.0021	50	50	100
21	0.1250	0.0625	0.1834±0.0052	0.4203±0.0009	0.4620±0.0008	50	50	100
22	0.0500	0.0625	0.1025±0.0086	0.4709±0.0009	0.4946±0.0007	25	25	50
23	−0.1000	0.0625	0.0680±0.0112	0.5525±0.0008	0.4893±0.0010	25	25	50
24	0.338	0.1250	0.9601±0.0010	0.0372±0.0007	0.0388±0.0008	50	50	128
25	0.2857	0.1250	0.9158±0.0018	0.0746±0.0012	0.0797±0.0014	50	50	121
26	0.2500	0.1250	0.8683±0.0023	0.1116±0.0014	0.1210±0.0016	50	50	105
27	0.2381	0.1250	0.8383±0.0022	0.1340±0.0013	0.1459±0.0015	50	50	105
28	0.2273	0.1250	0.8094±0.0025	0.1533±0.0014	0.1681±0.0017	50	50	105
29	0.2174	0.1250	0.7678±0.0054	0.1788±0.0024	0.1972±0.0028	50	50	100
30	0.2000	0.1250	0.6917±0.0053	0.2249±0.0023	0.2502±0.0028	50	50	100
31	0.1667	0.1250	0.5303±0.0050	0.3084±0.0019	0.3445±0.0022	50	50	100
32	0.1250	0.1250	0.3415±0.0045	0.3880±0.0013	0.4286±0.0014	50	50	100
33[a]	0.1250	0.1250	0.3369±0.0019	0.3892±0.0006	0.4297±0.0007	50	50	100
34	0.0750	0.1250	0.2145±0.0060	0.4424±0.0010	0.4722±0.0010	25	25	50
35	−0.1000	0.1250	0.0920±0.0096	0.5505±0.0009	0.4871±0.0011	25	25	50
36	−0.2000	0.1250	0.0613±0.0023	0.6266±0.0023	0.4342±0.0031	25	25	50
37	0.2857	0.2500	0.9390±0.0013	0.0562±0.0009	0.0587±0.0010	50	50	109
38	0.2381	0.2500	0.8966±0.0018	0.0909±0.0012	0.0969±0.0013	50	50	103
39	0.2000	0.2500	0.8127±0.0024	0.1548±0.0014	0.1670±0.0016	50	50	102
40	0.1250	0.2500	0.5655±0.0034	0.3082±0.0015	0.3341±0.0016	50	50	100
41	0.0750	0.2500	0.3965±0.0053	0.3938±0.0016	0.4175±0.0016	25	25	50
42	0.0625	0.2500	0.3678±0.0050	0.4091±0.0013	0.4303±0.0013	25	25	50
43	0.0250	0.2500	0.2896±0.0054	0.4486±0.0012	0.4586±0.0011	25	25	50
44	−0.0500	0.2500	0.1964±0.0038	0.5056±0.0007	0.4798±0.0006	50	50	100
45	−0.1000	0.2500	0.1566±0.0038	0.5394±0.0007	0.4785±0.0007	50	50	100
46	−0.1500	0.2500	0.1264±0.0040	0.5729±0.0008	0.4674±0.0008	50	50	100
47	−0.2000	0.2500	0.1030±0.0037	0.6171±0.0013	0.4366±0.0015	50	50	100
48	−0.2250	0.2500	0.0829±0.0025	0.6673±0.0034	0.3840±0.0041	25	25	61
49	−0.2500	0.2500	0.0564±0.0014	0.7960±0.0024	0.2274±0.0029	50	50	122
50	−0.2750	0.2500	0.0423±0.0019	0.8655±0.0023	0.1460±0.0026	25	25	67
51	−0.3000	0.2500	0.0297±0.0017	0.9033±0.0020	0.1040±0.0023	25	25	81
52	−0.3500	0.2500	0.0173±0.0007	0.9504±0.0010	0.0519±0.0011	50	50	106
53	−0.5000	0.2500	0.0064±0.0006	0.9876±0.0006	0.0126±0.0006	25	25	79
54	0.1875	0.7500	0.9473±0.0014	0.0498±0.0010	0.0513±0.0011	25	25	62
55	0.1000	0.7500	0.8424±0.0026	0.1405±0.0016	0.1452±0.0017	25	25	52
56	0.0500	0.7500	0.7450±0.0030	0.2165±0.0019	0.2216±0.0019	25	25	51
57	−0.0250	0.7500	0.5793±0.0041	0.3381±0.0015	0.3324±0.0016	25	25	50
58	−0.1000	0.7500	0.4350±0.0058	0.4377±0.0017	0.4001±0.0013	25	25	50
59	−0.1500	0.7500	0.3624±0.0064	0.4915±0.0018	0.4206±0.0015	25	25	50
60	−0.1875	0.7500	0.3143±0.0059	0.5331±0.0021	0.4223±0.0016	25	25	50
61	−0.2500	0.7500	0.2199±0.0029	0.6561±0.0037	0.3416±0.0038	25	25	55
62	−0.3000	0.7500	0.1157±0.0021	0.8374±0.0028	0.1613±0.0030	25	25	89
63	−0.3750	0.7500	0.0469±0.0014	0.9408±0.0012	0.0589±0.0012	25	25	67
64	0.338	1.250	0.9921±0.0004	0.0078±0.0003	0.0079±0.0003	50	50	101
65	0.2000	1.250	0.9834±0.0007	0.0163±0.0005	0.0165±0.0005	50	50	101
66	0.0500	1.250	0.9053±0.0014	0.0902±0.0019	0.0911±0.0014	50	50	100
67	−0.0500	1.250	0.7670±0.0020	0.2103±0.0012	0.2065±0.0012	50	50	100
68	−0.1000	1.250	0.6778±0.0023	0.2840±0.0011	0.2697±0.0011	50	50	100
69	−0.1500	1.250	0.5852±0.0027	0.3578±0.0011	0.3239±0.0009	50	50	100
70	−0.2000	1.250	0.5016±0.0033	0.4261±0.0012	0.3598±0.0008	50	50	100
71	−0.2500	1.250	0.4221±0.0035	0.4960±0.0013	0.3754±0.0008	50	50	100

TABLE II.—*Continued.*

ID	*K*	*L*	*S*	*f*(1)	*f*(2)	M_0	ΔM	Total
72	−0.2750	1.250	0.3786±0.0061	0.5384±0.0027	0.3679±0.0021	25	25	50
73	−0.3000	1.250	0.3131±0.0022	0.6216±0.0026	0.3169±0.0023	50	50	129
74	−0.3250	1.250	0.2282±0.0033	0.7358±0.0032	0.2280±0.0028	25	25	52
75	−0.3500	1.250	0.1690±0.0020	0.8088±0.0020	0.1711±0.0018	50	50	116
76	−0.4000	1.250	0.0961±0.0020	0.8940±0.0017	0.0993±0.0015	25	25	58
77	−0.5000	1.250	0.0298±0.0007	0.9686±0.0006	0.0307±0.0006	50	50	134
78	0.2000	0.4000	0.8834±0.0022	0.1035±0.0015	0.1091±0.0017	25	25	64
79	0.1750	0.3500	0.8171±0.0033	0.1543±0.0018	0.1645±0.0021	25	25	79
80	0.1500	0.3000	0.7121±0.0047	0.2251±0.0025	0.2430±0.0028	25	25	95
81	−0.0500	0.1000	0.1080±0.0123	0.5233±0.0008	0.4964±0.0009	25	25	50
82	−0.1000	0.2000	0.1364±0.0086	0.5443±0.0011	0.4821±0.0012	25	25	50
83	−0.1500	0.3000	0.1544±0.0076	0.5687±0.0013	0.4675±0.0014	25	25	50
84	−0.2000	0.4000	0.1604±0.0052	0.6035±0.0018	0.4341±0.0015	25	25	51
85	−0.2500	0.5000	0.1260±0.0025	0.7397±0.0040	0.2806±0.0049	25	25	60
86	−0.3000	0.6000	0.0863±0.0017	0.8622±0.0021	0.1404±0.0023	25	25	87
87	−0.0500	0.2000	0.1679±0.0084	0.5133±0.0012	0.4871±0.0011	25	25	50
88	−0.1000	0.4000	0.2468±0.0067	0.5182±0.0016	0.4629±0.0013	25	25	50
89	−0.1500	0.6000	0.2915±0.0059	0.5245±0.0018	0.4410±0.0013	25	25	50
90	−0.2000	0.8000	0.3204±0.0055	0.5362±0.0020	0.4168±0.0015	25	25	50
91	−0.2500	1.0000	0.3260±0.0055	0.5665±0.0024	0.3823±0.0017	25	25	50
92	−0.3000	1.2000	0.2889±0.0030	0.6466±0.0039	0.3046±0.0036	25	25	111
93	−0.0500	0.4000	0.3023±0.0052	0.4759±0.0014	0.4536±0.0013	25	25	50
94	−0.1000	0.8000	0.4638±0.0051	0.4226±0.0015	0.3886±0.0012	25	25	50
95	−0.1500	1.2000	0.5635±0.0051	0.3736±0.0018	0.3357±0.0014	25	25	50
96	−0.2000	1.6000	0.6321±0.0042	0.3313±0.0016	0.2958±0.0012	25	25	50
97	−0.2500	2.0000	0.6898±0.0038	0.2906±0.0015	0.2574±0.0014	25	25	50
98	−0.3000	2.4000	0.7282±0.0033	0.2595±0.0014	0.2314±0.0012	25	25	50

pertaining to the configuration which occurred M_0 iterations before the present point, is thrown out to make space for the information on the new configuration. Thus, the order parameters for the last M_0 configurations in each chain are retained. As each new configuration is added to the *LT* sequence and to the *HT* sequence the convergence test is repeated. When the convergence test is finally passed the third stage of the computations begins. All of the values for $f_i(1)$ in the two sequences are collected into one sum, $\sum f_i(1)$, and similarly for $f_i(2)$, and S_i. As each new configuration is generated in each sequence the values for $f_i(1)$, $f_i(2)$, and S_i are added to the corresponding sum. When ΔM new configurations in each sequence have been generated then the sums are divided by $2(M_0+\Delta M)=R$, the number of configurations in the ensemble, to obtain the final estimates of the quantities $f(1)$, $f(2)$, and S. In the third stage of the computations the sum of squares of each of the parameters is also generated in order to calculate the standard deviations.

This process does not ensure against the possibility of obtaining an erroneous answer because of the chains being trapped in a set of metastable states. It can be expected, however, that the chance of detecting such a situation is better than it would be if only one sequence was considered. Of course, if memory space and computing time permit, then one can extend this method to include a still larger number of independent chains.

The model for almost all of the computations had 8 sites on an edge, and thus contained a total of 512 sites. Some computations were done on a model with 16 sites on an edge but because of the large amounts of computing time required this work was rather limited. The amount of time required to complete one iteration for the 8×8×8 array was about 6 sec and it was approximately eight times this for the 16×16×16 array. As with the square lattice, periodic boundary conditions were always imposed. The sites were selected systematically for detailed consideration in an analogous fashion to the method used with the square lattice: successive sites in a row were treated, then successive rows and finally successive planes.

The results of this computation are compiled in Table II. Order parameters S, $f(1)$, and $f(2)$ are shown as functions of K in Figs. 1, 2, and 3, respectively. In Table II the numbers in the first column are simply identification numbers (*ID*). The next two columns contain the energy parameters K and L. In the last three columns the numbers M_0 and ΔM appearing

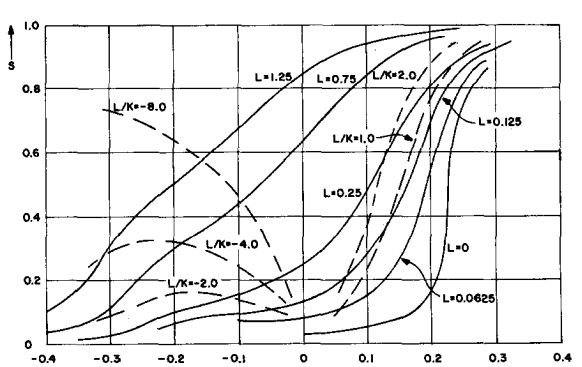

FIG. 1. Long-range order S in the simple cubic lattice shown as a function of K for different L (solid curves) and L/K (dashed curves).

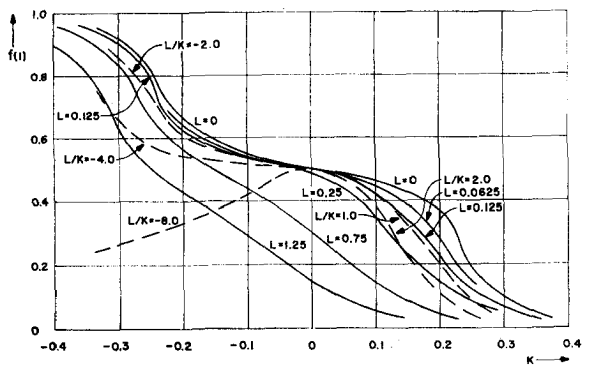

FIG. 2. First-neighbor short-range order parameter $f(1)$ in the simple cubic lattice shown as a function of K for different L (solid curves) and L/K (dashed curves).

there have already been defined and the numbers in the column headed "total" are the total number of iterations performed. Hence, in row 1, the values $M_0=50$, $\Delta M=50$, total=156 mean that 50 configurations were generated in the LT sequence and 50 configurations were generated in the HT sequence before the convergence tests began (since $M_0=50$); next, after convergence 50 more configurations were generated in each sequence (since $\Delta M=50$) to make a total of 200 configurations in the ensemble; since the total number of iterations was 156 it follows that the convergence test was passed upon completion of the one hundred and sixth iteration and therefore the first 56 configurations in each sequence were discarded. In the columns following K and L the order parameters S, $f(1)$, and $f(2)$ are presented. The spread indicated for each order parameter is the standard deviation of the mean. The results are listed in order of decreasing K (i.e., increasing temperature for fixed J) in groups in which L is fixed. Near the end of the table, starting at $ID=78$, some results are grouped together in which L/K is fixed: note that $L/K=H/J$, the ratio of the external field coupling energy to the nearest-neighbor coupling energy, which is independent of T. In the figures the results are plotted for fixed L (solid lines) and for fixed L/K (dashed lines). Although the order parameters can be computed analytically at $K=0$, the curves for constant L/K have not been extended through $K=0$ because computations were not performed in the region of $K=0$ and an extrapolation of these curves from the available data did not seem appropriate.

It will be noted that three different values for M_0 appear in the table: 100, 50, and 25. In the very first (chronologically) computations the large value of M_0 was used together with $\Delta M=300$, but because of the large amounts of computing time being absorbed it was decided to set M_0 and ΔM both at 50. Still later, for reasons of economizing on computing time the still smaller values $M_0=25$ and $\Delta M=25$ were introduced. In the first two cases the convergence test parameters were given the value 0.02: $\epsilon_1=\epsilon_2=0.02$. In the last case, in an attempt to compensate for the relatively small value of $M_0=25$ the parameters were given the value 0.01: $\epsilon_1=\epsilon_2=0.01$.

Since the size effect can be expected to be most significant in the zero field case in the neighborhood of the apparent critical temperature, some computations for a $16\times16\times16$ array were made in this region: these have identification numbers 8, 11, and 13. When these are compared against the corresponding results for the $8\times8\times8$ system, it will be noted that there is a marked difference in the value of S for the two cases. The differences in the results obtained for the short-range order parameters on the other hand are relatively slight. Hence, it appears that although the estimate of of S is a crude approximation of its value for the infinite system in this region, the estimates of $f(1)$ and $f(2)$ are rather good. In the case of $L=0.125$ a computation was made with the large lattice ($ID=33$) in the region where S can be expected to be most sensitive to size effects. Comparison of these results with those for the $8\times8\times8$ lattice shows that the difference in the results obtained for S, as well as for $f(1)$ and $f(2)$, is slight. Hence, for this value and higher values of L the estimates of S can be expected to be fairly a good approximation to the value for an infinite system.

In the antiferromagnetic region the value $L/K=-6$ is a critical one. For L/K greater in magnitude than this value the external field coupling dominates the nearest-neighbor coupling and at low temperatures the system tends to the state in which all spins are parallel to the external field. For L/K smaller in magnitude than this value, the nearest-neighbor coupling dominates, and at low temperatures the system tends to the state in which all nearest-neighbor spins are antiparallel. The series of computations at $L/K=-4$ and $L/K=-8$ illustrate the alternate behavior in the order parameters as the parameter K tends to large negative values (i.e., as $T\rightarrow0$ for J equal to a negative constant). It is interesting to notice that a relatively large number of iterations had to be performed in the computation with $ID=92$ before the convergence con-

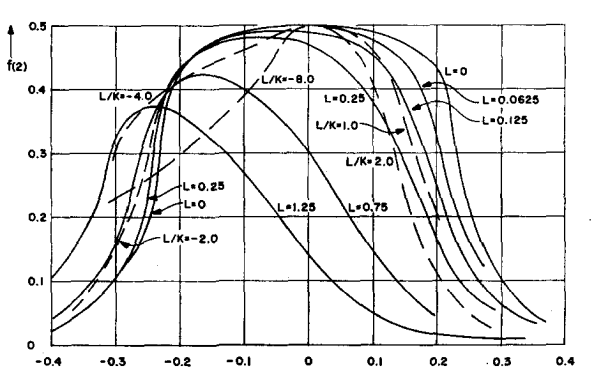

FIG. 3. Second-neighbor short-range order parameter $f(2)$ in the simple cubic lattice shown as a function of K for different L (solid curves) and L/K (dashed curves).

TABLE III. Comparison of results obtained from the Monte Carlo calculation with results obtained from evaluation of series expressions (footnote reference 3) for the simple cubic lattice.

	S		$f(1)$	
$K(L=0)$	Monte Carlo	Series	Monte Carlo	Series
0.5000	0.9919	0.9945	0.0079	0.0054
0.4000	0.9699	0.9795	0.0280	0.0195
0.3380	0.9424	0.9504	0.0516	0.0453
0.2000	...	...	0.3713	0.3756
0.1665	...	...	0.4048	0.4050
0.1430	...	...	0.4210	0.4217

dition was satisfied; in computations 85 and 86 a similar situation is noticed. The reason for this is that we are near the critical magnitude of K while below the critical magnitude of L/K. In the region $K<0$, $-6K>L\geq 0$ there are two states of minimum energy (the two states in which all neighbors are antiparallel), just as there are on the line $K>0$, $L=0$. Thus, there will again be a critical value of K, around which configurations consisting of mixtures of these two states will tend to persist. As the line $L/K=-6$ is approached, a third configuration also assumes importance, that in which all spins are parallel to the external field. This explains why computation 92 ($L/K=-4$) is even slower in convergence than computations 85 and 86 ($L/K=-2$). It is significant that the existence of this situation is strongly brought to one's attention because of the rules which have been set up for convergence testing. In a test based on the examination of one sequence of configurations there is a greater chance that one would fail to observe this near-critical situation since the sequence might remain entirely in one set of configurations during the computations. Furthermore, it should be noticed that the standard deviations in these cases do not indicate anything unusual. One can infer from the small standard deviations that once the convergence conditions were satisfied both the sequences remained in the one class of states which was most probable.

For regions in which series expansions of the long-range order and short-order can be used a comparison with results obtained from the present Monte Carlo method is possible. The series given in footnote reference 3[11] have been evaluated for a few cases and a comparison with the Monte Carlo results is shown in Table III.

V. BODY-CENTERED CUBIC LATTICE

The computations for this lattice are not as extensive as those for the simple cubic lattice. The convergence test is the same as the one used with the simple cubic lattice except that since only S_i and $f_i(1)$ are calculated during each iteration, just the first inequality (8a) is required to be satisfied. The value of ϵ_1 was set at 0.02. In the third stage of the generation of the configuration sequence (i.e., after the convergence condition was satisfied) the results for the two independent sequences were not combined but instead they were left separate. Thus, two separate estimates of the average were computed, one based on the HT sequence and the other based on the LT sequence. The LT sequence for positive K was started with an initial configuration in which all of the spins were parallel to one another and parallel to the external field. When K was negative the initial configuration for the LT sequence was one in which all nearest-neighbors were antiparallel. In the simple cubic lattice computations the LT sequence always started with a configuration in which all spins were parallel to each other and to the external field, even when K was negative.

In most of the calculations a model consisting of eight unit cells on an edge, and therefore 1024 sites, was used. In a few cases a smaller lattice having four unit cells on an edge was used. Twenty seconds was required to complete one iteration on the larger lattice. This program was not quite as efficient in its use of computer time as the simple cubic lattice program.

The sites were selected for detailed consideration in a systematic fashion. The two sublattices of the system were processed separately: that is, all the "center" sites were first processed sequentially, then all "corner" sites were processed sequentially to complete one iteration for one lattice.

The results are compiled in Table IV and displayed graphically in Figs. 4 and 5. In Table IV the upper value for S and for $f(1)$ is obtained from the LT sequence and the lower value for S and for $f(1)$ is obtained from the HT sequence. In the initial calculations the larger samples with $M_0=50$ and $\Delta M=50$ were computed, but to conserve on computer time this was later reduced

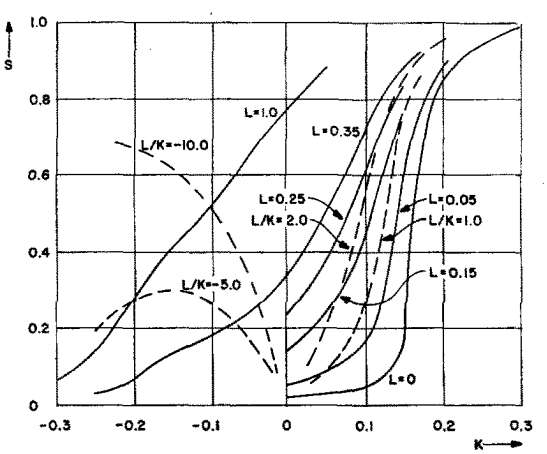

FIG. 4. Long-range order S in the body-centered cubic lattice shown as a function of K for different L (solid curves) and L/K (dashed curves).

[11] It should be pointed out that the series expressions given in Eqs. (7.5)-(7.7) of this paper are incorrect. Each series is given in the form $z^{-\alpha}$ (polynomial in z) whereas it should be $-\alpha \ln z$ + (polynomial in z).

TABLE IV. Summary of results from the body-centered cubic lattice. The superscript "a" on the *ID* number indicates the small (four unit cells on an edge) lattice.

ID	*K*	*L*	*S*	*f*(1)	M_0	ΔM	Total
1	0.3	0	0.9804 ± 0.0007	0.0189 ± 0.0007	50	50	101
			0.9800 ± 0.0018	0.0188 ± 0.0012			
2[a]	0.225	0	0.8966 ± 0.0060	0.0908 ± 0.0045	50	50	100
			0.8931 ± 0.0129	0.0866 ± 0.0064			
3	0.225	0	0.9067 ± 0.0019	0.0831 ± 0.0015	50	50	103
			0.9032 ± 0.0032	0.0847 ± 0.0016			
4	0.1875	0	0.8023 ± 0.0028	0.1607 ± 0.0018	50	50	101
			0.7850 ± 0.0095	0.1662 ± 0.0034			
5[a]	0.175	0	0.7127 ± 0.0109	0.2173 ± 0.0063	50	50	100
			0.6717 ± 0.0202	0.2251 ± 0.0088			
6	0.175	0	0.6884 ± 0.0050	0.2251 ± 0.0027	50	50	109
			0.6697 ± 0.0102	0.2321 ± 0.0035			
7	0.1625	0	0.5216 ± 0.0091	0.2981 ± 0.0033	50	50	100
			0.5246 ± 0.0103	0.2982 ± 0.0032			
8	0.15	0	0.2159 ± 0.0132	0.3734 ± 0.0027	50	50	101
			0.1602 ± 0.0101	0.3850 ± 0.0017			
9	0.1375	0	0.0938 ± 0.0095	0.4030 ± 0.0023	50	50	100
			0.1091 ± 0.0071	0.4099 ± 0.0013			
10	0.125	0	0.0784 ± 0.0071	0.4179 ± 0.0017	50	50	100
			0.0684 ± 0.0051	0.4223 ± 0.0013			
11[a]	0.1	0	0.1366 ± 0.0103	0.4430 ± 0.0030	50	50	100
			0.1433 ± 0.0107	0.4376 ± 0.0029			
12	0.1	0	0.0441 ± 0.0038	0.4404 ± 0.0012	50	50	100
			0.0484 ± 0.0036	0.4430 ± 0.0010			
13[a]	0.05	0	0.1031 ± 0.0098	0.4635 ± 0.0036	50	50	100
			0.1169 ± 0.0074	0.4714 ± 0.0022			
14	0.05	0	0.0453 ± 0.0062	0.4684 ± 0.0026	50	50	101
			0.0318 ± 0.0024	0.4767 ± 0.0018			
15	0.2	0.05	0.8809 ± 0.0029	0.1044 ± 0.0023	20	30	58
			0.8823 ± 0.0031	0.1030 ± 0.0024			
16	0.1625	0.05	0.7136 ± 0.0056	0.2190 ± 0.0032	20	30	56
			0.7137 ± 0.0052	0.2185 ± 0.0030			
17	0.15	0.05	0.6190 ± 0.0093	0.2715 ± 0.0042	20	30	50
			0.6018 ± 0.0073	0.2769 ± 0.0036			
18	0.125	0.05	0.3135 ± 0.0102	0.3869 ± 0.0027	20	30	50
			0.3165 ± 0.0130	0.3853 ± 0.0039			
19	0.1	0.05	0.1530 ± 0.0076	0.4351 ± 0.0014	20	30	50
			0.1516 ± 0.0094	0.4330 ± 0.0023			
20	0.15	0.15	0.7657 ± 0.0047	0.1906 ± 0.0031	20	30	51
			0.7663 ± 0.0047	0.1902 ± 0.0028			
21	0.125	0.15	0.6051 ± 0.0062	0.2857 ± 0.0034	20	30	53
			0.6113 ± 0.0063	0.2836 ± 0.0030			
22	0.1	0.15	0.4388 ± 0.0077	0.3652 ± 0.0031	20	30	50
			0.4441 ± 0.0086	0.3632 ± 0.0033			
23	0.05	0.15	0.2338 ± 0.0064	0.4517 ± 0.0018	20	30	50
			0.2349 ± 0.0066	0.4472 ± 0.0016			
24	0.15	0.25	0.8371 ± 0.0034	0.1416 ± 0.0025	20	30	51
			0.8338 ± 0.0034	0.1430 ± 0.0024			
25	0.125	0.25	0.7361 ± 0.0047	0.2130 ± 0.0032	20	30	50
			0.7379 ± 0.0042	0.2119 ± 0.0027			
26	0.1	0.25	0.6084 ± 0.0052	0.2905 ± 0.0029	20	30	50
			0.6056 ± 0.0053	0.2946 ± 0.0030			
27	0.05	0.25	0.3789 ± 0.0055	0.4089 ± 0.0023	20	30	50
			0.3712 ± 0.0069	0.4108 ± 0.0024			
28	0.175	0.35	0.9222 ± 0.0037	0.0720 ± 0.0030	20	30	50
			0.9198 ± 0.0020	0.0741 ± 0.0018			
29	0.15	0.35	0.8718 ± 0.0033	0.1143 ± 0.0026	20	30	50
			0.8795 ± 0.0027	0.1081 ± 0.0023			
30	0.125	0.35	0.8135 ± 0.0034	0.1596 ± 0.0026	20	30	50
			0.8076 ± 0.0040	0.1649 ± 0.0029			
31	0.1	0.35	0.7105 ± 0.0042	0.2329 ± 0.0027	20	30	50
			0.7164 ± 0.0040	0.2293 ± 0.0026			
32	0.05	0.35	0.4970 ± 0.0051	0.3621 ± 0.0025	20	30	50
			0.4943 ± 0.0049	0.3613 ± 0.0024			
33	−0.05	0.35	0.2451 ± 0.0060	0.4922 ± 0.0023	20	30	50
			0.2463 ± 0.0085	0.4916 ± 0.0018			
34	−0.1	0.35	0.1849 ± 0.0050	0.5304 ± 0.0020	20	30	50
			0.1870 ± 0.0059	0.5361 ± 0.0017			
35	−0.15	0.35	0.1387 ± 0.0033	0.5954 ± 0.0021	20	30	51
			0.1360 ± 0.0028	0.6010 ± 0.0034			
36	−0.2	0.35	0.0681 ± 0.0024	0.8300 ± 0.0050	20	30	63
			0.0646 ± 0.0021	0.8501 ± 0.0028			
37	−0.25	0.35	0.0296 ± 0.0011	0.9386 ± 0.0047	20	30	51
			0.0268 ± 0.0015	0.9461 ± 0.0017			
38	−0.05	0.5	0.3431 ± 0.0057	0.4591 ± 0.0025	20	30	50
			0.3467 ± 0.0065	0.4601 ± 0.0157			
39	−0.15	0.75	0.3060 ± 0.0054	0.5278 ± 0.0032	20	30	50
			0.3044 ± 0.0036	0.5321 ± 0.0167			
40	0.05	1.0	0.8729 ± 0.0025	0.1174 ± 0.0021	20	30	55
			0.8725 ± 0.0025	0.1180 ± 0.0022			
41	−0.1	1.0	0.5127 ± 0.0067	0.3947 ± 0.0044	20	30	50
			0.5133 ± 0.0088	0.3945 ± 0.0063			
42	−0.15	1.0	0.4107 ± 0.0056	0.4715 ± 0.0036	20	30	50
			0.4119 ± 0.0074	0.4738 ± 0.0044			
43	−0.2	1.0	0.2704 ± 0.0035	0.6472 ± 0.0055	20	30	56
			0.2734 ± 0.0028	0.6472 ± 0.0037			
44	−0.25	1.0	0.1299 ± 0.0038	0.8472 ± 0.0054	20	30	52
			0.1222 ± 0.0025	0.8604 ± 0.0028			
45	−0.3	1.0	0.0655 ± 0.0066	0.9247 ± 0.0091	20	30	50
			0.0548 ± 0.0016	0.9409 ± 0.0016			
46	−0.25	1.25	0.2003 ± 0.0078	0.7744 ± 0.0091	20	30	50
			0.1921 ± 0.0025	0.7877 ± 0.0028			
47	−0.2	2.0	0.6734 ± 0.0057	0.2991 ± 0.0051	20	30	50
			0.6728 ± 0.0064	0.3002 ± 0.0055			

to $M_0=20$ and $\Delta M=30$. In all cases the convergence parameter had the value $\epsilon_1=0.02$.

Comparison of the results for the smaller lattice, having four unit cells on an edge, with those for the larger lattice shows again the relatively strong dependence of the estimate of long-range order on lattice size. On the other hand the short-range order estimate again is not so sensitive to change in lattice size. Comparison of results for the LT sequences with those for the HT sequences indicates that although the differences between the pairs of results are not large, they are frequently larger than the spread indicated by the standard deviations. This is not surprising, since the correlation that exists between the sequential configurations tends to reduce the standard deviation from what it would be if the configurations in the chain were completely independent.

It is well known[3] that the Ising model is equivalent to a model of a binary substitutional alloy and from this viewpoint one can compare the present results with the results of the measurement of long-range order in

TABLE V. Comparison of results obtained from the Monte Carlo calculation with results obtained from evaluation of series expressions (footnote reference 3) for the body-centered cubic lattice.

	S		$f(1)$	
$K(L=0)$	Monte Carlo	Series	Monte Carlo	Series
0.300	0.9804 0.9800	0.9806	0.0189 0.0188	0.0188
0.150	...	...	0.3734 0.3850	0.3955
0.050	...	...	0.4684 0.4767	0.4742

β-brass made by Chipman and Warren.[12] This comparison appears in Fig. 6. The results of the Bethe[13] second approximation, according to Chipman and Warren, are also shown. In this figure the results have been normalized so as to make T_c coincide with the experimentally observed value of 465°C. For this normalization we set $K_c=0.1616$; this choice is discussed in the next section. The Monte Carlo results seem to fall closer to the experimental results than do the results of the Bethe second approximation at low temperatures. However, the relative position of the Monte Carlo curve depends on the normalization and the apparent deviation of the Monte Carlo curve from the Bethe curve and the experimental curve must be viewed with this in mind. Experimental measurements of the short-range order which may be compared with the Monte Carlo results do not seem to be available.

VI. ESTIMATION OF THE CRITICAL POINT

The finite size of the model precludes the existence of a true critical point. Nevertheless the model does

[12] D. Chipman and B. Warren, J. Appl. Phys. **21**, 696 (1950).
[13] H. A. Bethe, Proc. Roy. Soc. (London) **A150**, 552 (1935).

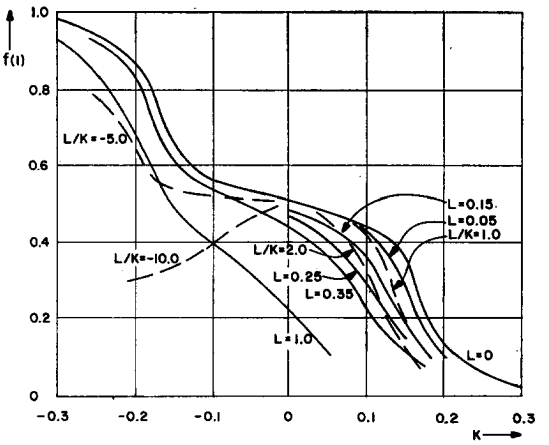

FIG. 5. First-neighbor short-range order parameter $f(1)$ in the body-centered cubic lattice shown as a function of K for different L (solid curves) and L/K (dashed curves).

exhibit a behavior in a fairly narrow range of K which is similar to that associated with a critical point or "Curie point" in these systems: the long-range order in the ferromagnetic case falls off sharply to very small values and the specific heat, as indicated by the standard deviation of the short-range order, appears to go through a maximum. It would be desirable to obtain from this information on a finite lattice an estimate of the critical point K_c. One procedure for obtaining such an estimate would be to record the values of K at which the specific heat becomes maximal as the size N of the system is increased. Treating these results as a function of $1/N$ and extrapolating to $1/N=0$ would provide the desired estimate. Unfortunately this procedure demands such enormous amounts of computing time that it does not seem to be practical at the present time. Since it is convenient to have some definition of the apparent critical point simply so that it may be discussed without ambiguity we use here the value of K at which $S=\frac{1}{2}$; the value of K defined in this fashion is denoted by $K_{\frac{1}{2}}$. This parameter has the virtue that it can be computed relatively accurately from a small number of calculations of S in the neighborhood of the apparent critical point

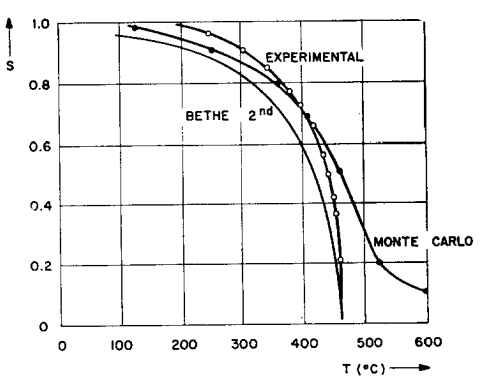

FIG. 6. Comparison of long-range order with experimental results on β-brass by Chipman and Warren, and with the results of the Bethe second approximation.

and there is no need to make the disagreeable subjective decision which attends extrapolation of the long-range order above K_c to zero, or extrapolation of the inverse susceptibility below K_c to zero. In addition to this it seems reasonable to expect that $K_{\frac{1}{2}}$ will be close to K_c for the following two reasons. The abrupt vanishing of the long-range order at K_c which one expects to find in three-dimensional lattices on the basis of the exact results for the infinite two-dimensional square lattice implies that $K_{\frac{1}{2}}$ for an infinite three-dimensional lattice will lie very close to K_c. Yang's[9] exact calculation of the long-range order in an infinite two-dimensional lattice yields a value for the fractional error $(K_{\frac{1}{2}}-K_c)/K_c$ of about 1.6×2^{-9}. The tighter coupling of the three-dimensional lattice should make this error still smaller. The second reason is that the portion of the long-range order curve in the neighborhood of $S=\frac{1}{2}$ is, relative to the "tail" of this curve, fairly insensitive to changes in lattice size and it is therefore expected that $K_{\frac{1}{2}}$ for the finite lattice of the model used here lies close to $K_{\frac{1}{2}}$ and hence close to K_c for an infinite lattice. Unfortunately a quantitative estimate of this is lacking, but the qualitative behavior of our results indicates that this conjecture is reasonable. For the simple cubic lattice, performing a linear interpolation between $ID=7$ and $ID=9$ of Table II, we find

$$K_{\frac{1}{2}}=0.2275 \quad \text{(simple cubic)}.$$

For the body-centered cubic lattice, performing a linear interpolation between $ID=7$ and $ID=8$ of Table IV (averaging the upper and lower values of S first), we find

$$K_{\frac{1}{2}}=0.1616 \quad \text{(body-centered cubic)}.$$

This figure has been identified as K_c for the normalization of the long-range order curve of Fig. 6 which was introduced in the preceding section.

VII. CONCLUSION

The Monte Carlo method appears to be well suited to the computation of short-range order in an Ising lattice. In the two-dimensional lattice, where the results can be checked against the exact treatment, the accuracy of the Monte Carlo results for a 20×20 array is quite good. With one exception, which occurs in the immediate neighborhood of the critical temperature, the errors are 5% or less. In the three-dimensional systems, arrays of 512 sites for the simple cubic and 1024 sites for the body-centered cubic provided estimates of the short-range order which were relatively insensitive to changes in the size of the array even in the neighborhood of the critical temperature. The long-range order, on the other hand, was found to be rather sensitive to changes in size of the array near the critical temperature. The long-range order results in this region are therefore rather crude. An accurate estimate of long-range order very near the critical temperature by the present method does not seem to be feasible with present computing equipment, because the large arrays that seem to be required demand unreasonably long computation times.

As pointed out by Newell and Montroll,[3] the primary reason for the continued interest in the Ising model is that it provides a simple testing ground for new approximate methods of investigating systems of interacting particles. The present results, except perhaps the long-range order in the zero field case near K_c, are felt to be the most accurate ones now available and they have been tabulated here in considerable detail so that they may be readily compared with the results of other methods. Computations using a simple cubic lattice with second- and third-neighbor interactions are now in progress.

Erratum

Linearized Plasma Oscillations in Arbitrary Electron Distributions

[J. Math. Phys. 1, 178 (1960)]

GEORGE E. BACKUS
Massachusetts Institute of Technology, Cambridge, Massachusetts

Dr. F. Meyer of the Max-Planck-Institut für Physik und Astrophysik in München has kindly pointed out that the proof of theorem 3 is incorrect. The exponents in the denominators of Eqs. (27) and (28) should be $\frac{1}{2}$ instead of 1, and an upper rather than a lower bound on $|\mathcal{L}(s)|$ is needed.

A correct proof can be given if $g_0'(u)$ is bounded and integrable and satisfies a Hölder condition. These hypotheses, although stronger than those stated for theorem 3 in the paper, are satisfied by the Maxwell distribution; and they permit the application of Muskelishvili's results, thus implying that there are positive numbers y_0 and M such that $|\mathcal{L}(s)|<M$ if $0<\Im s<y_0$. Then in the inequality following (29), α must be replaced by M.

With (27) and (28) rendered valueless, w is so far unspecified. We choose it so that $g_0'(w)\neq 0$. Then, since $g_0'(u)$ is continuous, there are positive numbers ϵ and δ such that $|g_0'(u)|>\epsilon$ if $|u-w|<\delta$. Hence

$$\int_{-\infty}^{\infty}\left|\frac{g_0'(u)}{u-s}\right|du \geq \epsilon\int_{w-\delta}^{w+\delta}\frac{du}{|u-w+iy|}=2\epsilon\{\ln[\delta+(\delta^2+y^2)^{\frac{1}{2}}]-\ln y\}.$$

Thus inequality (30) must be rewritten as

$$My|\ln y|^{\frac{1}{2}}|K(s)| \geq 2\epsilon\{\ln[\delta+(\delta^2+y^2)^{\frac{1}{2}}]-\ln y\}.$$

This inequality contradicts $|K(s)|y\leq m$ and proves theorem 3.