JOURNAL OF
MATHEMATICAL
PHYSICS

VOLUME 7, NUMBER 9 SEPTEMBER 1966

Lattice Constant Systems and Graph Theory*

M. F. SYKES
Wheatstone Physics Laboratory, King's College, England
J. W. ESSAM
Westfield College, London, England
B. R. HEAP
Mathematics Division, National Physical Laboratory, Teddington, England
AND
B. J. HILEY
Birkbeck College, London, England
(Received 13 July 1965)

The two principal systems of lattice constants that have arisen in the study of cooperative phenomena and related problems on crystal lattices are the *strong* (low-temperature) and the *weak* (high-temperature) systems. The two systems are defined in terms of the concepts of graph theory, and a general theorem relevant to cluster expansions is stated. The interrelation of the two systems is studied and exploited to derive configurational data for the face-centered cubic lattice. All star graphs with up to seven points (vertices) or nine lines (edges) that are embeddable on the face-centered cubic are described. A general classification of stars with cyclomatic index 3 is given.

1. INTRODUCTION

THE last few years have seen a most promising development in the calculation of exact series expansions needed for the investigation of the Ising and Heisenberg models of a ferromagnet and antiferromagnet and more generally for the statistical mechanics of interacting systems on crystal lattices. Considerable effort has also been devoted to closed form approximations and to improved systematic techniques of extrapolation. It is now recognized that reliable information on the physical properties of three-dimensional systems can be obtained from such studies, and an extensive literature exists. For a bibliography, reference should be made to the reviews by Domb,[1] Fisher,[2] and Helfand.[3]

With few exceptions the more successful methods of extracting reliable information about the critical region depend on the provision of increasing amounts of structural data for the lattice studied. It is found in practice that, as seems reasonable on physical grounds, the most satisfactory sequences of approximants are obtained by grouping the configurational data in a reasonably unbiased way. Essentially, most methods depend implicitly, and sometimes explicitly, on the enumeration and counting of the multiply connected graphs that can occur on the lattice studied. The number of graph types is not finite, and successive approximants correspond to the grouping of the data by the number of lines or the number of points in each graph.

The enumeration of all the possible multiply connected graphs (or stars) of a given number of lines or points is a pure mathematical problem of importance in the Mayer theory of gas condensation and has been much studied, notably by Uhlenbeck

* This research has been supported (in part) by the U. S. Department of the Army through its European Research Office.

[1] C. Domb, Advan. Phys. 9, Nos. 34 and 35 (1960).
[2] M. E. Fisher, J. Math. Phys. 4, 278 (1963).
[3] E. Helfand, Ann. Rev. Phys. Chem. 14, 117 (1963).

and Ford.[4] For most problems connected with interacting systems on a lattice, the problem takes on a new character as at an early stage a large number of the theoretically possible topologies are usually excluded by the structural limitations of the lattice. For each new approximation, or series coefficient, the number of graphs increases rapidly, but far less rapidly than for the most general possible lattice, which would admit all graphs, and to which the graph theoretical enumeration problem properly applies. Progress in obtaining high-order approximation has been relatively slow, largely because of the difficulty of supplying the configurational data; this is particularly the case for closely packed lattices such as the face-centered cubic.

As a natural consequence of different studies of the Ising model of a ferromagnet two different systems of configurational data or *lattice constants* have arisen. The *low*-temperature, or *strong*, system occurs in the derivation of expansions applicable to the ordered state below the transition temperature and corresponds to a systematic enumeration of disordered states in the ordered assembly. The *high*-temperature, or *weak*, system occurs in the derivation of expansions applicable above the transition temperature and corresponds to the systematic enumeration of ordered states in the disordered assembly.

The construction of lattice constant tables of useful proportions for the more usual crystal lattices is of fundamental importance and such data are applicable to a wide range of problems. In particular, we quote the elucidation of the critical properties of the specific heat, magnetization, and susceptibility of the Ising[5-9] and Heisenberg[10,11] models of a ferromagnet and antiferromagnet, the lattice gas problem,[2] the percolation problem,[12,13] cluster size,[14] the dilute magnet,[15-17] and the excluded volume problem.[18-20]

To provide lattice constant tables a variety of specialized techniques have been developed and many of these are described by Domb.[1] Increasing use is being made of electronic computers.[21-25] Recent improvements in cluster techniques[11,17,26] presage the study of more ambitious lattice model problems all depending on a knowledge of the same essential configurational data.

In the present paper we define the concepts of weak and strong lattice constants in the language of graph theory and state precisely some general properties of the two systems which are of particular interest in physical applications. The first of these, implicit in the literature, is the existence of the so-called cluster expansions (for quantities satisfying the extensive relation) in terms of the connected constants only. The second, again implicit in the literature, is the interrelation of the two systems which we exploit to obtain configurational data on the fcc lattice.

We show that the two systems can be related by a matrix, some of whose properties we study. The actual counting of the individual graphs on the more important crystal lattices has been studied simultaneously by Martin,[27] and extensive tables of data will be published elsewhere. The present paper is partly intended as an introduction to this publication. In particular, we discuss briefly the problem of classification and show that the interrelation of the systems can be used to effect an economy of effort. We subsequently apply the general results to the theory of percolation.[28]

2. GENERAL CONCEPTS

We employ as far as possible the graph definitions of Ore[29] with only slight additions dictated by the

[4] G. E. Uhlenbeck and G. W. Ford, *Studies in Statistical Mechanics*, J. de Boer and G. E. Uhlenbeck, Eds. (North-Holland Publishing Company, Amsterdam, 1962), Vol. 1, p. 123. See also W. G. Hoover and A. G. de Rocco, J. Chem. Phys. **36**, 3141 (1962).
[5] A. J. Wakefield, Proc. Cambridge Phil. Soc. **47**, 419 (1950); *ibid.* **47**, 799 (1951).
[6] C. Domb and B. J. Hiley, Proc. Roy. Soc. (London) **A268**, 506 (1962).
[7] C. Domb and M. F. Sykes, Proc. Roy. Soc. (London) **A235**, 247 (1956).
[8] M. E. Fisher and M. F. Sykes, Physica **28**, 939 (1962).
[9] G. A. Baker, Phys. Rev. **124**, 768 (1961).
[10] C. Domb and M. F. Sykes, Phys. Rev. **128**, 168 (1962).
[11] C. Domb and D. W. Wood, Phys. Letters **8**, 20 (1964).
[12] M. F. Sykes and J. W. Essam, Phys. Rev. **133**, A310 (1964).
[13] M. F. Sykes and J. W. Essam, J. Math. Phys. **5**, 1117 (1964).
[14] C. Domb and M. F. Sykes, Phys. Rev. **122**, 77 (1961).
[15] B. R. Heap, Proc. Phys. Soc. (London) **82**, 252 (1963).
[16] D. J. Morgan and G. S. Rushbrooke, Mol. Phys. **6**, 477 (1963).
[17] G. S. Rushbrooke, J. Math. Phys. **5**, 1106 (1964).
[18] M. E. Fisher and M. F. Sykes, Phys. Rev. **114**, 45 (1959).
[19] B. J. Hiley and M. F. Sykes, J. Chem. Phys. **34**, 1531 (1961).
[20] M. F. Sykes, J. Math. Phys. **2**, 52 (1961).
[21] G. S. Rushbrooke and J. Eve, J. Chem. Phys. **31**, 1333 (1959).
[22] G. S. Rushbrooke and J. Eve, J. Math. Phys. **3**, 185 (1962).
[23] J. L. Martin, Proc. Cambridge Phil. Soc. **58**, 92 (1962).
[24] C. Domb, J. Chem. Phys. **38**, 2957 (1963).
[25] C. Domb, I. Gillis, and G. Wilmers, Proc. Phys. Soc. (London) **85**, 625 (1965).
[26] See in particular Ref. 1, Secs. 5.2.10 and 5.3.2.
[27] J. L. Martin (private communication). Work in progress at the Mathematics Division of the National Physical Laboratory, Teddington, England.
[28] J. W. Essam and M. F. Sykes, J. Math. Phys. **7**, 1573 (1966).
[29] O. Ore, *Theory of Graphs*, American Mathematical Society Colloquium Publications (American Mathematical Society, Providence, Rhode Island, 1962), Vol. XXXVIII.

FIG. 1. An undirected linear graph.

special emphasis of our studies. For a detailed treatment of the fundamental concepts, reference should be made to Ore[29] and to Berge[30] and for the topological background to Patterson[31] and Kelley[32] and particularly Veblen.[33]

We illustrate in Fig. 1 an *undirected linear graph.* It has six *vertices* which constitute the vertex set V, four *edges*, and three connected *components*. The graph is the abstract concept represented by the vertex set and the associated pairings of the vertices represented by the edges. The number of edges having a given vertex as end point is the *degree* of that vertex. In our example there are three vertices of degree two, two of degree one, and one of degree zero.

For a connected graph C, a vertex is called an *articulation point* or *cut-point* (or cut-vertex) if the graph obtained by deleting it, and all its incident edges, is not connected. A connected graph with two or more vertices and no cut-points is called a *star*. A connected graph other than a single vertex, or a single edge, with no cut-points is called *multiply connected* (see Fig. 2). Multiple connectivity may be defined in other ways, and for proofs of the equivalence of various definitions, reference should be made to the works cited.

A set of vertices V' is called an *articulation* or *cut-set of vertices* of C if the graph obtained by deleting V', and all its incident edges, is not connected.

An edge is called a *cut-edge* or *cut-bond* of C if the graph obtained by deleting it is not connected.

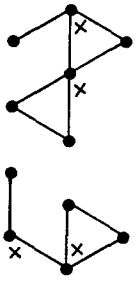


FIG. 2. Examples of graphs with cut-points. Cut-points are marked X.

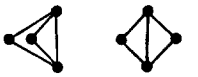
[30] C. Berge, *Théorie des graphes et ses applications* (Dunod, Paris, 1958) (English edition: Methuen, London, 1962). The latter translates *sous-graphe* as *sub-graph* instead of *section graph* and *sous-graphe partiel* as *partial sub-graph* instead of *sub-graph*.

[31] E. M. Patterson, *Topology* (Oliver and Boyd, London, 1963).

[32] J. L. Kelley, *General Topology* (D. Van Nostrand Company, Inc., Princeton, New Jersey, 1955).

[33] O. Veblen, *Analysis Situs*, American Mathematical Society Colloquium Publications (American Mathematical Society, Providence, Rhode Island, 1931), Vol. V, Pt. 2.

FIG. 3. Example of two isomorphic graphs.



The end points of the edge are not deleted, and therefore the graph consisting of a single edge with its end points has a cut-edge. A *cut-set of edges* is a set of edges whose deletion (without deletion of their end points) from C leaves a graph which is not connected.

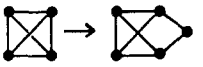
Two graphs G and G' are said to be isomorphic when there is a one–one correspondence between their vertex sets V and V' such that corresponding vertices are joined by edges in one of them only if they are joined in the other (Ore)[29]; see Fig. 3.

For many applications a graph is thought of as a network and the vertices as junctions. Vertices of degree greater than 2 are called the *nodes or principal points*; those of degree 2 or 1 the *antinodes*.

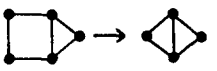
If G is any graph and A some vertex of degree 2, then we define the *suppression* of the vertex A to be the deletion of A from the vertex set V of G and the identification of the two edges incident upon A. The reverse process of replacing an edge incident upon two vertices B and C by two edges (one incident upon B, one upon C, and both upon a new vertex A) is called the *insertion* of a second-order vertex on the edge BC.

If G is any graph and G' a graph derived from G by the insertion or suppression of any number of vertices of degree 2, then G and G' are said to be *homeomorphic* and each is a *homeomorph* of the other. Figure 4 gives examples.

Homeomorphs have the same basic topology and the concept of homeomorphism is of value in many physical applications. For example, the number of self-avoiding walks on a lattice may be expressed



Suppression of vertex of degree 2



The graphs  and  are homeomorphic.

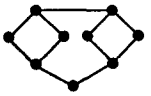
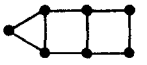
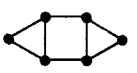
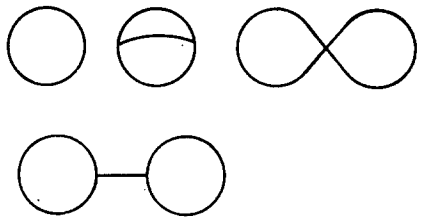
 and  are homeomorphs of 

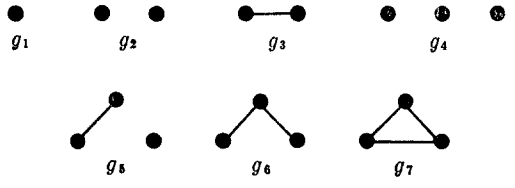
FIG. 4. Examples. Insertion of vertex of degree 2.

in terms of all the homeomorphs of the four topological types.[20]



It is convenient to assign reference symbols to the various types of linear graph and we use g_i and order the subscripts as follows.

We denote the number of vertices and edges in g_i by $v(g_i)$, $l(g_i)$, respectively, and abbreviate in general to v_i, l_i. We order the subscripts so that g_i occurs before g_j if $v_i < v_j$; or if $v_i = v_j$ then if $l_i < l_j$. Thus the sequence begins



(2.1)

If $v_i = v_j$ and $l_i = l_j$ of which the first example is the pair  and  , we suppose that the suffixes have been arbitrarily assigned.

In a similar manner, we assign symbols c_i, s_i to the various types of connected and star graphs, respectively, arranged in the same graph dictionary order. In general, we abbreviate $v(c_i)$, $l(c_i)$, $v(s_i)$, $l(s_i)$ to v_i, l_i whenever it is clear from the context which graph dictionary is being used. Again, when using other symbols, G for example, we usually abbreviate $v(G)$, $l(G)$ to v_G, l_G, and likewise with other quantities whenever the meaning is clear.

We denote the number of connected components in g_i by $n(g_i)$ or n_i. The *cyclomatic number* $C(g_i)$ of a graph is defined to be

$$C(g_i) = l_i - v_i + n_i. \tag{2.2}$$

Evidently this is invariant under the suppression or insertion of vertices of degree 2, and homeomorphic graphs have the same cyclomatic number.

If g_r and g_s are two graphs with disjoint vertex sets and disjoint edge sets, we denote by $g_r \cup g_s$ the graph whose vertex and edge sets are the unions of the respective sets of g_r and g_s. Thus, for example, $g_5 = g_1 \cup g_3$ and $g_4 = g_1 \cup g_1 \cup g_1 = g_1 \cup g_2$.

A graph H is a *subgraph* of G when the vertex set $V(H)$ is contained in the vertex set $V(G)$ and all the edges of H are edges of G. If A is a subset of the vertex set of G, the *section graph* $G(A)$ is defined as the subgraph whose vertex set is A and whose edges are *all* those edges of G which connect two vertices in A (Ore).[29]

A *partial graph* P of G is a subgraph of G whose vertex set is the vertex set of G.

Two subgraphs are *vertex-disjoint* if they have no vertices, hence no edges, in common. Two subgraphs are *edge-disjoint* if they have no edges in common. If H_1 and H_2 are two subgraphs of G we define their *sum graph* $H = H_1 + H_2$ to be the subgraph formed from all the vertices and edges of H_1 and H_2 or both.

Any subgraph G' of a graph G which is isomorphic with a graph g is said to represent an *embedding* of g in G in the *weak* sense (*weak embedding*). Any section graph G^* of G which is isomorphic with g is said to represent an embedding of g in G in the *strong* sense (*strong embedding*). Evidently a strong embedding is also a weak embedding, but the converse statement is not necessarily true.

Any weak embedding of g in G defines a subset V' of the vertices of G which are also vertices of g embedded in G. We call the section graph which has V' as its vertex set the *associated section graph* of the embedding.

Suppose that a graph G is the sum graph of two or more of its subgraphs, for example, of the three subgraphs H_1, H_2, H_3 isomorphic with some g_i, g_j, g_k of (2.1). Then

$$G = H_1 + H_2 + H_3. \tag{2.3}$$

Now H_1, H_2, H_3 are weak embeddings of g_i, g_j, g_k in G, and we call (2.3) an *overlap partition* of G into g_i, g_j, g_k. In general, there will be more than one possible choice of embeddings of g_i, g_j, g_k in G having G as its sum graph, and we define the total number of such choices as the number of overlap partitions of G into g_i, g_j, g_k and write this number

$$\{g_i + g_j + g_k = G\}. \tag{2.4}$$

For example, there are three overlap partitions of the triangle g_7 into g_3 and g_6 and three into g_5 and g_6 or

$$\begin{aligned} \{g_3 + g_6 &= g_7\} = 3, \\ \{g_5 + g_6 &= g_7\} = 3. \end{aligned} \tag{2.5}$$

We call these quantities overlap partitions because in general the component graphs do overlap and have some, or all, of their edges and vertices in common.

The *lattice constant* of a graph g on a graph G is defined for both weak and strong embeddings as follows:

(1) Weak sense (Weak or high temperature lattice constant) $(g; G)$ = Number of subgraphs of G isomorphic with g.
(2) Strong sense (Strong or low-temperature lattice constant) $[g; G]$ = Number of section graphs of G isomorphic with g.

We follow the convention of Domb in representing the weak or high-temperature condition by round brackets and the strong or low-temperature condition by square brackets. $(g; G)$ is the number of weak embeddings of g in G, $[g; G]$ the number of strong embeddings. For any graph G with v vertices and l edges we have two sets of lattice constants:

(1) The weak set $(g_i; G)$ for all g_i with $l_i \leq l$, $v_i \leq v$.
(2) The strong set $[g_i; G]$ for all g_i with $l_i \leq l$, $v_i \leq v$.

Obviously any constant for g_i with $l_i > 1$ and or $v_i > v$ must be zero, and by excluding these the set of lattice constants as defined is a finite set. For example, the lattice constants of the triangle g_7 are

$$(g_1; g_7) = 3 \qquad [g_1; g_7] = 3,$$
$$(g_2; g_7) = 3 \qquad [g_2; g_7] = 0,$$
$$(g_3; g_7) = 3 \qquad [g_3; g_7] = 3,$$
$$(g_4; g_7) = 1 \qquad [g_4; g_7] = 0,$$
$$(g_5; g_7) = 3 \qquad [g_5; g_7] = 0,$$
$$(g_6; g_7) = 3 \qquad [g_6; g_7] = 0,$$
$$(g_7; g_7) = 1 \qquad [g_7; g_7] = 1.$$

3. GENERAL THEOREMS

We now state a result which relates the strong and weak set of lattice constants for any graph G.

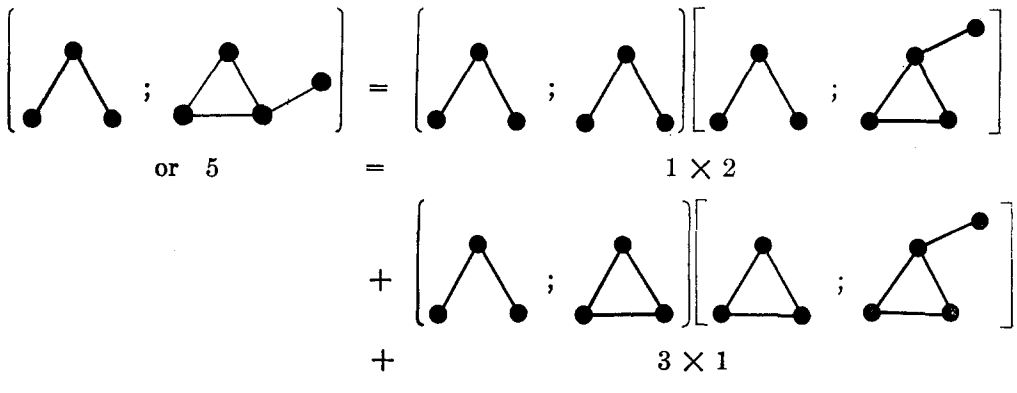
Theorem I: If $g_i^{(r)}$ is a graph with r vertices and G any graph

$$(g_i^{(r)}; G) = \sum_j (g_i^{(r)}; g_j^{(r)})[g_j^{(r)}; G], \tag{3.1}$$

where the summation is taken over all graphs $g_j^{(r)}$ with r vertices.

Proof: Any strong embedding of $g_i^{(r)}$ in G will be the associated section graph of $(g_i^{(r)}; g_j^{(r)})$ weak embeddings of $g_i^{(r)}$ in G. By definition the number of these strong embeddings is $[g_j^{(r)}; G]$ and since every embedding of $g_i^{(r)}$ has one, and only one, associated section graph, the result follows.

We illustrate the fact that any weak constant can be expressed as a linear sum of strong constants with the same number of vertices by an example.



(3.2)

Since (3.1) is true for any graph it is convenient to condense the notation and write

$$(g_i; G) = p_i, \quad [g_i; G] = P_i, \quad (g_i; g_j) = a_{ij}, \tag{3.3}$$

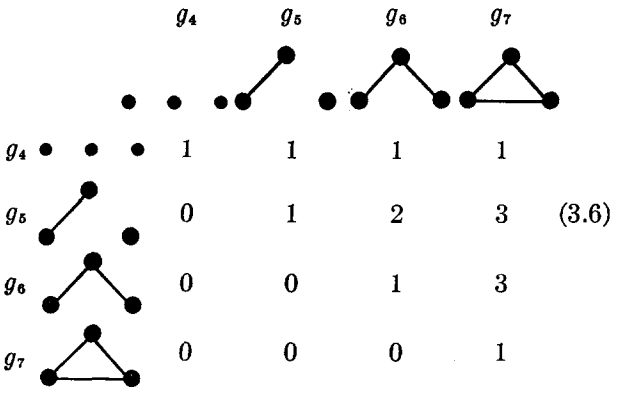
and in this notation (3.1) becomes

$$p_i = \sum_j a_{ij} P_j, \qquad v_i = v_j. \tag{3.4}$$

If we group together all equations of type (3.4) which apply to graphs with r vertices they may be written in matrix notation as

$$\mathbf{p}(r) = \mathbf{A}(r)\mathbf{P}(r), \tag{3.5}$$

where $\mathbf{p}(r)$ and $\mathbf{P}(r)$ are column vectors the elements of which are the weak and strong constants, respectively, of all graphs with r vertices and $\mathbf{A}(r)$ is a square matrix which we call the rth-order *conversion matrix*. As an example we illustrate the third-order conversion matrix.



	g_4	g_5	g_6	g_7
g_4	1	1	1	1
g_5	0	1	2	3
g_6	0	0	1	3
g_7	0	0	0	1

(3.6)

The conversion matrix is triangular with the lower triangle zero since we have ordered the graphs by the number of their edges and

$$(g_i; g_j) = 0, \qquad l_i > l_j. \tag{3.7}$$

Further, all the diagonal elements are unity, and therefore $\mathbf{A}(r)$ is nonsingular and can be inverted recursively to yield a reciprocal conversion matrix $\mathbf{B}(r)$. The elements of $\mathbf{B}(r)$ can be written

$$b_{ij} = (-1)^{l_i - l_j} a_{ij} \tag{3.8}$$

a result due to Martin.[34]

To prove (3.8) we construct further matrices $\mathbf{C}(r, k); k \geq 1$ where

$$\begin{aligned} c_{ij}(r, k) &= a_{ij}(r) \quad \text{if} \quad l_j = l_i + k \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{3.9}$$

Denote a graph g_i with v vertices and l edges by $g_i\{v, l\}$. Now, by removing one edge from a graph g_j with $l + k$ edges and $k - 1$ edges from all the resulting graphs, we find

$$\begin{aligned} &(g_i\{r, l\}; g_j\{r, l + k\}) \\ &= \frac{1}{k} \sum_h (g_i\{r, l\}; g_h\{r, l + k - 1\}) \\ &\quad \times (g_h\{r, l + k - 1\}; g_j\{r, l + k\}), \end{aligned} \tag{3.10}$$

where the sum is over all graphs with $l + k - 1$ edges. By the definition (3.9) this can be rewritten

$$c_{ij}(r, k) = \frac{1}{k} \sum_h c_{ih}(r, k - 1) c_{hj}(r, 1). \tag{3.11}$$

In matrix notation (3.11) becomes

$$\mathbf{C}(r, k) = (1/k)\mathbf{C}(r, k - 1)C(r, 1), \tag{3.12}$$

and therefore

$$\mathbf{C}(r, k) = (1/k!)[C(r, 1)]^k. \tag{3.13}$$

[34] J. L. Martin (private communication).

Now

$$\begin{aligned} \mathbf{A}(r) &= \mathbf{I} + \sum_{k=1}^{\frac{1}{2}r(r-1)} \mathbf{C}(r, k) \\ &= \mathbf{I} + \sum_{k=1}^{\frac{1}{2}r(r-1)} [\mathbf{C}(r, 1)]^k/k! = \exp [\mathbf{C}(r, 1)], \end{aligned} \tag{3.14}$$

since

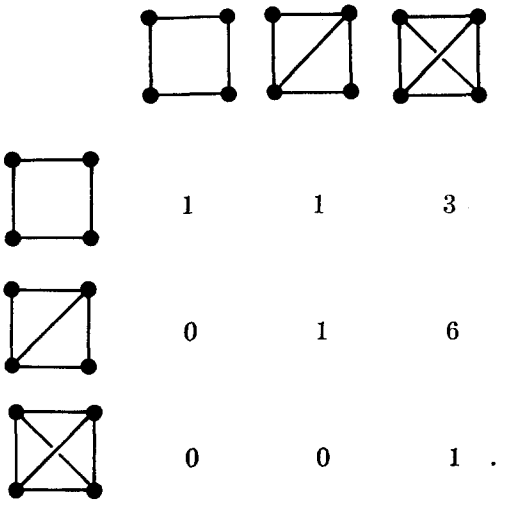
$$[\mathbf{C}(r, 1)]^k = 0 \quad \text{for} \quad k > \tfrac{1}{2}r(r - 1).$$

But

$$\begin{aligned} \mathbf{B}(r) &= [\mathbf{A}(r)]^{-1} = \exp [-\mathbf{C}(r, 1)] \\ &= \mathbf{I} + \sum_{k=1}^{\frac{1}{2}r(r-1)} (-1)^k \mathbf{C}(r, k) \end{aligned} \tag{3.15}$$

from which the result (3.8) follows immediately. The recurrence relation (3.12) may be used as a practical means of computation of $\mathbf{A}(r)$ given $\mathbf{C}(r, 1)$.

It follows by repeating the arguments of this section for the set of connected constants and further for the set of star constants that we may define corresponding nonsingular conversion matrices for the weak and strong sets of these quantities. We need only the additional observation that the associated section graph of any embedding of a connected graph must be a connected graph and that of a star must be a star. As an example we illustrate the fourth-order star conversion matrix.



(3.16)

We now state a theorem which we use to show that the lattice constant on a graph G for any graph with more than one component (separated lattice constant) can be calculated from the connected lattice constants of G.

Theorem II: If g_i and g_j are two graphs $g_i \neq g_j$ and G any graph then

$$(g_i \cup g_j; G) = (g_i; G)(g_j; G) - \sum_k \{g_i + g_j = g_k^{(r)}\}(g_k^{(r)}; G), \qquad (3.17)$$

where the summation is taken over all $g_k^{(r)}$ with $r < v_i + v_k$.

Proof: The total number of pairs of embeddings of g_i and g_j is $(g_i; G)(g_j; G)$ and these divide into two disjoint classes:

(1) Pairs for which the respective embeddings of g_i and g_j are vertex disjoint. The number of these is by definition $(g_i \cup g_j; G)$.

(2) Pairs for which the respective embeddings have at least one vertex in common. Then if $g_k^{(r)} = g_i + g_j$ is the sum graph of any pair of embeddings the number of vertices in g_k must be less than $v_i + v_j$. The number of pairs which have any given g_k as sum graph is simply the number of overlap partitions of g_k into g_i and g_j. Further every pair of embeddings defines some sum graph and therefore the number of pairs in this class is

$$\sum_k \{g_i + g_j = g_k^{(r)}\}(g_k^{(r)}; G), \qquad (3.18)$$

where the summation is taken over all $g_k^{(r)}$ with $r < v_i + v_j$.

If $g_i = g_j$ the result must be modified as the number of pairs is then $\frac{1}{2}(g_i; G)^2$.

The graphs g_i and g_j are not necessarily connected and by successive applications of Theorem II the number of embeddings of a many component graph may be expressed in terms of the embeddings of single component graphs. We therefore have the important result that the weak constant for a graph g_i with connected components c_r, c_s, c_t, $\cdots$ can be expressed as a polynomial, of degree equal to the number of components of g_i, in weak connected constants and that the resultant expression is *independent of* G.

Thus any (*weak*) separated lattice constant $(g_i; G)$ may be expressed in terms of the connected (*weak*) constants $(c_j; G)$ with $v_j \leq v_G$. It follows, by the existence and properties of the conversion matrices, that the statement is true with (*weak*) replaced by (*strong*) in either or both positions.

An important application of this result is to the development of cluster expansions for extensive quantities associated with a graph.

4. THE CONSTRUCTION OF CONVERSION MATRICES

We devote attention in this section to a description of methods employed in the construction of the rth-order conversion matrix $\mathbf{A}(r)$ defined in (3.3), (3.4), and (3.5), and illustrated for $r = 3$ in (3.6).

To begin with we note that provided the g_i are ordered as described in Sec. 2 then $\mathbf{A}(r)$ can be conveniently partitioned into block matrices $\mathbf{A}_{s,t}(r)$ $[0 \leq s, t \leq \frac{1}{2}r(r-1)]$ where if the element corresponding to $(g_i; g_j)$ lies within $\mathbf{A}_{s,t}(r)$ then besides $v_i = v_j$ we have $l_i = s$ and $l_j = t$. The matrix $\mathbf{A}_{s,t}(r)$ is thus of order $n(r, s) \times n(r, t)$, where $n(r, s)$ is the number of distinct graphs having r vertices and s edges. Also,

$$\mathbf{A}_{s,s}(r) = I_{n(r,s)}; \qquad \mathbf{A}_{s,t}(r) = 0 \quad \text{if} \quad t < s. \qquad (4.1)$$

It has already been mentioned in the proof of (3.8) that the relation (3.12) may be used as a practical means of computation of $\mathbf{A}(r)$. In terms of the $\mathbf{A}_{s,t}(r)$ this relation becomes

$$\mathbf{A}_{s,t}(r) = \mathbf{A}_{s,s+1}(r)\mathbf{A}_{s+1,t}(r)/(t-s) \qquad (4.2)$$

for $t > s + 1$. Thus to evaluate the $\mathbf{A}_{s,t}(r)$ and thereby ultimately the complete matrix, $\mathbf{A}(r)$, it is sufficient to obtain the submatrices $\mathbf{A}_{s,s+1}(r)$ $[0 \leq s \leq \frac{1}{2}r(r-1)]$.

Let us consider the evaluation of $(g_i; g_j)$, where $v_i = v_j = r$ and $l_j = l_i + 1$. This is equal to the number of ways of removing a single edge from the graph g_j to give the graph g_i. For small values of r this evaluation can be carried out straightforwardly by inspection once a complete table of graphs having r vertices has been drawn up. The conversion matrices for $r \leq 5$ are easily computed in this way, and the same general method can be used for the evaluation of the sixth-order matrix. In this latter case for large values of s, that is when the graphs have a large number of edges, the recognition of a graph is made simpler by considering its complement. However, when we consider the seventh-order conversion matrix, the total number of graphs involved, namely 1044, becomes too large for such methods and recourse has to be made to a computer.

There are two main problems in connection with the use of a computer to enumerate graphs, namely, the actual representation of a graph in the computer and the identification of a particular graph among a dictionary of possibly some hundreds of graphs. The most convenient method of representing a graph appears to be by means of its adjacency matrix. If the vertices of a graph having a total of v vertices are labeled 1, 2, $\cdots$, v, then the *adjacency matrix* of the graph is the $v \times v$ matrix $D = (d_{r,s})$ in which $d_{r,s} = 1$ if there is an edge connecting vertex r to vertex s and $d_{r,s} = 0$ otherwise. Since in this work

we are dealing with unlabeled graphs, a particular labeling must be assigned to each graph in order to be able to represent it in this manner. A convenient way of doing this is to assign the label 1 to the vertex having the highest degree, 2 to the one having the next highest, and so on. This labeling is not normally unique since if several vertices have the same degree, the assignment for these vertices is made arbitrarily.

When we wish to identify a particular graph, we compare its adjacency matrix with those of all graphs having the same set of degrees. Should the graph not be so identified, we interchange two rows and columns of the adjacency matrix (i.e., interchange the labels of two vertices of the graph). This is done according to a scheme which preserves the set of degrees as before and which ultimately runs through all possible labelings of the graph.[35] The process is then repeated until the graph is identified.

In order to obtain the list of all s-edged graphs which are obtained from a particular $(s+1)$-edged graph, it is only necessary to remove each edge of the graph in turn, possibly relabel the graph to conform with the representation scheme, and then apply the identification process.

By means of these techniques, the matrices $\mathbf{A}_{s,s+1}$ (7) $(0 \leq s \leq 20)$ were computed in approximately four hours using the Ace computer of the National Physical Laboratory. Full details of the program and other results concerning seven-point graphs are given elsewhere.[36] Once these matrices had been evaluated, the construction of the complete matrix $\mathbf{A}$ (7) followed using (4.2).

5. CLUSTER EXPANSIONS

Suppose that $\phi = \phi(G)$ is some quantity determined explicitly by the set of lattice constants of G, weak or strong, say $\Pi_i(G)$, and certain independent parameters. For example, we may take ϕ to be the configurational free energy L of the Ising model of a set of spins located on the vertices of G and interacting in pairs delineated by the edges of G. This is determined by G and the variable parameters defining the absolute temperature T and applied magnetic field H, together with constant parameters such as the exchange interaction J. A general formulation[1] in terms of configurations (subgraphs) can be given and the quantity ϕ will be

$$\phi = \phi(G) = L[\Pi_i(G); T, H]. \tag{5.1}$$

As a second example we take the mean number of clusters in a random mixture. Suppose the vertices of G are colored black with probability p and white with probability $1-p$. Then any realization of the probability distribution on G defines a linear graph R_B which is the section graph of G whose vertices are all the black vertices. Now G contains $[g_i; G]$ section graphs isomorphic with g_i, and the probability of any one of these being R_B is just

$$p^{v_i}(1-p)^{v_G - v_i}, \tag{5.2}$$

and if g_i has n_i connected components the mean number is defined to be

$$K = \langle n_i \rangle = \sum_{g_i} n_i p^{v_i}(1-p)^{v_G - v_i}[g_i; G], \tag{5.3}$$

and the quantity ϕ will be

$$\phi = \phi(G) = K(\Pi_i(G); p). \tag{5.4}$$

[We recall that in (5.2) $v_G = (g_1; G)$ is also a lattice constant of G].

Now by Theorem II if $\phi(G)$ is determined by the set of all lattice constants it will equally be determined by the set of all *connected* constants alone. We thus suppose generally that

$$\phi = \phi(G) = \Psi[\Pi_i(G); t], \tag{5.5}$$

where $\Pi_i(G)$ with $i = 1, 2, \cdots$ now denotes the set of *connected* lattice constants $(c_i; G)$ or $[c_i; G]$ listed in the conventional graph dictionary order of Sec. 2 and t denotes a set of parameters.

We now suppose that $\phi(G)$ has an *extensive property*, namely,

$$\phi(G \cup G') = \phi(G) + \phi(G'), \tag{5.6}$$

that is, the quantity ϕ for two graphs considered together is the sum of the quantities for the two graphs considered separately. The extensive property of the free energy is well known. For the mean number function it is obvious.

It is evident that each connected constant $\Pi_i(G)$ satisfies the extensive property since

$$\begin{aligned} &(c_i; G \cup G') = (c_i; G) + (c_i; G') \\ \text{and} \quad &[c_i; G \cup G'] = [c_i; G] + [c_i; G']. \end{aligned} \tag{5.7}$$

Consequently, we have

$$\Psi[\Pi_i(G) + \Pi_i(G'); t] = \Psi[\Pi_i(G); t] + \Psi[\Pi_i(G'); t]. \tag{5.8}$$

Now, we use this functional relation to show that $\Psi(\Pi_i; t)$ is a *linear* function of the nonnegative

[35] Details of a similar scheme which runs through all possible permutations of N objects by means of interchanges has been given by B. R. Heap, Computer J. 6, 293 (1963).

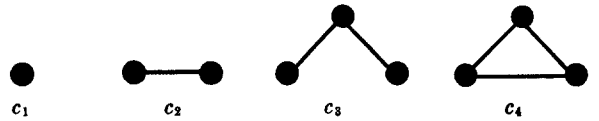
[36] B. R. Heap, National Physical Laboratory, Mathematics Division Report No. Ma. 64 (to be published).

integers Π_i. If the Π_i were independently allowed, all positive integral values this result would be trivial since by repeated application of (5.8) we find the unique solution

$$\begin{aligned}\Psi(\Pi_1, \Pi_2, \cdots ; t) &= \Psi(0, 0, 0, \cdots ; t) \\ &+ \Pi_1\Psi(1, 0, 0, \cdots ; t) \\ &+ \Pi_2\Psi(0, 1, 0, \cdots ; t) \\ &+ \Pi_3\Psi(0, 0, 1, \cdots ; t) + \cdots . \end{aligned} \tag{5.9}$$

But the lattice constants must satisfy certain inequalities since, for example, a graph with one edge must have at least two vertices and more generally the presence of any subgraph implies the presence of all the subgraphs of that subgraph.

To prove a result allowing for such restrictions, we consider the set of graphs all of whose lattice constants Π_j are zero for $j > k$ where k is some integer. For simplicity we present the proof for the case $k = 4$ when the relevant connected graphs are



(5.10)

We proceed by defining a new set of variables Π_j^* which may independently take all nonnegative integral values and are related linearly to the set Π_i. Consider the highest graph in the list, namely, c_4. Since it is not contained in any of the lower-order graphs, its lattice constant is unrestricted so we take

$$\Pi_4^* = \Pi_4. \tag{5.11}$$

Now consider the next highest graph in the list, namely, c_3. This lattice constant must be at least $(c_3; c_4)\Pi_4^*$ or $[c_3; c_4]\Pi_4^*$ in the weak or strong systems, respectively, since it is a subgraph of c_4. The number of independent occurrences of c_3 will therefore be

$$\Pi_3^* = \Pi_3 - (c_3; c_4)\Pi_4^* \tag{5.12}$$

in the weak system with a similar expression in the strong system. To simplify the manipulations and to treat both strong and weak systems together we write $(c_i; c_j)$ and $[c_i; c_j]$ as $\Delta_{i,j}$ it being understood that, as with the Π_i, weak or strong constants are chosen throughout. The result (5.12) is now

$$\Pi_3^* = \Pi_3 - \Delta_{3,4}\Pi_4^*. \tag{5.13}$$

Considering the next highest graph c_2, we see by the same argument that there must be at least $\Delta_{2,3}\Pi_3^* + \Delta_{2,4}\Pi_4^*$ occurrences. The independent occurrences will thus be

$$\Pi_2^* = \Pi_2 - \Delta_{2,3}\Pi_3^* - \Delta_{2,4}\Pi_4^*. \tag{5.14}$$

Finally we have similarly

$$\Pi_1^* = \Pi_1 - \Delta_{1,2}\Pi_2^* - \Delta_{1,3}\Pi_3^* - \Delta_{1,4}\Pi_4^*. \tag{5.15}$$

The generalization for $k > 4$ is straightforward and reflection will show that Π_j^* is simply the number of components of G isomorphic with c_j. Since $\Pi_j^*(G)$ is a linear combination of the set $\Pi_i(G)$ it clearly also satisfies the extensive property and so we may write

$$\Psi(\Pi_i; t) = \Psi^*(\Pi_j^*; t), \tag{5.16}$$

where

$$\begin{aligned}&\Psi^*[\Pi_j^*(G) + \Pi_j^*(G'); t] \\ &\quad = \Psi^*[\Pi_j^*(G); t) + \Psi^*(\Pi_j^*(G'); t].\end{aligned} \tag{5.17}$$

Since the Π_j^* are independent this functional relation may be solved to yield a solution of the form (5.9), where

$$\begin{aligned}&\Psi^*(1, 0, 0, 0; t) = \Psi(1, 0, 0, 0; t) = \phi(c_1), \\ &\Psi^*(0, 1, 0, 0; t) = \Psi(\Delta_{1,2}, 1, 0, 0; t) = \phi(c_2), \\ &\Psi^*(0, 0, 1, 0; t) = \Psi(\Delta_{1,3}, \Delta_{2,3}, 1, 0; t) = \phi(c_3), \\ &\Psi^*(0, 0, 0, 1; t) = \Psi(\Delta_{1,4}, \Delta_{2,4}, \Delta_{3,4}, 1; t) = \phi(c_4),\end{aligned} \tag{5.18}$$

and generally

$$\begin{aligned}&\Psi^*(0, 0, \cdots , 1_n, \cdots ; t) \\ &\quad = \Psi(\Delta_{1,n}, \Delta_{2,n}, \cdots \Delta_{n-1,n}, 1_n, 0, 0, \cdots ; t) \\ &\quad = \phi(c_n).\end{aligned} \tag{5.19}$$

On substituting and collecting terms, therefore, we finally obtain Ψ in terms of the Π_i, and we have proved:

Theorem III: If $\phi = \Psi(\Pi_i(G); t)$ satisfies the extensive property (5.6) then

$$\Psi(\Pi_i; t) = \sum_{i=1}^{k} \Pi_i f_i(t), \tag{5.20}$$

where

$$\begin{aligned}f_1(t) &= \phi(c_1), \\ f_2(t) &= \phi(c_2) - \Delta_{1,2}\phi(c_1), \\ f_3(t) &= \phi(c_3) - \Delta_{2,3}\phi(c_2) + \{\Delta_{1,2}\Delta_{2,3} - \Delta_{1,3}\}\phi(c_1), \\ f_4(t) &= \phi(c_4) - \Delta_{3,4}\phi(c_3) + \{\Delta_{2,3}\Delta_{3,4} - \Delta_{2,4}\}\phi(c_2) \\ &\quad + \{-\Delta_{1,2}\Delta_{2,3}\Delta_{3,4} + \Delta_{1,2}\Delta_{2,4} \\ &\quad + \Delta_{1,3}\Delta_{3,4} - \Delta_{1,4}\}\phi(c_1),\end{aligned} \tag{5.21}$$

and generally in $f_n(t)$ the coefficient of $\phi(c_m)$ is evidently the sum over all products of the form

$$(-\Delta_{m,m+i})(-\Delta_{m+i,m+j}) \cdots (-\Delta_{m+l,n}), \qquad (5.22)$$

where

$$0 < i < j \cdots < l < n - m.$$

In practice it is more often convenient to determine the $f_i(t)$ directly from the recursion

$$f_{i+1}(t) = \phi(c_{i+1}) - \sum_{j=1}^{i} \Delta_{j,i+1} f_j(t). \qquad (5.23)$$

The weight functions f_i are thus obtained by substitution of the values of ϕ on each graph in turn. The result (5.20) is of great practical use in deriving series expansions.

As examples we quote the method of Domb[1] for the free energy of the Ising and Heisenberg model and of Domb and Wood[11] for the Heisenberg model. In both these problems the technique is immediately applicable to the dilute magnet by a simple modification of the constants which must be weighted by the appropriate probability functions. This is essentially the method of Rushbrooke[17] and of Abe.[37] The general method appears to have been developed by many authors and can be derived in other ways as, for example, by the methods of Kubo[38] and of Strieb *et al.*[39]

The result (5.20) is valid in both the weak and the strong systems, and we denote the weight functions of a connected graph by w_i and W_i, respectively, and write

$$\phi(G) = \sum_i p_i w_i(t) = \sum_i P_i W_i(t). \qquad (5.24)$$

The appropriate weak or strong weight functions can be obtained by direct substitution in (5.21) or recursively from (5.23). The two systems of weak and strong weights are related by the conversion matrix. Denote the column vectors corresponding to w_i, W_i of all graphs of r vertices by $\mathrm{w}(r)$, $\mathbf{W}(r)$. Then

$$\mathrm{w}(r) = \tilde{\mathbf{B}}(r)\mathbf{W}(r), \qquad (5.25)$$

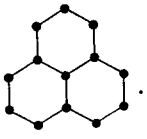
where $\tilde{\mathbf{B}}(r)$ denotes the transpose of $\mathbf{B}(r)$.[40] We exploit (5.24) to derive expansions for the mean number function.[28]

[37] R. Abe, Progr. Theor. Phys. (Kyoto) **31**, 412 (1964).
[38] R. Kubo, J. Phys. Soc. Japan **17**, 1100 (1962).
[39] B. Strieb, H. B. Callen, and G. Horwitz, Phys. Rev. **130**, 1798 (1963).
[40] W. L. Ferrar, *Algebra* (Oxford University Press, New York, 1941), Sec. 7.5 and Example 11.

6. LATTICE CONSTANTS FOR AN INFINITE GRAPH

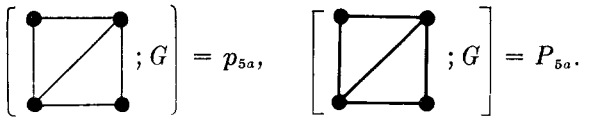
On an infinite graph the definition of a lattice constant requires modification. The infinite graphs that arise in physical applications are usually infinite crystal lattices. It is customary to denote by N the number of sites and to suppose that as $N \to \infty$ the edge effects become negligible. No difficulty arises with the more usual crystal lattices, which have a regular structure, and by a suitable choice of periodic boundary condition may be looped on a torus. For an infinite graph we only define connected lattice constants. The number of embeddings of a connected graph will be proportional to N and it is usual to redefine the corresponding lattice constant *per site*.

On an individual lattice the construction of a lattice constant table presents at the outset certain difficulties of notation. As remarked in the Introduction, although the number of types increases rapidly, this increase is far less rapid than that of the number of theoretically possible types, many of which are not embeddable in the lattice considered. For example, on the honeycomb lattice there is only one star with 13 vertices that is embeddable in the strong sense—the graph



The determination of the appropriate suffix or reference number for this if all the stars with up to 13 vertices were arranged in graph dictionary order is a procedure that rapidly becomes very difficult and is not practically feasible. In previous sections we have used the conveniently short notation p_i, P_i for theoretical purposes.

Quite extensive lattice constant data are to be found in the review by Domb,[1] who employs a system of symbols of type p_{nx}, P_{nx} for weak and strong constants where n denotes the number of edges and x serves to distinguish different topologies with the same number of edges and is arbitrarily asssigned. For example,



The system is quite convenient for some applications such as the expression of the first few coefficients of a particular series expansion in a form applicable to any graph but is not practical for more

extensive tables. We give in the next section a summary of the weak embeddings of every star of up to nine lines on the face-centered cubic in a scheme which can be applied without modification to other lattices as for example the diamond lattice.

7. THE ENUMERATION PROBLEM ON THE FACE-CENTERED CUBIC

The configurational problem presented by a close-packed lattice such as the face-centered cubic is complex. The most important constants are the star lattice constants, and in Table I we give an analysis of those with up to seven vertices. The order of magnitude of the counting problem, or actual determination of the value of the constants, as opposed to the enumeration problem, or listing of the possible graph types, can be gauged by the size of the total count by which we mean the sum of all the lattice constants in any class considered (in our example all stars with a given number of vertices). The number of theoretical possible stars increases rapidly with the number of vertices. The number of these stars that can be embedded in the face-centered cubic increases less rapidly, and this is especially the case for strong embeddings and for these latter the total count is also smaller. If we employ the method of space types the maximum asymmetry on the face-centered cubic of any type is $48N$ and since the total count of the 47 seven vertex stars in the strong system is only $3972N$ these correspond to only a small number of types each. We have therefore counted the strong star constants in this way and we give in the Appendix the 71 star graphs with from three to seven vertices that have nonzero strong embeddings for the face-centered cubic.

We have derived the corresponding set of 306 nonzero weak star lattice constants by employing the star conversion matrices. For most applications of weak constants, an edge grouping rather than a vertex grouping is appropriate. (For example, for the specific heat of the Ising model or the mean number of clusters in bond percolation processes.) The data from the star conversion matrices is then usefully supplemented by further graphs with low cyclomatic number. For most applications also drawings provide a cumbersome and somewhat unsatisfactory description, and we have therefore classified the weak constants as follows.

We first consider the graphs that are homeomorphs of a Jordan curve. In lattice statistics these are usually called polygons and denoted by p_n, where n denotes the number of edges. To avoid confusion with our previous notation and to remain consistent with our subsequent notation for more complex graphs, we use the symbol $(n)_p$ to represent *both* the graph *and* its weak lattice constant. (The strong constant will be denoted by $[n]_p$.) We follow this convention throughout the remainder of this paper.

On the face-centered cubic following the convention of Sec. 6 that for an infinite lattice these quantities are defined *per site*,

$$(3)_p = 8,$$
$$(4)_p = 33,$$
$$(5)_p = 168,$$
$$(6)_p = 970,$$
$$(7)_p = 6168,$$
$$(8)_p = 42\,069,$$
$$(9)_p = 301\,376.$$

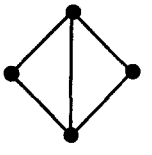
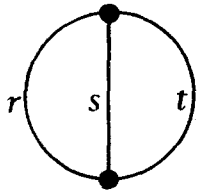
After the Jordan curves, which have cyclomatic number 1, we list graphs with cyclomatic number 2. The first example is the graph  and all the others are homeomorphs of this. We call such graphs θ graphs. A θ graph is defined if the number of edges in each of the three "bridges" connecting the nodes is given. Denoting these by r, s, t we describe these graphs by the notation represented by the scheme

TABLE I. Data for face-centered cubic lattice.

Vertices	Number of star constants			Total count	
	Theoretical	Nonzero Weak	Nonzero Strong	Weak	Strong
3	1	1	1	$8N$	$8N$
4	3	3	3	$71N$	$29N$
5	10	8	4	$882N$	$126N$
6	56	41	16	$13832N$	$685N$
7	468	253	47	$242476N$	$3972N$

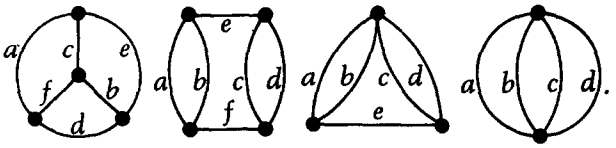


$= (r, s, t)_\theta \qquad r \leq s \leq t$

example  $= (1, 3, 3)_\theta$.

We list in Table II the θ graphs on the face-centered cubic with up to nine edges together with the corresponding weak lattice constants. The ordering of the parameters $r \leq s \leq t$ is purely for convenience of reference. $(1, 3, 3)_\theta$ and $(3, 1, 3)_\theta$ are the same graph.

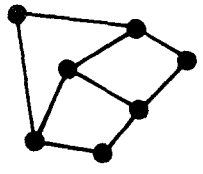
Graphs with cyclomatic index 3 are all homeomorphs of four distinct topologies[41] which we name as follows:



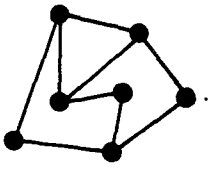
α graph β graph γ graph δ graph

The α graph (or tetrahedral graph) is described by the following scheme. The edge lengths of bridges between distinct pairs of nodes are recorded as number pairs in dictionary order. The final pair is *then* reversed, if necessary, to ensure that the first entries for each pair are all incident on one vertex.

Thus in general we write $(a, b; c, d; e, f)$ with $a \leq b$, $c \leq d$, $a \leq c$, $c \leq \min(e, f)$ and the bridges corresponding to a, c, e are all incident on one node. For example, we illustrate two distinct α graphs,



$(1, 2; 1, 2; 1, 2)_\alpha$



$(1, 2; 1, 2; 2, 1)_\alpha$

TABLE II. Face-centered cubic. θ graphs.

$l=5$	$(1,2,2)_\theta$	36	$l=9$	$(1,2,6)_\theta$	122592
$l=6$	$(1,2,3)_\theta$	384		$(1,3,5)_\theta$	78420
	$(2,2,2)_\theta$	36		$(1,4,4)_\theta$	34380
$l=7$	$(1,2,4)_\theta$	2400		$(2,2,5)_\theta$	28560
	$(1,3,3)_\theta$	966		$(2,3,4)_\theta$	40848
	$(2,2,3)_\theta$	600		$(3,3,3)_\theta$	5252
$l=8$	$(1,2,5)_\theta$	16464			
	$(1,3,4)_\theta$	11616			
	$(2,2,4)_\theta$	3888			
	$(2,3,3)_\theta$	3132			

[41] See, for example, G. W. Ford and G. E. Uhlenbeck, Proc. Natl. Acad. Sci. U. S. 43, 163 (1957).

TABLE III. Face-centered cubic. α graphs.

$l = 6$	$(1,1;1,1;1,1)_\alpha$	2
$l = 7$	$(1,1;1,1;1,2)_\alpha$	48
$l = 8$	$(1,1;1,1;1,3)_\alpha$	312
$l = 8$	$(1,1;1,1;2,2)_\alpha$	78
$l = 8$	$(1,1;1,2;1,2)_\alpha$	456
$l = 9$	$(1,1;1,1;1,4)_\alpha$	2088
$l = 9$	$(1,1;1,1;2,3)_\alpha$	840
$l = 9$	$(1,1;1,2;1,3)_\alpha$	5328
$l = 9$	$(1,1;1,2;2,2)_\alpha$	1392
$l = 9$	$(1,2;1,2;1,2)_\alpha$	696
$l = 9$	$(1,2;1,2;2,1)_\alpha$	632

TABLE IV. Face-centered cubic. β graphs.

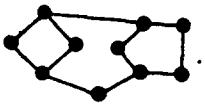
$l = 8$	$(1,2;1,2;1,1)_\beta$	564
$l = 9$	$(1,2;1,2;1,2)_\beta$	6696
$l = 9$	$(1,2;1,3;1,1)_\beta$	5472
$l = 9$	$(1,2;2,2;1,1)_\beta$	1656

As before the ordering of the parameters is one of convenience and provides a unique description. The essential part of the convention is that the *first* entries of each pair are all incident on one node. Thus, for example, the last example $(1, 2; 1, 2; 2, 1)_\alpha$ is the same as $(1, 2; 2, 1; 1, 2)_\alpha$.

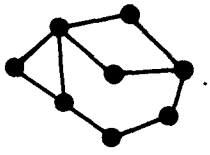
A description of the remaining three topologies presents no ambiguities and we write

$(a, b; c, d; e, f)_\beta$ with $a \leq b$, $c \leq d$, $e \leq f$, $a \leq c$

and the pair e, f last as drawn. A specific example is

$(2, 2; 2, 3; 1, 2)_\beta =$  .

For γ graphs we write $(a, b; c, d; e)_\gamma$ with e last as drawn, and a specific example is

$(1, 2; 2, 2; 3)_\gamma =$  .

For δ graphs we write as an obvious extension of

TABLE V. Face-centered cubic. γ graphs.

$l = 7$	$(1,2;1,2;1)_\gamma$	192
$l = 8$	$(1,2;1,2;2)_\gamma$	1056
$l = 8$	$(1,2;1,3;1)_\gamma$	2016
$l = 8$	$(1,2;2,2;1)_\gamma$	528
$l = 9$	$(1,2;1,2;3)_\gamma$	6384
$l = 9$	$(1,2;1,3;2)_\gamma$	10128
$l = 9$	$(1,2;1,4;1)_\gamma$	12144
$l = 9$	$(1,2;2,2;2)_\gamma$	3024
$l = 9$	$(1,2;2,3;1)_\gamma$	5472
$l = 9$	$(1,3;1,3;1)_\gamma$	5040
$l = 9$	$(1,3;2,2;1)_\gamma$	2568
$l = 9$	$(2,2;2,2;1)_\gamma$	312

TABLE VI. Face-centered cubic. δ graphs.

$l = 7$	$(1,2,2,2)_\delta$	= 24
$l = 8$	$(1,2,2,3)_\delta$	= 384
$l = 8$	$(2,2,2,2)_\delta$	= 9
$l = 9$	$(1,2,2,4)_\delta$	= 2352
$l = 9$	$(1,2,3,3)_\delta$	= 1848
$l = 9$	$(2,2,2,3)_\delta$	= 192

that for θ graphs

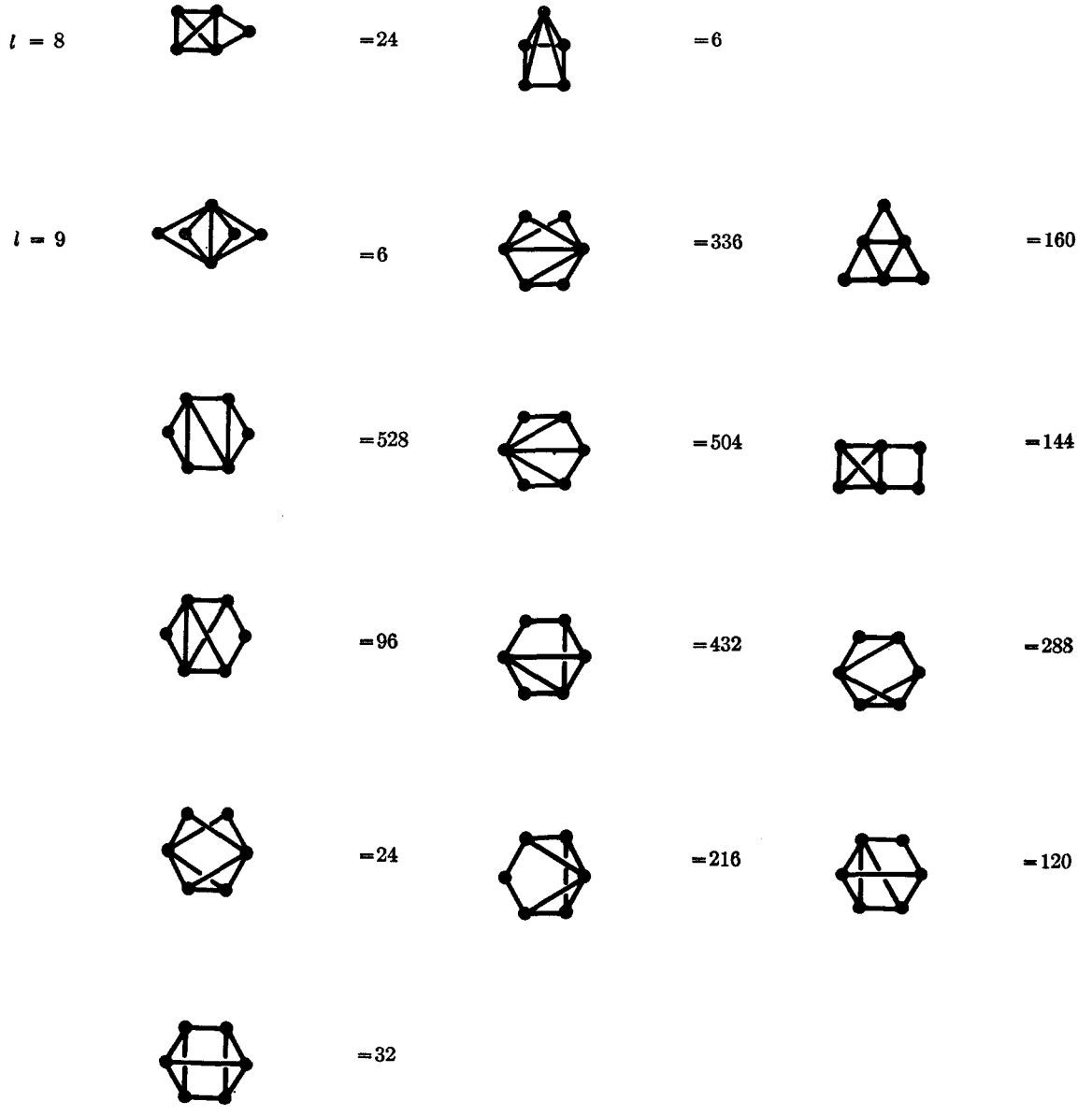
$$(a, b, c, d)_\theta \qquad a \leq b \leq c \leq d.$$

We give in Tables III–VI the graphs of cyclomatic index 3 on the face-centered cubic with up to nine edges. For $l \leq 8$ they are given by Domb, for $l = 9$ we have used the seven vertex conversion matrix and the data of the Appendix.

For cyclomatic number 4 there are 17 topological classes, and since there are only 15 such graphs that are embeddable in the face-centered cubic for up to nine edges, we simply list in Table VII the individual graphs. The list exhausts all the weak stars, 71 in all, for $l \leq 9$, and to this order there are no stars with cyclomatic number 5 or more embeddable in the face-centered cubic.

That the enumeration of the 47 strong star em-

TABLE VII. Stars of cyclomatic number 4, with up to nine edges, that are embeddable in the face-centered cubic lattice in the weak sense.

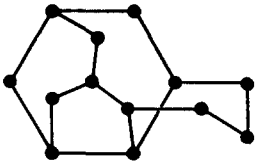


beddings with seven vertices is complete is confirmed by their yielding after conversion to the weak system the correct values of $(7)_p$ and the four θ graphs with $l = 8$ (Table II) which are given by Domb (p_{8a}, p_{8b}, p_{8c}, p_{8d} in his notation). This provides a check on the completeness of the strong, and therefore the weak, enumeration.

The 468 × 468 star conversion matrix of seventh order has been evaluated as described in Sec. 4. In this way we obtain a complete list of weak stars with up to seven vertices. To extend the *vertex grouping* it is necessary to enumerate all stars with eight vertices. This is feasible in the strong system, as is the setting up of the eight-order conversion matrix, and we are undertaking this work. To extend the *edge grouping* it is more efficient to obtain constants of low cyclomatic number directly in the weak system. We have already done this for $l = 9$ by adding the polygon and θ graphs. For $l = 10$ we need only the polygon, θ graphs, and α, β, γ, δ graphs. These are conveniently counted by an extension of the techniques of Martin.[23]

To proceed further and describe stars with more than nine edges embeddable in the face-centered cubic it is desirable to develop a notation for those of cyclomatic number 4 of which there are 17 topological classes. We are undertaking this.[42]

The notation of this section is immediately applicable to other lattices. For example, on the diamond lattice there is only one star with cyclomatic number above 3 for $l \leq 16$ the graph



= 12,

and we list in Table VIII the remaining 45 stars up to this order.

For practical purposes, actual drawings of these graphs are unwieldy and seldom, if ever, required. We think our classification is convenient for the presentation of data to a computer, and if extensive

TABLE VIII. Weak star lattice constants for the diamond lattice.

Jordan curves			
$(6)_p$ = 2			
$(8)_p$ = 3			
$(10)_p$ = 24			
$(12)_p$ = 94			
$(14)_p$ = 582			
$(16)_p$ = 3126			
θ-graphs			
$l = 10$	$(2,4,4)_\theta$ = 6	$l = 16$	$(2,4,10)_\theta$ = 876
$l = 11$	$(1,5,5)_\theta$ = 18		$(2,6,8)_\theta$ = 276
$l = 12$	$(2,4,6)_\theta$ = 12		$(4,4,8)_\theta$ = 72
$l = 13$	$(1,5,7)_\theta$ = 72		$(4,6,6)_\theta$ = 138
	$(3,3,7)_\theta$ = 24		
	$(3,5,5)_\theta$ = 12		
$l = 14$	$(2,4,8)_\theta$ = 204		
	$(2,6,6)_\theta$ = 6		
	$(4,4,6)_\theta$ = 24		
$l = 15$	$(1,5,9)_\theta$ = 576		
	$(1,7,7)_\theta$ = 60		
	$(3,3,9)_\theta$ = 72		
	$(3,5,7)_\theta$ = 72		
	$(5,5,5)_\theta$ = 8		
α-graphs		γ-graphs	
$l = 12$	$(2,2;2,2;2,2)_\alpha$ = 1	$l = 15$	$(1,5;2,4;3)_\gamma$ = 48
$l = 13$	none		$(1,5;4,4;1)_\gamma$ = 24
$l = 14$	$(1,1;1,3;4,4)_\alpha$ = 12	$l = 16$	$(1,5;1,5;4)_\gamma$ = 48
	$(1,3;1,3;2,4)_\alpha$ = 12		$(2,4;2,4;4)_\gamma$ = 12
$l = 15$	$(1,4;1,4;1,4)_\alpha$ = 16		
	$(1,2;2,3;2,5)_\alpha$ = 24	δ-graphs	
$l = 16$	$(1,1;1,3;4,6)_\alpha$ = 24	$l = 16$	$(1,5,5,5)_\delta$ = 4
	$(1,3;1,3;2,6)_\alpha$ = 12		
	$(1,3;1,5;2,4)_\alpha$ = 24		
	$(1,5;2,2;3,3)_\alpha$ = 12		
	$(2,2;2,2;2,6)_\alpha$ = 24		
β-graphs			
$l = 14$	$(2,4;2,4;1,1)_\beta$ = 6		
$l = 15$	$(1,5;2,4;1,2)_\beta$ = 72		
$l = 16$	$(1,5;1,5;1,3)_\beta$ = 108		
	$(1,5;1,5;2,2)_\beta$ = 54		
	$(2,4;2,4;2,2)_\beta$ = 6		
	$(2,4;2,6;1,1)_\beta$ = 12		
	$(3,3;3,3;2,2)_\beta$ = 6		

[42] The classes are illustrated in Ref. 28, Appendix.

data tables are to be exploited computers are likely to be employed. Many properties that arise in applications are also soluble in quite general form once the topology of the graph is known. For example, the star subgraphs of a β graph are

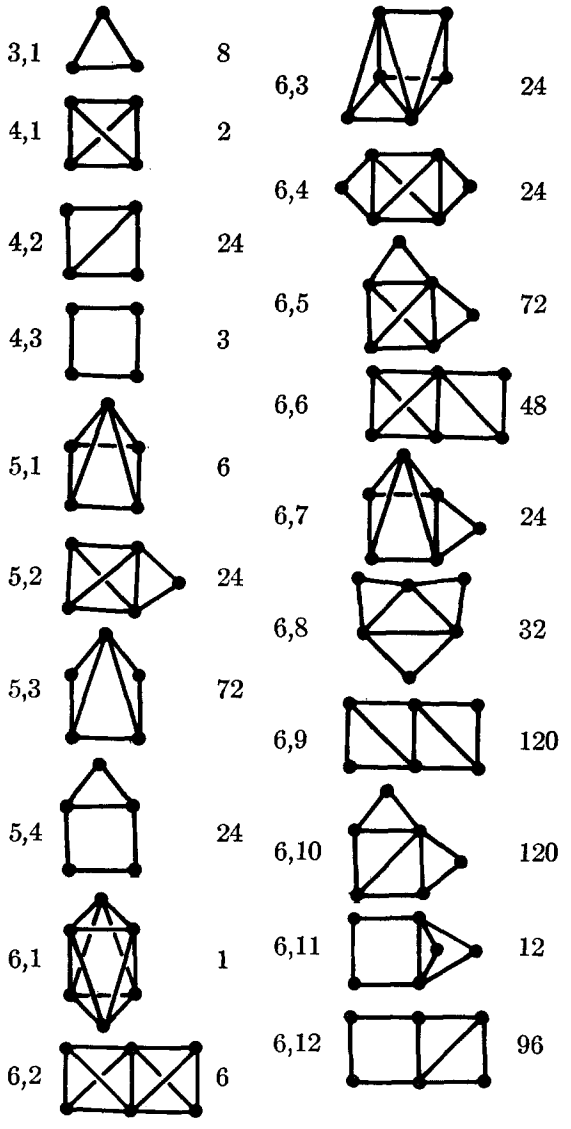
$$(a, b, e + f + c)_\theta \quad (a, b, e + f + d)_\theta$$
$$(c, d, e + f + a)_\theta \quad (c, d, e + f + b)_\theta$$
$$(a + b)_p \quad (c + d)_p \quad (a + e + f + c)_p$$
$$(a + e + f + d)_p \quad (b + e + f + c)_p$$
$$(b + e + f + d)_p.$$

ACKNOWLEDGMENTS

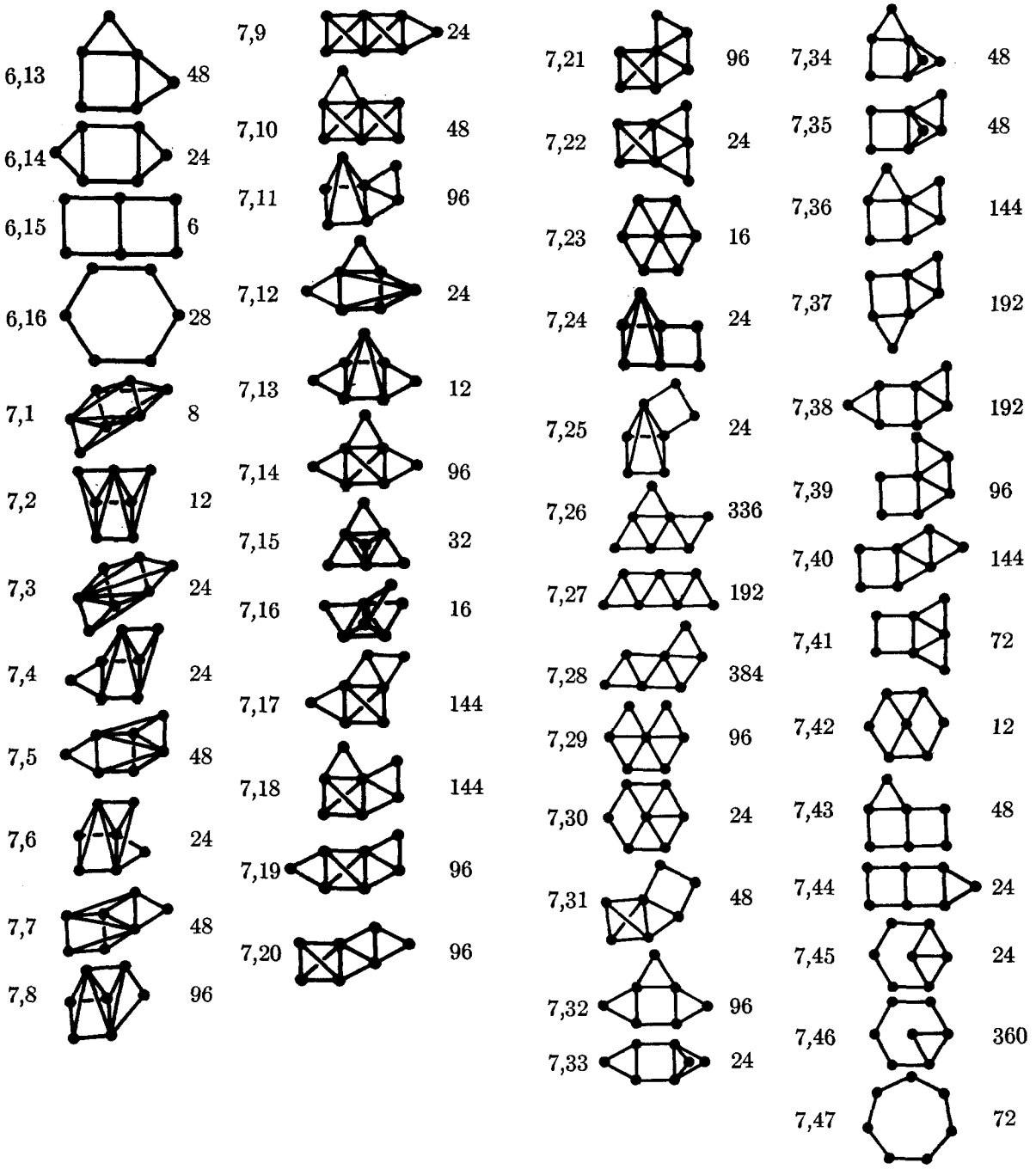
The authors are particularly indebted to Professor C. Domb and Professor M. E. Fisher, of King's College, London, for much constructive criticism of the theoretical section of this paper. The work described in Sec. 4 forms part of the research program of the National Physical Laboratory (Teddington, England) and is published by permission of the Director. We are especially grateful to Dr. J. L. Martin of the Mathematics Division of the National Physical Laboratory for his assistance at all stages of this research and for the machine counting of many lattice constants.

APPENDIX

The 71 stars with from 3 to 7 vertices that have nonzero strong embeddings in the face-centered cubic lattice are shown in the following.



(*Star diagrams continued on next page*)


6,13 48
6,14 24
6,15 6
6,16 28
7,1 8
7,2 12
7,3 24
7,4 24
7,5 48
7,6 24
7,7 48
7,8 96
7,9 24
7,10 48
7,11 96
7,12 24
7,13 12
7,14 96
7,15 32
7,16 16
7,17 144
7,18 144
7,19 96
7,20 96
7,21 96
7,22 24
7,23 16
7,24 24
7,25 24
7,26 336
7,27 192
7,28 384
7,29 96
7,30 24
7,31 48
7,32 96
7,33 24
7,34 48
7,35 48
7,36 144
7,37 192
7,38 192
7,39 96
7,40 144
7,41 72
7,42 12
7,43 48
7,44 24
7,45 24
7,46 360
7,47 72

Percolation Processes. I. Low-Density Expansion for the Mean Number of Clusters in a Random Mixture*

J. W. ESSAM
Westfield College, London, England
AND
M. F. SYKES
Wheatstone Physics Laboratory, King's College, London, England
(Received 13 July 1965)

A cluster expansion, valid at low densities, is derived for the mean number of clusters in a random mixture of sites or bonds on a graph. It is shown that only clusters without a cut-point (stars) are required, and a number of general theorems for determining the weights are proved.

1. INTRODUCTION

IN this paper we examine the derivation of series expansions, valid at low densities, for the mean number of clusters in a random mixture. We have introduced this problem in a previous paper,[1] hereafter referred to as I. For a detailed introduction to the problem, and to the closely related percolation problem, reference should be made to I, to Fisher and Essam,[2] and for a general survey, to Hammersley and Frisch.[3] We make use of the general concepts of graph theory which we have described in a paper,[4] hereafter referred to as II.

We begin with a statement of the problem in the formal terminology of graph theory and then derive a cluster expansion for the mean number function. We show that the expansion depends only on connected clusters without cut-points (stars) and that the corresponding weights are particularly simple in the system of weak lattice constants. We prove a number of general theorems of use in determining the weights of individual clusters and, as an example, derive series expansions for the site and bond problem on the plane triangular lattice.

We subsequently apply these expansions to a study of the mean number function and extend the series developments for the mean size of clusters in a random mixture.

2. STATEMENT OF THE PROBLEM

(A) Site Problem

In site mixtures the sites, or vertices, of a linear graph G are supposed colored black with probability p or white with complementary probability q. We adopt the convention that, in such a random mixture of two species represented by the black and white sites, the primary species is the black, and we refer to small p as *low density*. The bonds, or edges, of G are regarded as colored black if they connect two black sites, white if they connect two white sites, uncolored if they connect sites of different colors.

Any realization R of the probability distribution on G defines two linear graphs R_B and R_W which are, respectively, the ensembles of black and white clusters. More precisely, R_B is the section[5] graph of G defined by all the sites of G that are black in R and the term black cluster is used to describe any connected component of R_B. In general, R_B has many connected components, and it is the expectation value of the number of these that we study. Denoting the number of connected components of R_B by $n(R_B)$, we define the mean number function K by

$$K(p; G) = \langle n(R_B) \rangle. \tag{2.1}$$

(B) Bond Problem

In bond mixtures the bonds, or edges, of a linear graph G are supposed colored black with probability $\bar{p}$ or white with complementary probability $\bar{q}$. A realization $\bar{R}$ of the bond probability distribution on G defines two linear graphs $\bar{R}_B$ and $\bar{R}_W$ which are, respectively, the ensemble of black and white bond clusters. More precisely, $\bar{R}_B$ is the subgraph of G defined by the edges of G that are black in $\bar{R}$, together with their end points, and a black bond cluster is a connected component of $\bar{R}_B$. We define the corresponding mean number function by

$$\bar{K}(\bar{p}; G) = \langle n(\bar{R}_B) \rangle. \tag{2.2}$$

* This research has been supported (in part) by the U. S. Department of the Army through its European Research Office.

[1] M. F. Sykes and J. W. Essam. J. Math. Phys. 5, 1117 (1964).

[2] M. E. Fisher and J. W. Essam, J. Math. Phys. 2, 609 (1961).

[3] H. L. Frisch and J. M. Hammersley, J. Soc. Indust. Appl. Math. 11, 894 (1963).

[4] M. F. Sykes, J. W. Essam, B. R. Heap, and B. J. Hiley, J. Math. Phys. 7, 1557 (1966).

[5] Defined in II, Sec. 2.

In this simple conceptual form the sites of the graph are not assigned a color. The bond problem may be studied as a site problem in which each bond is made to correspond to a site on a suitably defined covering graph.[2]

In many applications of bond mixtures the bonds are primarily considered as connections between the sites, and a simplification results if we adopt the following convention which we call the *null-cluster convention.*

Suppose that for any realization $\bar{R}$ two sites are defined as connected if they are joined by a black bond. We employ the term *black-connected* cluster to describe any connected component of the *partial*[5] graph, $\bar{P}_B$ of G whose edge set is the edge set of $\bar{R}_B$. Some black-connected clusters may reduce to isolated sites (null-clusters). We define the mean number function $\bar{K}_0$ for this convention to be the expectation value for the number of black-connected site clusters. We write

$$\bar{K}_0(\bar{p}; G) = \langle n(\bar{P}_B)\rangle, \tag{2.3}$$

where the suffix on $\bar{K}_0$ denotes the operation of the null-cluster convention. If we denote the mean number, or expectation value, of the isolated sites or null-clusters by $\langle$n.c.$\rangle$, then

$$\bar{K}_0(\bar{p}; G) = \bar{K}(\bar{p}; G) + \langle \text{n.c.}\rangle. \tag{2.4}$$

3. LOW-DENSITY CLUSTER EXPANSION FOR $K(p; G)$ (SITE PROBLEM)

We now describe a method of obtaining the mean number function which avoids a detailed specification of cluster perimeters[6] required by the alternative perimeter method. It has been outlined in II and is readily formalized.

To recapitulate, any realization, R, of the probability distribution defines a section graph R_B of G which is the graph of the black sites and bonds. In the notation of II, G contains $[g_i; G]$ section graphs isomorphic with g_i and the probability of any one of these being R_B is just

$$p^{v_i}(1-p)^{vG-v_i}. \tag{3.1}$$

Thus if g_i has n_i connected components

$$K(p; G) = \sum_i n_i p^{v_i}(1-p)^{vG-v_i}[g_i; G], \tag{3.2}$$

where the summation is taken over all the strong lattice constants of G. Now these latter can be expressed in terms of the connected constants only, and because mean number is an extensive property it follows, by the arguments of Sec. 5 of II, that the resultant expression is linear in the connected constants and that a cluster expansion may be developed for $K(p; G)$. Thus we may write, denoting connected graphs by c_i,

$$K(p; G) = \sum_i W_i(p)[c_i; G], \tag{3.3}$$

where W_i is the appropriate weight function of c_i.

It is evident from the form of (3.2) that the weight functions are polynomials in p. They can be evaluated by carrying out the substitutions for the separated lattice constants, and the coefficient of any connected constant $[c_i; G]$ will come from two sources (II, Theorem II):

(1) from c_i itself. The presence of the factor p^{v_i} in (3.1) ensures that this contribution is always of degree at least v_i in p.

(2) from separated constants such as, for example, a three component graph $c_r \cup c_s \cup c_t$. In the reduction of these it is evident that c_i cannot occur as an overlap partition unless $c_r \cup c_s \cup c_t$ has at least v_i vertices, and therefore again by (3.1) the contribution is of degree at least v_i in p.

Alternatively, the weight functions may be evaluated from the mean number functions of the individual connected constants. Since the weight of any graph c_i is expressible in terms of the $K(p; G)$ of c_i and all its connected subgraphs [Eqs. (5.19) and (5.20) of II], and this expression is linear, and further, each function is a polynomial in p of degree at most v_i, it follows that $W_i(p)$ is of degree at most v_i in p. Thus by virtue of the previous result, $W_i(p)$ can only have one nonzero coefficient—that of the v_ith power of p, and we state the result as a theorem.

Theorem I: The strong weight function $W_i(p)$ of $[c_i; G]$ can be written $K_i p^{v_i}$, where K_i is independent of p.

The strong weight functions arise quite naturally in the site problem since the clusters studied are all section graphs. However, we often find it convenient to work with the corresponding weak weight functions which we introduce by the following theorem.

Theorem II: The weak weight function $w_i(p)$ of $(c_i; G)$ can be written $k_i p^{v_i}$, where k_i is independent of p.

Proof: The result follows at once by conversion of the strong weight functions $K_i p^{v_i}$ into the weak weight functions by means of the conversion matrix for weights, which is just the transpose of the reciprocal conversion matrix for the connected con-

[6] Defined in I, Sec. 2, and in Ref. 2, Sec. 2.

stants, and which relates constants with the same number of *vertices*.

We now define $K(c_i) = K_i$ and $k(c_i) = k_i$ to be the (strong) K-wt and (weak) k-wt of c_i, respectively. Following the convention of II, Sec. 2, we also abbreviate $K(s_i)$, $k(s_i)$ to K_i, k_i whenever it is clear from the context which graph dictionary is being used, and further, we sometimes write K_G, k_G for $K(G)$, $k(G)$. For any graph G we have

$$K(p; G) = \sum_i [c_i; G]K_i p^{v_i} \tag{3.4}$$

$$= \sum_i (c_i; G)k_i p^{v_i}, \tag{3.5}$$

where the summations are taken over all the connected constants of G. The form of (3.4) and (3.5) which results from Theorems I and II makes it possible, when G is an infinite graph, to derive series developments in powers of p as appropriate to low densities. By the methods of this section, the weights of the site and the bond are found to be $+1$ and -1, respectively (in both systems), and we use this result as a lemma to prove the next theorem.

Theorem III: For any graph G

$$\sum_{i>2} [c_i; G]K_i = \sum_{i>2} (c_i; G)k_i = C(G), \tag{3.6}$$

where $C(G)$ denotes the cyclomatic number (or circuit rank) of G and the summation is taken over all the connected constants of G except the site and the bond. (In graph dictionary order these will have suffixes 1 and 2.)

Proof: In both the weak and the strong systems, the site has weight $+1$ and the bond, weight -1. Thus from (3.4)

$$K(p; G) = v_G p - l_G p^2 + \sum_{i>2} [c_i; G]K_i p^{v_i}. \tag{3.7}$$

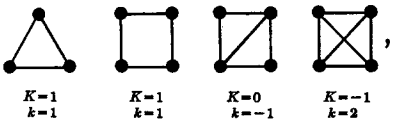
But for $p = 1$ the expected number of clusters must reduce to the number of components in G or $K(1; G) = n_G$, and on setting $p = 1$ in (3.7),

$$\sum_{i>2} [c_i; G]K_i = n_G - v_G + l_G = C(G) \tag{3.8}$$

by definition. Likewise the result holds for weak weights.

Theorem III may be used to derive the weak and strong weights by successive application to the connected constants arranged in a suitable graph dictionary order (i.e., in order of ascending cyclomatic number). *We may thus take (3.6) as defining the quantities K_i, k_i associated with a graph c_i.*

The weights of all graphs with three and four vertices are found to contain only four with nonzero contributions,



, (3.9)

which suggests that only stars have nonzero weight—a result we now prove.

Theorem IV: If c_i has a cut point, $K_i = k_i = 0$.

Proof: We prove the result in the strong system, and the result for the weak system follows by changing to weak weights throughout.

Suppose a graph G has a cut point at the vertex A. Then by definition the deletion of A, together with all its incident edges, leave a graph with at least two connected components. Denote the vertex set of any one of these by V' and that of all the others by V''. Denote by G' the section graph of G with the vertex set $V' + A$ and by G'' that with vertex set $V'' + A$. Then $G = G' + G''$. By application of (3.6) to G, G' and G'',

$$\sum_{i>2} [c_i; G]K_i = C(G), \tag{3.10}$$

$$\sum_{i>2} [c_i; G']K_i = C(G'), \tag{3.11}$$

$$\sum_{i>2} [c_i; G'']K_i = C(G''), \tag{3.12}$$

and since A is an articulation point

$$C(G) = C(G') + C(G''). \tag{3.13}$$

The constants of G result from embeddings of each c_i in G, and these may be grouped into three mutually disjoint classes: those that lie wholly in G', those that lie wholly in G'', and those that lie neither wholly in G nor wholly in G' and which necessarily correspond to those c_i with cut points. Denoting the contribution of this third class of embeddings by an asterisk, we must have [from (3.10)–(3.13)]

$$\sum_{i>2} [c_i; G]^* K_i = 0, \tag{3.14}$$

and since the only connected graph of three points with a cut point has weight zero, the result follows inductively by successive application of (3.14) to all graphs with cut points.

Theorem IV enables the definitive equation (3.6) for weights to be restricted to multiply connected graphs, and we may write, for any such graph M,

$$\sum_{i>1} [s_i; M]K_i = \sum_{i>1} (s_i; M)k_i = C(M), \tag{3.15}$$

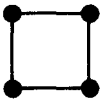
where the summation is taken over all stars except the bond (s_1). For theoretical purposes it is most

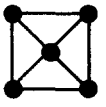
convenient to study the weak weights; we do so in Sec. 5 after we have developed the cluster expansion for the bond problem to which, as we show, the weak weights also apply. For practical purposes the strong weights are useful in the actual derivation of series expansions, and we prove two theorems, restricted to strong weights, that simplify the derivation of individual weights.

Theorem V: The strong K-wt of a graph G is the coefficient of the v_Gth power of p in $K(p; G)$.

Proof: The result is evident since there is only one nonzero constant with v_G vertices in the strong system—the graph itself.

As examples of the application of this theorem we quote

$G =$  $K(p; G) = 4p - 4p^2 + p^4, \qquad K_G = +1, \tag{3.16}$

$G =$  $K(p; G) = 5p - 8p^2 + 4p^3 + p^4 - p^5, \quad K_G = -1. \tag{3.17}$

We remark, parenthetically, that the corresponding k-wts can be found by use of the conversion matrix. The method is capable of elaboration and an appropriate technology can be developed to obtain $K(p; G)$ for an individual graph by the use of recurrence relations. It is found that many stars have zero K-wt, and most of these are accounted for by the next theorem.

Theorem VI: If G is a graph with a cut set of n vertices, and further the section graph which has these n vertices as vertex set is a complete graph of n vertices, then $K_G = 0$.

Proof: Denote the vertex set of the cut set by N. Denote the vertex set of one of the connected components that result from the deletion of N and its incident edges by V', and that of all the others by V''. Denote by G' the section graph with vertex set $N + V'$, and by G'' that with vertex set $V'' + N$. If in a realization R of the distribution any of the vertices in N are black, they must all be members of the same component of R_B since the section graph of G defined by N is a complete graph. But $G = G' + G''$ and the number of components in the sum graph G is the sum of the components in G' and G'' if the cut set contains no components (probability q^n). If the cut set contains a component it contributes both to G' and G''. Thus

$$K(p; G' + G'') = K(p; G') + K(p; G'') + q^n - 1. \tag{3.18}$$

By Theorem V the required weight is the coefficient of the $(v_{G'} + v_{G''} - n)$th power of p, and this exceeds max $(v_{G'}, v_{G''})$ since $v_{G'} > n$, $v_{G''} > n$. Therefore $K_G = 0$.

4. LOW-DENSITY CLUSTER EXPANSION FOR $\bar{K}(\bar{p}; G)$ (BOND PROBLEM)

The cluster expansion method applies to bond mixtures in an analogous manner to the treatment of site mixtures. If g_i is any subgraph of G, with no isolated vertices, the probability of g_i being a realization of the bond distribution probability is now

$$\bar{p}^{l_i}(1 - \bar{p})^{l_G - l_i}, \tag{4.1}$$

and the argument proceeds formally as for the site problem. Thus we now write

$$\bar{K}(\bar{p}; G) = \sum_i (c_i; G)\bar{w}_i(\bar{p}) \tag{4.2}$$

and obtain in place of Theorem I:

Theorem VII: The bond weight function $\bar{w}_i(\bar{p})$ for $(c_i; G)$ can be written $\bar{k}_i\bar{p}^{l_i}$, where $\bar{k}_i$ is independent of $\bar{p}$.

Because the conversion matrix for weights converts from constants with r vertices to constants with r *vertices* (and not *edges*), the *strong* bond weight function of a graph, $\bar{W}_i(\bar{p})$, is in general a polynomial in $\bar{p}$.

By including the subgraphs of G with isolated vertices we obtain corresponding results for the null-cluster convention, and we denote the corresponding weights, independent of $\bar{p}$ in Theorem VII by $\bar{k}_i^0$. The bond weights of the site and the bond are found to be, for the site $\bar{k} = 0$, $\bar{k}^0 = +1$ and,

for the bond $\bar{k} = +1$, $\bar{k}^0 = -1$. The null-cluster convention leads directly to:

Theorem VIII: For any graph c,

$$\bar{k}_i^0 = k_i, \tag{4.3}$$

or with the null cluster convention, the bond and site weights are identical.

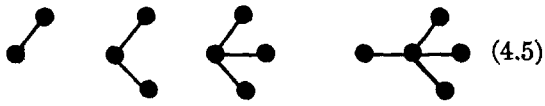
Proof: The result follows because under the null-cluster convention the site has weight $+1$, the bond, weight -1, and the number of clusters for $p = 1$ is the number of components in G. Therefore by the arguments of Theorem III the weights are defined by

$$\sum_{i>2} (c_i; G)\bar{k}_i^0 = C(G), \tag{4.4}$$

which is identical with the definitive equation (3.6) for weak weights.

We examine the properties of weak weights in Sec. 5 and conclude this section with a closer study of the relation between $\bar{k}_i$ and $\bar{k}_i^0$.

To relate the two systems of weak bond weights, we first observe that corresponding to Theorem V we have the result that the weak $\bar{k}$-wt of a graph G is the coefficient of the l_Gth power of $\bar{p}$ in $\bar{K}(\bar{p}; G)$. By an argument closely parallel to that of Theorem VI, it can be shown that graphs with a cut-vertex whose deletion, together with all its incident edges, leaves a graph with at least one edge have zero $\bar{k}$-wt. The only graphs with cut-vertices that do not satisfy this latter condition are those with the obvious general topology



(4.5)

To determine the weight of the general graph a_s of this type, we use

$$\bar{K}(\bar{p}; a_s) = 1 - \bar{q}^s = \binom{s}{1}\bar{p} - \binom{s}{2}\bar{p}^2 + \cdots + (-1)^{s+1}\binom{s}{s}\bar{p}^s, \tag{4.6}$$

and therefore from the last coefficient the $\bar{k}$-wt of $a_s = (-1)^{s+1}$.

For a general graph with no repeated bonds (i.e., not a multigraph) the contribution of graphs of type (4.5) is

$$\sum_{\substack{\text{all}\\ \text{sites}}} \sum_{s} (-1)^{s+1}\binom{z}{s}\bar{p}^s - l_G\bar{p}, \tag{4.7}$$

where the sum runs over all the sites of G and z is the number of edges incident on a site. Every a_s can be associated with its center point except the bond a_1, which is counted twice.

For the null cluster convention we must add the number of isolated clusters, and with the same restriction on multigraphs,

$$\langle \text{n.c.} \rangle = \sum \bar{q}^s. \tag{4.8}$$

On adding (4.8) and (4.7) we are left with

$$1 - l_G\bar{p}, \tag{4.9}$$

and the contribution from terms of type (4.5) cancel except for the site and the bond which now have weights $+1$ and -1, respectively, and we may write

$$\bar{k}_i^0 = \bar{k}_i = k_i \quad \text{for all stars} \quad i > 2. \tag{4.10}$$

On a regular lattice of coordination number z, with N sites, we may write

$$\langle \text{n.c.} \rangle = Nq^s. \tag{4.11}$$

5. PROPERTIES OF THE WEAK WEIGHTS

We now establish a number of theorems applicable to the weak k-wts defined for multiply connected graphs by

$$\sum_{i\geq 2} (s_i; M)k_i = C(M). \tag{5.1}$$

Theorem IX: If two graphs are homeomorphic they have equal k-wts.

Proof: The result is more or less obvious from the definitive equation (5.1). A tedious proof is readily constructed, but we confine our treatment to examples from which it is evident that the result will follow inductively.

First, every Jordan[7] curve has weight $+1$ since there is only one multiply connected subgraph, the graph itself, and the cyclomatic number is 1. [The site and bond are excluded by (5.1).] Thus

$$k\text{-wt of } (n)_p = +1. \tag{5.2}$$

For stars of cyclomatic number 2 there is only one topological type—the θ graph. Any θ graph $(r, s, t)_\theta$ has three Jordan subgraphs:

$$(r + s)_p, \quad (r + t)_p, \quad (s + t)_p, \tag{5.3}$$

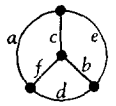
all with weight $+1$, and therefore

$$k\text{-wt of } (r, s, t)_\theta = -1. \tag{5.4}$$

For stars of cyclomatic number 3, there are

[7] The various types of graph are described in II, Sec. 7.

four topological types, and we work with only one, the tetrahedral or α graph:



or $(a, b; c, d; e, f)_\alpha$,

of which the relevant subgraphs are the seven Jordan curves

$$(a + c + f)_p, \quad (a + d + e)_p, \quad (b + c + e)_p,$$
$$(b + d + f)_p, \quad (a + b + c + d)_p,$$
$$(a + b + e + f)_p, \quad (c + d + e + f)_p, \tag{5.5}$$

which contribute $+7$ and the six θ graphs

$$(a, c + f, d + e)_\theta, \quad (b, c + e, d + f)_\theta,$$
$$(c, a + f, b + e)_\theta, \quad (d, a + e, b + f)_\theta, \tag{5.6}$$
$$(e, a + d, b + c)_\theta, \quad (f, a + c, b + d)_\theta,$$

which contribute -6. Therefore

$$k\text{-wt of } (a, b; c, d; e, f)_\alpha = +2. \tag{5.7}$$

Theorem IX effects a great reduction in the number of individual k-wts that need be worked, and we now state the k-wts of all stars with $C(S) = 3$:

$$\begin{aligned} &\alpha \text{ graphs} \quad k\text{-wt} = +2, \\ &\beta \text{ graphs} \quad k\text{-wt} = +1, \\ &\gamma \text{ graphs} \quad k\text{-wt} = +1, \\ &\delta \text{ graphs} \quad k\text{-wt} = +1. \end{aligned} \tag{5.8}$$

For $C(S) = 4$ there are 17 topological types and we list these, together with their weak weights in the Appendix.

In practice the determination of the weight of a graph from (5.1) becomes heavy as the cyclomatic number increases. Numerous results can be established to effect a reduction in the calculation, and we quote two theorems,[8] which we apply in a subsequent paper to the problem of high-density expansions.

Theorem X: If S is a planar topological star, that is, a planar star drawn to conform with the planar condition, and $\sum^*$ denotes summation over all substars that are not finite faces, then

$$\sum_{i\geq 2}^{*} (s_i; S)k_i = 0. \tag{5.9}$$

To avoid a special notation we adopt the convention that the restriction $\sum^*$ imposed on the summation implies the corresponding restriction on the lattice constants. Thus in (5.9) the constant $(s_i; S)$ is the constant for the embeddings of s_i that are not finite faces. Likewise, in the next theorem $\sum_B$ implies that $(s_i; S)$ is the constant for embeddings that contain the boundary.

Proof: The result follows from Euler's law of the edges which states that for a planar topological graph the cyclomatic number is equal to the number of finite faces. Each of these has weight unity in (5.1), and therefore the total contribution from stars that are not finite faces must be zero.

Theorem XI: If S is a planar topological star and $\sum_B$ denotes summation over all substars which contain the contour of the infinite face of S (the boundary of S), then

$$\sum_{i\geq 2}{}_B (s_i; S)k_i = 0, \quad C(S) > 1. \tag{5.10}$$

Proof: We show that if the result holds for $C(s_i) < n$ it will hold for $C(S) = n$. It is true for $C(S) = 2$. If S is a planar topological star, then so are all its substars, and we may divide these into mutually disjoint categories by the contours of their infinite faces. The members of any category are the substars of the graph bounded by the contour that contain the contour, and this graph must have cyclomatic number less than $C(S)$ unless the contour is the contour of S. Assuming the result holds for $C(s_i) < C(S)$, the contribution from each category is zero unless the contour reduces to a finite face. If we exclude these,

$$\sum^* (s_i; S)k_i = 0 \tag{5.11}$$

by the previous theorem, and therefore

$$\sum_B^* (s_i; S)k_i = 0, \tag{5.12}$$

but the asterisk, which excludes finite faces from the summation, is now redundant since $C(S) > 1$ and no finite face contains the boundary. Thus the result is proved.

As an example we can now simplify the calculation of the k-wt of the tetrahedral graph $(a, b; c, d; e, f)_\alpha$. The subgraphs which contain the boundary are: the graph itself and

$$\begin{aligned} &(a + d + e)_p && \text{contribution } +1 \\ &\left.\begin{aligned} &(a, c + f, d + e)_\theta \\ &(d, a + e, b + f)_\theta \\ &(e, a + d, c + b)_\theta \end{aligned}\right\} && \text{contribution } -3, \end{aligned}$$

[8] In these theorems we use the terms star and substar in place of multiply connected graph and multiply connected subgraph. It is clear from the context that the bond is not intended.

and therefore by (5.10)

$$k\text{-wt of } (a, b; c, d; e, f)_\alpha = +2. \qquad (5.13)$$

For completeness we now state the theorem for weak weights which corresponds to Theorem VI for strong weights.

Theorem XII: If G is a graph with a cut set of n vertices, and further the section graph G''' which has these n vertices as vertex set is a complete graph of n vertices, then with G' and G'' defined as in Theorem VI if k_G, $k_{G'}$, $k_{G''}$, $k_{G'''}$ denote the weak weights of G, G', G'', G''', respectively:

$$k_{G'''}k_G = k_{G'}k_{G''}. \qquad (5.14)$$

We omit the proof of this theorem since the result is most easily established by techniques we shall describe in a subsequent paper; the direct proof is long. It is a result of great practical use. For example, the k-wt of the graph formed by placing a tetrahedron on one of the triangular faces of a square-based pyramid as drawn,

(5.15)

is found by taking G' as the pyramid ($k = -3$), G'' as the tetrahedron ($k = 2$), and G''' as the triangle they have in common ($k = 1$), and therefore from (5.14), $k = -6$.

The scope of Theorem XII is much extended by Theorem IX, and all homeomorphs of (5.15) will have $k = -6$. A particularly useful application is to the large class of graphs which have a cut-set with $n = 2$, together with their homeomorphs. Then g''' is the bond, and the required weight is just the product $-k_G k_{G''}$. Thus the weight of a θ-graph is obtained as -1 since G' and G'' are polygons. The weights for β- and γ-graphs follow from the product of a polygon and a θ-graph.

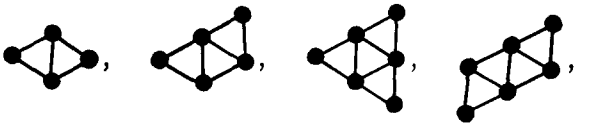
Theorems X and XI can be extended to K-wts, but in the present paper we do not elaborate further the theory of strong weights since the theory of weak weights is the more elegant and more generally useful.

6. EXPANSION FOR $K(p; G)$ FOR THE TRIANGULAR LATTICE (SITE PROBLEM)

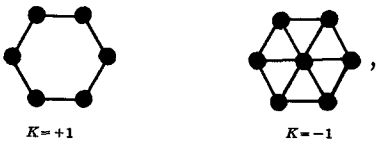
For an infinite graph we neglect edge effects and denote by $k(p; G)$ the mean number of clusters per site. The first contributions to a low-density expansion, valid for small p, come from the site, the bond, and the triangle. Denoting the triangular lattice by T, we have

$$k(p; T) = p - 3p^2 + 2p^3 + \cdots . \qquad (6.1)$$

To extend this series we examine contributions from star graphs with more than three sites. The number of stars embeddable in the triangular lattice increases very rapidly with the number of sites, but many of these have zero K-wt. This is because of the presence of a cut-set of the special type described in Theorem VI; such cut-sets are of frequent occurrence on the triangular lattice. Thus for example the graphs



all have zero K-wt. There are no strong embeddings of $(4)_p$ or $(5)_p$ and the next contributions to (6.1) arise from



of which there are one per site, respectively, to give $+p^6 - p^7$. We summarize in Table I the graphs with 8, 9, and 10 sites required to extend (6.1) to the term in p^{10} and so obtain

$$k(p; T) = p - 3p^2 + 2p^3 + p^6 - p^7 + 3p^8 - 4p^9 + 9p^{10} + \cdots . \qquad (6.2)$$

Further coefficients may be derived by direct enumeration, but this particular series is more readily extended by exploiting the exact matching relation which holds for the triangular lattice (I Sec. 3). That (6.2) is correct may be verified by comparison with Eq. (3.4) of I. For three-dimensional lattices no such matching relations exist, and the methods of this section enable the corresponding expansions for mean number for the simple cubic, body-centered cubic, and face-centered cubic lattices to be derived. We shall show in a subsequent paper that these expansions are of use in extending series expansions for the mean size of clusters at low densities.

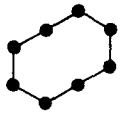
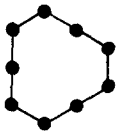
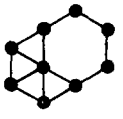
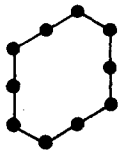
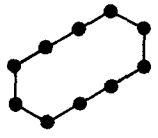
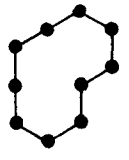
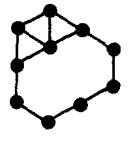
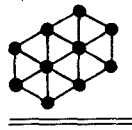
7. EXPANSION FOR $\bar{K}(\bar{p}; G)$ FOR THE TRIANGULAR LATTICE (BOND PROBLEM)

For an infinite lattice we derive the expansion for the mean number of bond clusters *per site.* The expansion for the triangular lattice starts with the

contribution from the graphs of type (4.5) which, per site, amounts to

$$3\bar{p} - 15\bar{p}^2 + 20\bar{p}^3 - 15\bar{p}^4 + 6\bar{p}^5 - \bar{p}^6. \qquad (7.1)$$

TABLE I. Some graphs on the triangular lattice that contribute to $k(p; T)$

Graph[a]	Number (per site)	K-weight	Contribution
	3	+1	$+3p^8$
	2	+1	$+2p^9$
	6	−1	$-6p^9$
	3	+1	$+3p^{10}$
	3	+1	$+3p^{10}$
	6	+1	$+6p^{10}$
	6	−1	$-6p^{10}$
	3	+1	$+3p^{10}$

[a]We illustrate the individual space-types.

TABLE II. Summary of stars on the triangular lattice and their contributions to $k(p; T)$.

	k-wt	$l = 7$	$l = 8$	$l = 9$	$l = 10$
polygons (Jordan curves)	+1	42	123	380	1212
θ-graphs	−1	42	165	609	2283
α-graphs	+2	0	0	20	120
β-graphs	+1	0	6	54	375
γ-graphs	+1	6	36	162	666
δ-graphs	+1	0	0	0	6
F-graphs	−3	0	0	0	15
H-graphs	−1	0	0	0	24
K-graphs	−1	0	0	6	48
L-graphs	−1	0	0	0	12
N-graphs	−1	0	0	6	48
O-graphs	−1	0	0	2	12
Contribution:		$+6\bar{p}^7$	$+0\bar{p}^8$	$+13\bar{p}^9$	$+27\bar{p}^{10}$

TABLE III. Weak lattice constants for the triangular lattice.

θ-graphs.	$l = 7$	$(1,2,4)_\theta$ =	30
		$(1,3,3)_\theta$ =	12
	$l = 8$	$(1,2,5)_\theta$ =	96
		$(1,3,4)_\theta$ =	60
		$(2,2,4)_\theta$ =	6
		$(2,3,3)_\theta$ =	3
	$l = 9$	$(1,2,6)_\theta$ =	312
		$(1,3,5)_\theta$ =	168
		$(1,4,4)_\theta$ =	69
		$(2,2,5)_\theta$ =	30
		$(2,3,4)_\theta$ =	30
	$l = 10$	$(1,2,7)_\theta$ =	1068
		$(1,3,6)_\theta$ =	516
		$(1,4,5)_\theta$ =	378
		$(2,2,6)_\theta$ =	132
		$(2,3,5)_\theta$ =	126
		$(2,4,4)_\theta$ =	63
α-graphs	$l = 9$	$(1,1;1,1;1,4)_\alpha$ =	6
		$(1,1;1,2;1,3)_\alpha$ =	12
		$(1,2;1,2;1,2)_\alpha$ =	2
	$l = 10$	$(1,1;1,1;1,5)_\alpha$ =	24
		$(1,1;1,2;1,4)_\alpha$ =	48
		$(1,1;1,3;1,3)_\alpha$ =	24
		$(1,2;1,2;1,3)_\alpha$ =	24
β-graphs	$l = 8$	$(1,2;1,2;1,1)_\beta$ =	6
	$l = 9$	$(1,2;1,2;1,2)_\beta$ =	30
		$(1,2;1,3;1,1)_\beta$ =	24
	$l = 10$	$(1,2;1,2;1,3)_\beta$ =	102
		$(1,2;1,2;2,2)_\beta$ =	51
		$(1,2;1,3;1,2)_\beta$ =	120
		$(1,2;1,4;1,1)_\beta$ =	60
		$(1,2;2,2;1,2)_\beta$ =	12
		$(1,2;2,3;1,1)_\beta$ =	6
		$(1,3;1,3;1,1)_\beta$ =	24
γ-graphs	$l = 7$	$(1,2;1,2;1)_\gamma$ =	6
	$l = 8$	$(1,2;1,2;2)_\gamma$ =	12
		$(1,2;1,3;1)_\gamma$ =	24
	$l = 9$	$(1,2;1,2;3)_\gamma$ =	30
		$(1,2;1,3;2)_\gamma$ =	48
		$(1,2;1,4;1)_\gamma$ =	60
		$(1,3;1,3;1)_\gamma$ =	24
	$l = 10$	$(1,2;1,2;4)_\gamma$ =	96
		$(1,2;1,3;3)_\gamma$ =	108
		$(1,2;1,4;2)_\gamma$ =	108
		$(1,2;1,5;1)_\gamma$ =	168
		$(1,2;2,2;3)_\gamma$ =	12
		$(1,2;2,3;2)_\gamma$ =	12
		$(1,2;2,4;1)_\gamma$ =	12
		$(1,3;1,3;2)_\gamma$ =	42
		$(1,3;1,4;1)_\gamma$ =	108
δ-graphs	$l = 10$	$(1,2,2,5)_\delta$ =	6

Up to six bonds the only star graphs are the following:

Graph	k-weight	Number	Contribution
$(3)_p$	$+1$	2	$+ 2\bar{p}^3$
$(4)_p$	$+1$	3	$+ 3\bar{p}^4$
$(5)_p$	$+1$	6	$+ 6\bar{p}^5$
$(1, 2, 2)_p$	-1	3	$- 3\bar{p}^5$
$(6)_p$	$+1$	15	$+15\bar{p}^6$
$(1, 2, 3)_p$	-1	12	$-12\bar{p}^6$

and on adding these contributions to (7.1)

$$\bar{k}(\bar{p}; T) = 3\bar{p} - 15\bar{p}^2 + 22\bar{p}^3 - 12\bar{p}^4 + 9\bar{p}^5 + 2\bar{p}^6 + \cdots . \quad (7.2)$$

We have extended (7.2) by enumerating every star graph on the triangular lattice with 7, 8, 9, and 10 lines, thus adding the terms

$$+6\bar{p}^7 + 0\bar{p}^8 + 13\bar{p}^9 + 27\bar{p}^{10}. \quad (7.3)$$

We give in Table II a summary of the contributions from different topological types, and in Table III we list the lattice constants for all graphs of cyclomatic number 3 and 4.

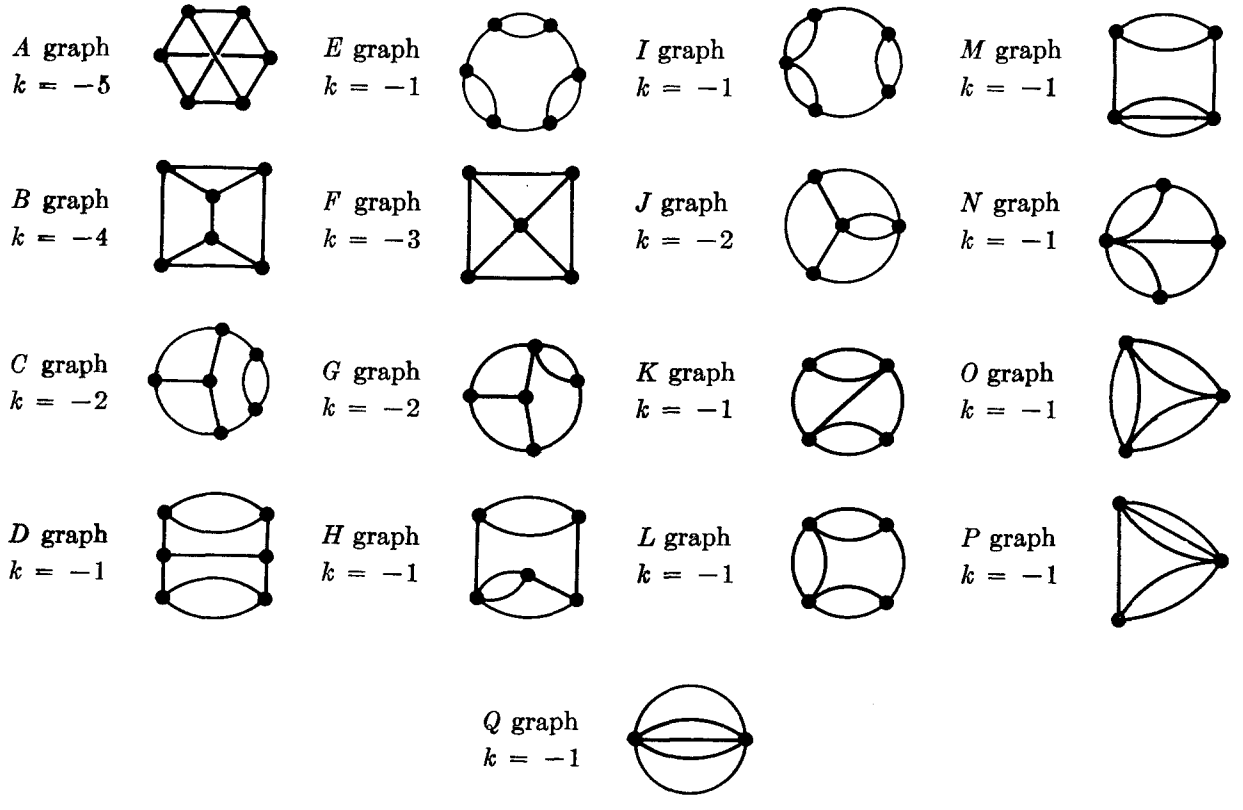
In an analogous manner expansions can be obtained for other lattices. The graphs required are numerous but not too difficult to count in the *weak* system on a computer .The weight problem is made manageable by the results of Sec. 5. We shall subsequently apply these series developments to a study of the mean number function and also to extending the corresponding mean size series.

We have verified (7.3) by deriving the high-density expansion for the matching lattice; that is, the high-density expansion for the bond problem on the honeycomb lattice.

ACKNOWLEDGMENTS

We are grateful to Professor C. Domb for suggesting the problem and to Professor M. E. Fisher for much constructive criticism.

APPENDIX: THE 17 TOPOLOGICAL TYPES OF STAR WITH $C(S) = 4$



The Enumeration of Homeomorphically Irreducible Star Graphs

B. R. HEAP
Mathematics Division, National Physical Laboratory, Teddington, Middlesex, England
(Received 21 February 1966)

In expressing properties of interacting systems in the form of series, in many cases only summation over star graphs is involved. The identification and classification of such graphs is simplified by reducing them to homeomorphically irreducible stars. These graphs can be regarded as being all the possible different topological types of star. A method is described which has been used to produce all the different homeomorphically irreducible stars which have cyclomatic numbers ≤ 5. In particular, it has been established that there are 118 such graphs with cyclomatic number 5, 111 of which are planar. Diagrams of these graphs are appended, and a table of their k weights, which are needed for obtaining series for percolation processes, is also given. An extension of the method has been used to count the numbers of homeomorphically irreducible stars containing up to 8 vertices and 12 edges, a table of which is given.

1. INTRODUCTION

A RECENT paper[1], hereafter referred to as I, discussed the application of graph theory to problems involving the obtaining of series expansions needed for the investigation of the statistical mechanics of interacting systems on crystal lattices. In a subsequent paper,[2] Essam and Sykes used some of the concepts introduced in I in order to obtain many general theorems for the problem of obtaining series expansions for percolation processes. For a full bibliography of this and other similar problems, the reader is referred to the reviews of Domb,[3] Fisher,[4] Uhlenbeck and Ford,[5] and to the papers mentioned previously. In many cases, the series expansion is essentially equivalent to a summation with appropriate weights over a restricted class of linear graphs known as stars. The problem of the classification of such graphs was considered in I, where it was shown that the classification could be simplified by reducing the stars to simpler graphs which we refer to as homeomorphically irreducible stars. The enumeration of homeomorphically irreducible graphs is one of the unsolved problems in the field of the counting of linear graphs.[6] In this paper we describe a method which has been used to enumerate and obtain all homeomorphically irreducible stars containing up to 8 vertices and 12 edges. A knowledge of these graphs will allow percolation series to be extended to higher orders.

We begin by introducing the graph-theoretical definitions and concepts needed for this work.

2. GRAPH-THEORETICAL CONCEPTS

A *graph* G consists of a set S of *vertices* together with (undirected) *edges* joining some of the vertices. A pair of vertices may be joined by more than one edge (*parallel edges*), but we do not allow edges which join a vertex to itself (*loops*). A graph is said to be *disconnected* if it is possible to divide the vertex set S into two disjoint subsets S_1, S_2 such that there are no edges joining any vertex εS_1 to any vertex εS_2; otherwise, the graph is said to be *connected*. A *cut vertex* of a connected graph G is a vertex εS whose removal from G together with its incident edges disconnects the graph. A connected graph containing no cut vertices is called *multiply-connected* and is usually referred to as a *star*.

The number of edges incident at a vertex of a graph is known as the *degree* of the vertex. Two graphs are *isomorphs* if there is a one-to-one correspondence between their vertex sets such that the corresponding vertices are joined by $m \geq 0$ edges in one if and only if they are so joined in the other. If a graph G contains a vertex of degree 2, A say, which is joined by edges to vertices B, C, the process of removing A from S together with the removal of edges AB, AC and the addition of an edge BC is known as the *suppression* of the vertex A. (It is possible that the suppression of a vertex leads to a graph containing a loop which is forbidden by our definition of a graph. It will become evident that this cannot happen for the special graphs that we are to consider in this work.) The reverse process of replacing an edge BC by a new vertex A and

[1] M. F. Sykes, J. W. Essam, B. R. Heap, and B. J. Hiley, J. Math. Phys. 7, 1557 (1966).
[2] J. W. Essam, and M. F. Sykes, J. Math. Phys. 7, 1573 (1966).
[3] C. Domb, Phil. Mag. Suppl. 9, 149 (1960).
[4] M. E. Fisher, J. Math. Phys. 4, 278 (1963).
[5] G. E. Uhlenbeck and G. W. Ford, *Studies in Statistical Mechanics*, J. de Boer and G. E. Uhlenbeck, Eds. (North-Holland Publishing Co., Inc., Amsterdam, 1962), Vol. 1, Part B.
[6] F. Harary, *Applied Combinatorial Mathematics*, F. Beckenbach, Ed. (John Wiley & Sons, Inc., New York, 1964), Chap. 6.

joining A, B and A, C by edges is known as the *insertion* of the (second-degree) vertex A. We then say that two graphs are *homeomorphs* if by means of the suppression and insertion of second-degree vertices one can be transformed into an isomorph of the other. Thus, homeomorphs can be regarded as having the same basic topology. A graph which contains no vertex of degree 2 (i.e., from which no vertices can be suppressed) is referred to as *homeomorphically irreducible*. (We normally abbreviate this to HI.)

As was mentioned in Sec. 1, in the applications of graph theory to physical problems we can normally restrict attention to stars. Accordingly, in the remainder of this paper we are concerned solely with such graphs. Let a given star contain n vertices and m edges. The *cyclomatic number* ν of the graph is defined by

$$\nu = m - n + 1. \tag{1}$$

Since the insertion of a second-degree vertex on an edge of such a star increases both m and n by unity, the cyclomatic number is unchanged. Thus, graphs which are homeomorphic have the same cyclomatic number. (The converse is not necessarily true.) It was shown in I that stars having small values of ν could conveniently be classified by specifying their basic topology, i.e., by suppressing all vertices of degree 2, thereby producing a *homeomorphically irreducible star* (HI star). In order to achieve the classification, it is necessary to enumerate all the different HI stars for small values of ν. For $\nu = 2$ there is only one such graph, for $\nu = 3$ there are 4 such graphs, and Essam and Sykes[2] have shown that there are 17 such graphs with $\nu = 4$ (these are illustrated in the Appendix of their paper). In Sec. 3 we describe a method for the evaluation of such graphs which has been used principally to enumerate the HI stars having cyclomatic number 5.

3. ENUMERATION OF HOMEOMORPHICALLY IRREDUCIBLE STARS

All the different HI stars having cyclomatic number ν ($\nu \geq 3$) can be obtained from those having cyclomatic number $\nu - 1$ by running through all possible ways of:

(a) joining any two distinct vertices by an edge;
(b) inserting a second-degree vertex on any edge and then joining it to any other vertex by an edge;
(c) inserting two second-degree vertices on any two (not necessarily distinct) edges and joining them by an edge.

That this procedure produces HI stars having cyclomatic number ν can be seen from Eq. (1), since in (a) m is increased by unity and n remains unchanged; in (b) m is increased by 2 and n by 1; and in (c) m is increased by 3 and n by 2. It is clear that the above procedure generates all the possible HI stars with cyclomatic number ν at least once. This method allows us to produce the graphs having $\nu \leq 4$ without too much difficulty just by using pencil and paper. However, when we consider the case $\nu = 5$ a more automatic process is required, since in running through the above procedure each possible HI star may be produced several times and the recognition and identification of a graph becomes extremely hazardous. Accordingly, we have used a computer to produce these graphs.

There are two main problems in connection with the use of a computer to enumerate graphs, namely the actual representation of a graph in the computer and the identification of a particular graph amongst a dictionary of possibly many graphs. A convenient method of representing a graph is by means of its *adjacency matrix*. If the vertices of a graph having n vertices are labeled 1, 2, $\cdots$, n, the adjacency matrix $A = (a_{ij})$ is the $n \times n$ matrix in which

$$a_{ij} = \text{number of edges joining vertex } i \text{ to vertex } j. \tag{2}$$

However, in order to produce this matrix it is first necessary to label the vertices—this can conveniently be done by assigning the label 1 to the vertex having the highest degree; the label 2 to the vertex having the next highest degree; etc. Clearly this labeling is not necessarily unique.

In identifying a particular graph, we compare its adjacency matrix with those of all previously known graphs having the same set of degrees. Should the graph not be so identified, we interchange two rows and columns of A (i.e., interchange the labels of two vertices of the graph) and repeat the comparison process. The interchanging is done according to a scheme which preserves the set of degrees as before and which ultimately runs through all possible labelings of the graph consistent with the labeling scheme. The method of doing this is an extension of a scheme of the author's[7] which runs through all possible permutations of N objects by means of interchanges.

By forming the graphs according to the scheme mentioned above and then using the identification procedure, we obtain the HI stars having $\nu = 5$ and find that there are 118 such graphs. Diagrams

[7] B. R. Heap, Computer J. 6, 293 (1963).

TABLE I. The k weights of the homeomorphically irreducible stars with cyclomatic number 5.

Graph number	k weight	Graph number	k weight	Graph number	k weight	Graph number	k weight
1	1	31	1	61	2	91	2
2	1	32	1	62	2	92	4
3	1	33	1	63	2	93	4
4	1	34	1	64	1	94	5
5	1	35	1	65	2	95	1
6	1	36	2	66	2	96	2
7	1	37	2	67	2	97	3
8	1	38	2	68	1	98	3
9	2	39	1	69	4	99	6
10	1	40	3	70	4	100	4
11	1	41	4	71	5	101	7
12	1	42	1	72	1	102	8
13	1	43	1	73	2	103	1
14	2	44	2	74	3	104	1
15	1	45	1	75	3	105	1
16	1	46	2	76	2	106	2
17	2	47	3	77	6	107	2
18	1	48	4	78	5	108	1
19	1	49	1	79	4	109	2
20	1	50	1	80	1	110	2
21	1	51	2	81	1	111	4
22	2	52	1	82	2	112	4
23	1	53	1	83	1	113	5
24	1	54	1	84	1	114	4
25	1	55	1	85	1	115	8
26	1	56	2	86	2	116	10
27	2	57	1	87	2	117	11
28	1	58	1	88	1	118	12
29	3	59	1	89	2		
30	2	60	1	90	2		

of these graphs are given in the Appendix. A more detailed discussion of the computer program is given in National Physical Laboratory Mathematics Division Report No. Ma. 57, which is available from the author.

In their work on series expansions for percolation problems, Essam and Sykes showed that, in the most useful case, all homeomorphs enter the series with the same weight (the k weight). The k-weight $k(G)$ of a graph G is defined recursively by

$$\sum_{g \subseteq G} k(g) = \nu, \tag{3}$$

where $g \subseteq G$ denotes any subgraph of G which is a star, and the summation extends over all such subgraphs.[2] (For a full discussion of k weights the reader is referred to the aforementioned paper.) Most of the 118 HI stars contain a pair of vertices which are joined by more than one edge—for such a graph, Theorem XII of Essam and Sykes gives the k weight immediately as -1 times the k weight of the graph which is obtained by removing one of the parallel edges. However, in 14 cases no such parallel edges occurred and the k weights were derived using (3). The k weights are given in Table I.

The enumeration of the graphs took approximately two hours using the English Electric-Leo KDF9 computer of the National Physical Laboratory. The enumeration of the HI stars having cyclomatic number 6 would not be feasible using the program outlined above, since the identification of graphs having 10 vertices, each of which has degree 3 would take an inordinate length of time. However, the use of different techniques for the identification of a graph which are at present under investigation should allow such graphs to be enumerated in the future.

As was mentioned in Sec. 1, the numbers of homeomorphically irreducible graphs having n vertices and m edges are unknown except for trivial cases. However, by using the methods mentioned above we have been able to obtain the numbers of HI stars which contain up to 8 vertices and 12 edges. For example, HI stars having n vertices and m edges can be formed (i) from those having n vertices and $m - 1$ edges by using procedure (a); (ii) from those having $n - 1$ vertices and $m - 2$ edges by using (b); and (iii) from those having $n - 2$ vertices and $m - 3$ edges by using (c). It is thus necessary to run through these three possible ways of forming the graphs. The numbers of the graphs are given in Table II.

It is also of interest to examine which of the HI

TABLE II. The numbers of homeomorphically irreducible stars having $n \leq 8$ vertices and $m \leq 12$ edges.

n	2	3	4	5	6	7	8
m							
3	1						
4	1						
5	1	1					
6	1	2	2				
7	1	3	5				
8	1	4	13	4			
9	1	6	26	24	5		
10	1	7	47	84	38		
11	1	9	78	233	216	23	
12	1	11	126	557	914	314	16

stars having a particular value of ν are planar since these are the different ways of dividing up the plane into $\nu + 1$ regions (with the condition that the ensuing graph is a star). For $\nu = 2, 3, 4, 5$ there are 1, 4, 16, 111 such graphs, respectively.

ACKNOWLEDGMENTS

The work reported in this paper forms a part of the research program of the National Physical Laboratory and the paper is published by permission of the Director of the Laboratory. The author would like to thank Dr. M. F. Sykes of King's College, London, for bringing this problem to his attention and for much valuable discussion. He is also indebted to Dr. J. L. Martin for reading a draft copy of the manuscript, and to Miss J. Wade for providing the drawings illustrated in the Appendix.

APPENDIX

We append here in Tables III, IV, and V diagrams of all HI stars with cyclomatic numbers $\nu \leq 5$. Those with $\nu \leq 4$ were previously illustrated by Essam and Sykes[2] and are included here for completeness. For $\nu = 2$, graph 1 in the list corresponds to the θ-graph of Essam and Sykes. For $\nu = 3$, graphs 1, 2, 3, 4 correspond to the δ, γ, β and α-graphs, while for $\nu = 4$, graphs 1, 2, $\cdots$, 17 correspond to the Q, P, $\cdots$, A graphs.

The vertices of the graphs have been labeled 1, 2, $\cdots$ according to an extension of the method described in Sec. 3 which was used in the computer program. The graphs have been ordered and assigned labels 1, 2, $\cdots$ according to an extension of a scheme suggested by Nagle.[8] Full details of these schemes are given in the report mentioned above.

TABLE III. Homeomorphically irreducible stars with cyclomatic numbers 2 and 3.

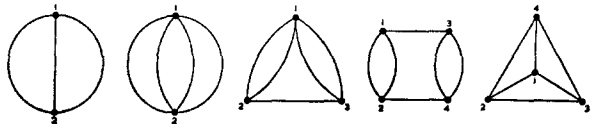
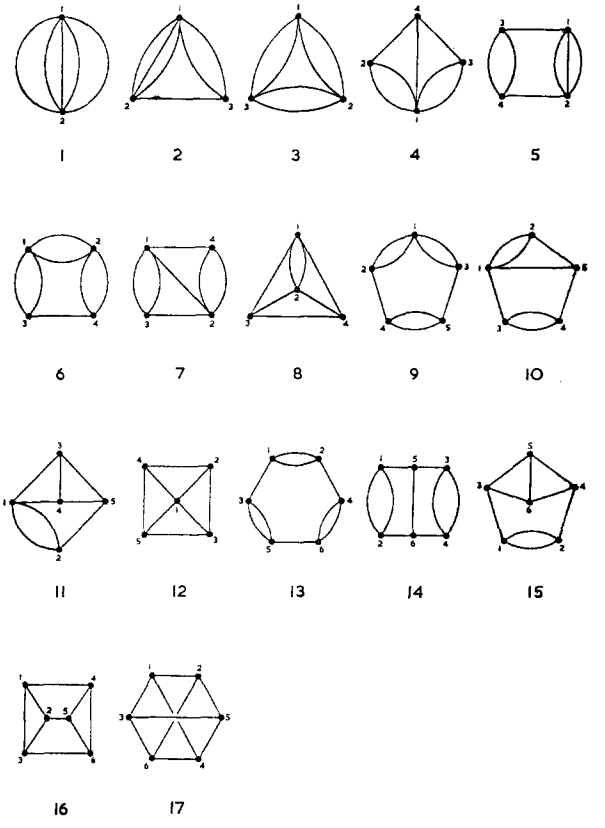
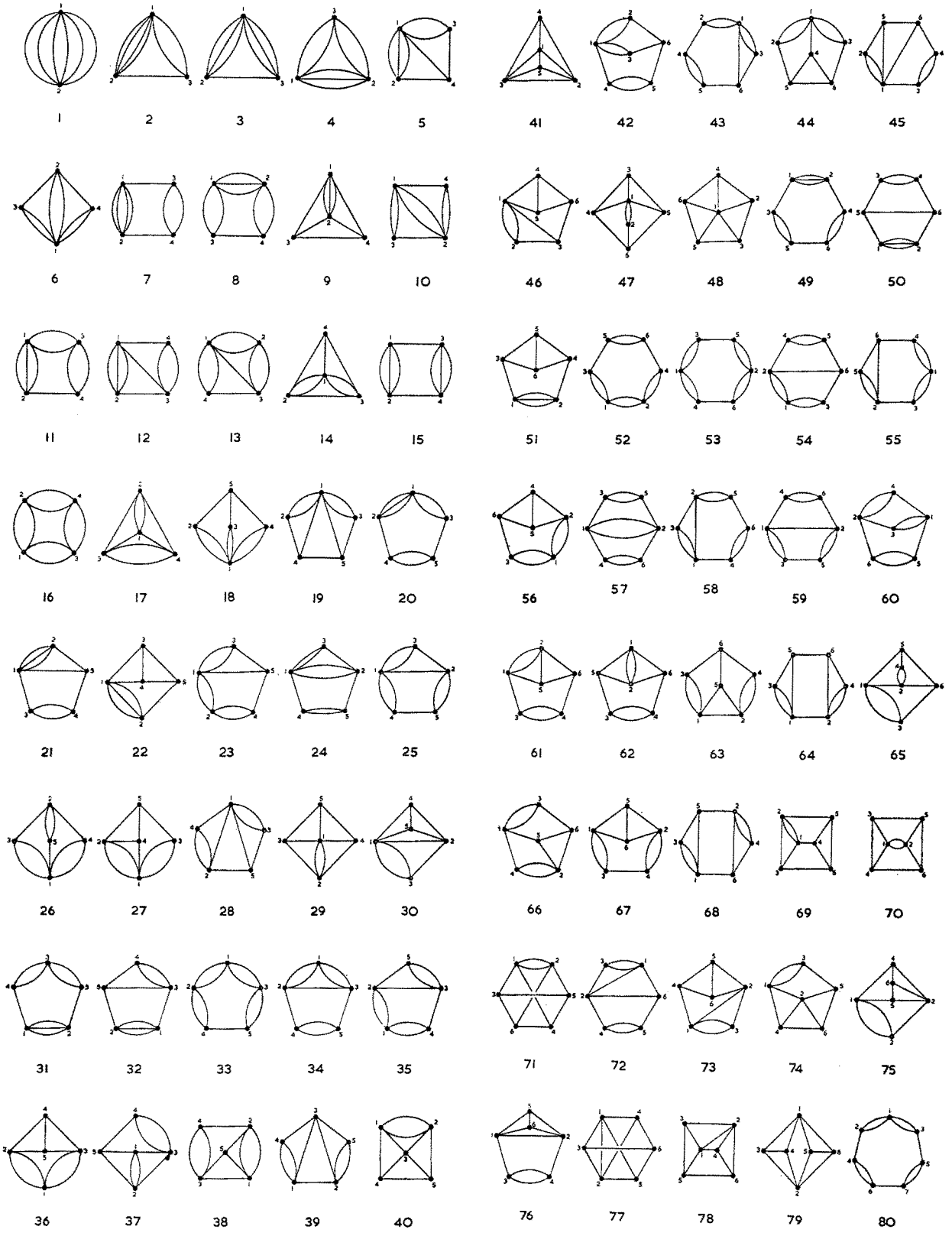


TABLE IV. Homeomorphically irreducible stars with cyclomatic number 4.

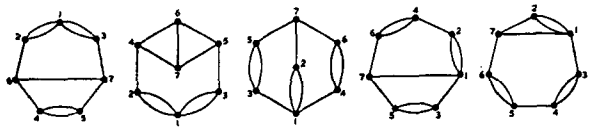


[8] J. F. Nagle, J. Math. Phys. 7, 1588 (1966).

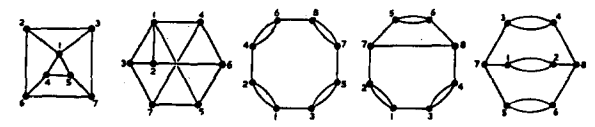
TABLE V. Homeomorphically irreducible stars with cyclomatic number 5.



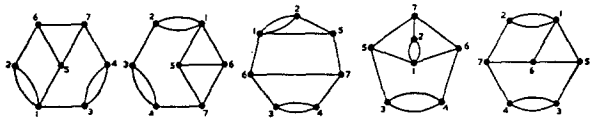
(TABLE V—Continued)



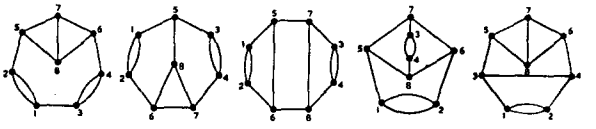
81 82 83 84 85



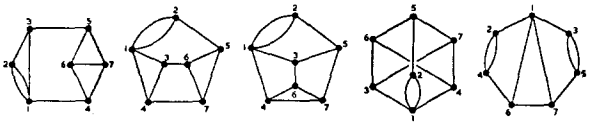
101 102 103 104 105



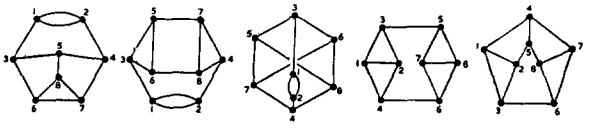
86 87 88 89 90



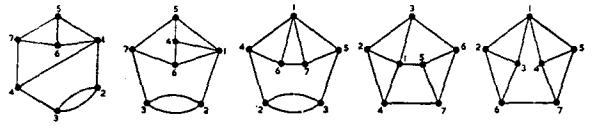
106 107 108 109 110



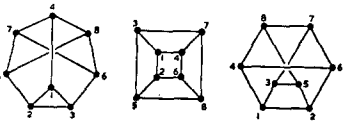
91 92 93 94 95



111 112 113 114 115



96 97 98 99 100



116 117 118

On Ordering and Identifying Undirected Linear Graphs

JOHN F. NAGLE*
Wheatstone Physics Laboratory, King's College, London, England
(Received 17 March 1966)

A general linear ordering relation is presented which completely orders any subset of the undirected linear graphs with the same number of vertices. The discussion is then specialized to the ordering and the identification of the unlabeled stars with no vertices of degree two, which faithfully represent the basic topologies of all stars.

INTRODUCTION

THE study of linear graphs is important to mathematical physics because graphs frequently provide a convenient representation of the terms in series (or perturbation) expansions which are used in discussing otherwise intractable problems. Particular problems in statistical mechanics which may be cited are the Mayer cluster expansion for the virial series of a gas,[1] the Ising-model series expansions,[2-4] the percolation problem,[5] the excluded volume problem,[6] and the residual entropy of ice problem.[7]

When it becomes necessary to consider complex linear graphs certain problems arise. First, it becomes difficult to obtain all the graphs of the type wanted. Secondly, there is the problem of identifying a graph with its isomorph in a list. The technique of drawing the graph in several different ways with the hope of making it look like one in the list is very tedious for even moderately complex graphs. The solution to the identification problem is related to the solution to the next problem. This is the question of how to order the graphs conveniently in a numbered list in such a way as to maximize the usefulness of the list to a large number of research workers.

Despite the obvious advantages of adopting a systematic ordering and notation for linear graphs, there have been two excellent reasons for not doing so. The first is that different problems have different "natural" groupings of linear graphs. For example, most problems require only well-defined subsets of linear graphs, such as the Mayer virial series for a gas which requires only the multiply connected (no articulation vertices) and unlabeled graphs called stars.[1] Furthermore, within these subsets different gross criteria are important for different problems. For example, for the high-temperature Ising-model series the number of edges e is of more importance than the number of vertices v, while the reverse is true for the residual entropy of ice series.[7] Other criteria for classification are the cyclomatic number; the degree of each vertex (number of edges incident to the vertex); the number of double, triple, $\cdots$, edges; and so on.

The second reason for not adopting a systematic ordering of linear graphs, even for use with specialized problems, is the difficulty in finding a completely systematic ordering. For example, if one uses any combination of the criteria mentioned above one finds that eventually there are distinct graphs not differentiated by the criteria which must then be ordered arbitrarily. This ultimate arbitrariness has undermined the effort to organize graph data in a systematic way.

The primary purpose of this paper is to show in Sec. I how the "ultimate arbitrariness" in ordering linear graphs can be resolved in a conceptually simple way through the use of what we will call the *canonical matrix* of a graph. Indeed, all undirected linear graphs with v vertices could be ordered using only the canonical matrix. However, the considerations mentioned above concerning the natural groupings of graphs for particular problems show that this is an unwise program. Rather, it is felt that some "initial arbitrariness" in ordering graphs ahould remain to suit the particular problem in hand. In this spirit, the more modest program of ordering the stars with no vertices of degree two is discussed in Sec. II as an example of what can be done for graphs relevant to particular problems.

* Present address: Department of Chemistry, Cornell University, Ithaca, N. Y.

[1] G. E. Uhlenbeck and G. W. Ford, *Studies in Statistical Mechanics*, J. De Boer and G. E. Uhlenbeck, Eds. (North-Holland Publishing Company, Amsterdam, 1962), Vol. 1, p. 123, and Proc. Natl. Acad. Sci. U. S. 42, 122, 203, 529 (1956).
[2] C. Domb, Advan. Phys. 9, 150 (1960).
[3] M. E. Fisher, J. Math. Phys. 4, 278 (1963).
[4] M. F. Sykes, J. W. Essam, B. R. Heap, and B. J. Hiley, J. Math. Phys. 7, 1557 (1966).
[5] J. W. Essam and M. F. Sykes, J. Math. Phys. 7, 1573 (1966).
[6] M. F. Sykes, J. Math. Phys. 2, 52 (1961).
[7] J. F. Nagle, J. Math. Phys. 7, 1484 (1966).

I. A LINEAR ORDER RELATION FOR GRAPHS WITH V VERTICES

A complete (linear) order relation[8] on a set X is defined as a binary relation, $\succ$, such that for x, y, and z belonging to X, we have

(1) $x \succ y$, and $y \succ z$ implies $x \succ z$,
(2) $x \succ y$, and $y \succ x$ implies $x = y$,
(3) $x \neq y$ implies either $x \succ y$ or $y \succ x$.

The matrix associated with a graph (adjacency matrix) is defined as the square array (a_{ij}), where a_{ij} is the number of edges joining the ith and jth vertices.[9] In general, there are many matrices associated with an unlabeled graph because the vertex numberings can be permuted. A procedure will be described which selects one of these matrices which then is defined as the canonical matrix of the graph.

Let us define an order relation for $v \times v$ matrices. Let $A = (a_{ij})$ and $B = (b_{ij})$. Then, we scan the two matrices row by row, comparing a_{ij} to b_{ij}, and stop the first time that $a_{ij} \neq b_{ij}$. If $a_{ij} > b_{ij}$, then we say that A precedes B. More formally, let us define $(m, n) < (i, j)$ if and only if $[(m < i)$ or $(m = i$ and $n < j)]$. Then, A precedes B if and only if there exists *no* (i, j) such that $\{(b_{ij} > a_{ij})$ and $[a_{mn} = b_{mn}$ for all $(m, n) < (i, j)]\}$. Clearly, either (A precedes B) or (B precedes A). If (A precedes B) and (B precedes A), then ($A = B$). Finally, it is trivial to verify that if (A precedes B) and (B precedes C), then (A precedes C). Thus, this definition produces a linear ordering of $v \times v$ matrices.

It is now possible to define the canonical matrix of a graph as that matrix associated with the graph which precedes all others. Since the number of matrices associated with a finite graph is finite, this definition is valid. However, all the matrices associated with the graph may not be acceptable candidates for the canonical matrix. For example, if graphs with one colored vertex and $v - 1$ ordinary vertices were under consideration, it would be natural to require that the colored vertex be labeled 1. This imposes a restriction which determines the acceptable matrices associated with the graph. Therefore, it is suitable to define the (restricted) canonical matrix as that acceptable matrix associated with the graph which precedes all others which also satisfy the restrictions in question. We note that the canonical matrix defines a canonical vertex numbering, modulo symmetry operations of the graph.

[8] J. L. Kelly, *General Topology* (D. Van Nostrand, Inc., Princeton, N. J., 1955), p. 14.

[9] C. Berge, *The Theory of Graphs*, A. Doigt, Translator (John Wiley & Sons, Inc., New York, 1962), p. 130.

The final obvious step is to order the graphs in the subset in question according to the matrix order of the canonical matrices. Since the correspondence between graphs and canonical matrices is one-to-one and the matrix ordering is complete, the corresponding graph ordering will also be complete.[10]

II. ON ORDERING AND IDENTIFYING THE STARS WITH NO VERTICES OF DEGREE TWO

The object of this section is to order the stars (undirected and unlabeled graphs with no articulation vertices[1]) with no vertices of degree two. The reason for singling out the stars with no vertices of degree two is that they faithfully represent the basic topologies of all stars. This follows because any star can be derived from one and only one of the stars with no vertices of degree two under the topologically invariant operation of inserting vertices of degree two. It has been suggested that stars be partially classified first by topological type.[4,11,12] With this use in mind, the stars with no vertices of degree two will be called *star types* in this paper.[13] The star types are useful for obtaining the stars, which are necessary for the calculation of lattice constants, which in turn are required for the Ising-model series expansions and various other physical problems.[2,4,5,7,12]

In an adjoining paper, B. R. Heap has discussed how the star types may be obtained using a computer.[11] In an appendix of Heap's paper[11] is a list of the star types for cyclomatic numbers three, four, and five ordered in the way suggested here. This list should be referred to for examples.

The canonical matrix ordering described in the last section can be applied to completely order any subset of the star types. Therefore, this section is mainly concerned with the pre-ordering of the star types into convenient subsets. Of course, this pre-

[10] We note that the restriction to subsets with a fixed number of vertices could be removed if there were no vertices of degree zero or, alternatively, if there were no loops, which are edges that connect a vertex to itself. In the latter case the adjacency matrix is invertebrate, i.e., $a_{ii} = 0$, and one can alter the matrix definition so that a_{ii} equals the degree of the ith vertex. Then, in both cases a matrix with smaller v could be compared to one with larger v by filling in some zeros. It should also be noted that as the matrix ordering is defined the more complex graphs precede the simpler ones. However, this is the way which is best suited to the application in Sec. II.

[11] B. R. Heap, J. Math. Phys. 7, 1582 (1966).

[12] G. W. Ford and G. E. Uhlenbeck, Proc. Natl. Acad. Sci. U. S. 43, 163 (1957).

[13] An alternative name to star type might be star complex since the topological space of a graph is a complex. A particular triangulation of the complex corresponds to a particular set of vertices of degree two. Unfortunately, the name star has a different meaning in topology, so it seems unwise to combine it with a topological term. Other names used in the literature are homeomorphically irreducible star by Heap[11] and homeomorphic type by Ford and Uhlenbeck.[12]

ordering is subject to the personal preferences of the individual doing the ordering. To obtain an ordering as objective as possible discussions have been held with other workers who use these graphs.

The most conspicuous quantifiable features of the star types are the number of vertices v, the number of edges e, and the cyclomatic number $c = e - v + 1$. Fortunately, it does not matter much which two of these criteria are chosen for ordering purposes, since the star types can easily be rearranged if one prefers a different choice. For the problems considered by Sykes, it is most convenient to arrange the star types first by increasing cyclomatic number and next by increasing number of vertices.[5]

One might consider the criterion of planarity to be more fundamental than the above criteria. However, two considerations eliminate this criterion. First, it is of no use in constructing lists of graphs. Second, for identification it is easier to count v and e than it is to convince oneself that a graph is nonplanar. Also, planarity does not lend itself to computer identification.

Next, a less obvious ordering criterion is introduced. Define the *degree-tuple* of a graph as $(D_1, \cdots, D_i, \cdots, D_v)$ where D_i, for $i = 1, 2, \cdots, v$, is the degree of the ith vertex and the numbering of the vertices is restricted by $i < j$ if $D_i > D_j$. Thus, the degree-tuple of a graph is a v-tuple of nonincreasing numbers, such as (5, 4, 4, 4, 3). Of course, $e = \frac{1}{2}(D_1 + D_2 + \cdots + D_v)$ so the degree- tuple supplies one with v, e, and c. Now, let us define an ordering relation for degree-tuples as $(D_1, \cdots, D_v)$ precedes $(D_1', \cdots, D_v')$ if there exists some i such that $D_i > D_i'$ and $D_j = D_j'$ for $j < i$. Let us agree to order subsets of star types with the same c and v by the degree-tuple ordering. For example, the set of star types with degree-tuple (5, 4, 4, 3) precedes the set with degree-tuple (4, 4, 4, 4) and (6, 5, 3) precedes (6, 4, 4).

The degree-tuple ordering will not be fully justified until the application of the canonical matrix is discussed. However, certain advantages can be mentioned now. The problem of counting the number of two overlapping polygons on a lattice[2] requires that each $D_i \leq 4$, the Ising-model magnetic problem requires precisely two odd D_i,[6] and the residual entropy of ice problem requires all $D_i = 4$.[7] Therefore, in each of these cases knowledge of the degree-tuple eliminates many subsets immediately. Furthermore, there are 38 star types in the subset $c = 5$ and $v = 6$: this is a large enough number to warrant additional labels to facilitate identification.

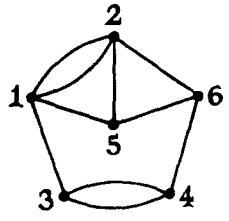


FIG. 1. Star type number 61 with $c = 5$.

So far, little has been said about the identification problem. The identification procedure which accompanies the ordering introduced thus far is: count the degree of each vertex in turn, compose the degree-tuple, and from the degree-tuple compute $e = \frac{1}{2}\Sigma_i D_i$ and $c = e - v + 1$. Now, when one is counting the vertex degrees it is easy to notice additional information concerning the numbers of single, double, triple, etc., edges incident to each vertex. (A double edge consists of two edges connecting the same two vertices.) Let s_i, d_i, $\cdots$, h_i, denote, respectively the number of single, double, $\cdots$, highest multiple, edges incident to the ith vertex. Then, this information may be recorded as follows:

Abstract Form	Star Type 61, $c = 5$ (See Fig. 1)
$\begin{bmatrix} h_1, h_2, \cdots, h_v \\ \vdots \\ d_1, d_2, \cdots, d_v \\ s_1, s_2, \cdots, s_v \end{bmatrix}$	$\begin{bmatrix} 1, 1, 1, 1, 0, 0 \\ 2, 2, 1, 1, 3, 3 \end{bmatrix}$
$(D_1, D_2, \cdots, D_v)$	(4, 4, 3, 3, 3, 3).

Clearly, $s_i + 2d_i + 3t_i + \cdots = D_i$. For the specific example shown in Fig. 1, which is number 61 in Heap's list for $c = 5$, the labeling of the vertices 3 and 6 can be permuted and the array in braces becomes

$$\begin{bmatrix} 1, 1, 0, 1, 0, 1 \\ 2, 2, 3, 1, 3, 1 \end{bmatrix}$$

(4, 4, 3, 3, 3, 3)

and the degree-tuple is unchanged. Clearly an order relation for arrays in braces can be defined similarly to the way the order relation was defined for matrices. Simply scan the array column by column until there is disagreement. The array with the larger entry in this position of first disagreement precedes the other. Therefore, it is possible to define the *decomposed degree-tuple* to be that array in braces of the star type which precedes all other arrays with the restriction that only arrays are to be con-

sidered which satisfy $s_i + 2d_i + 3t_i + \cdots = D_i$ as defined by the degree-tuple. It is easily verified that the first array presented for star type shown in Fig. 1 gives the decomposed degree-tuple. Other examples of decomposed degree-tuples which precede all the other possible arrays under the permissible permutations are

$$\begin{bmatrix} 0, 1, 1, 1, 1, 0 \\ 5, 2, 2, 2, 2, 3 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1, 1, 0, 0 \\ 1, 0, 2, 1 \\ 0, 1, 0, 1 \end{bmatrix}$$

$$(5, 4, 4, 4, 4, 3) \qquad (5, 4, 4, 3).$$

Now, the obvious ordering for subsets of star types with the same degree-tuple is by the order relation for decomposed degree-tuples. For example, the ordered decomposed degree-tuples for the (4, 3, 3, 3, 3) degree-tuple are

$$\begin{bmatrix} 2, 1, 1, 1, 1 \\ 0, 1, 1, 1, 1 \end{bmatrix}, \quad \begin{bmatrix} 1, 1, 1, 1, 0 \\ 2, 1, 1, 1, 3 \end{bmatrix}, \quad \begin{bmatrix} 1, 1, 0, 0, 0 \\ 2, 1, 3, 3, 3 \end{bmatrix},$$

$$\text{and} \quad \begin{bmatrix} 0, 0, 0, 0, 0 \\ 4, 3, 3, 3, 3 \end{bmatrix}.$$

Now, many star types for $c \leq 5$ can be uniquely determined from the decomposed degree-tuple. The largest subset with the same decomposed degree-tuple in $c = 5$ consists of eight star types. (These are numbered 61 to 68 in Heap's list.) Thus, it seems that further pre-ordering criteria are not really useful for the highest cyclomatic numbers presently obtainable. Rather, it seems that the pre-ordering may be too extensive. One justification for the decomposed degree-tuple is that it is easy to obtain. Another justification will become apparent in the application of the canonical matrix ordering relation.

In the definition of the decomposed degree-tuple, only vertex numberings were allowed which did not alter the degree-tuple. Now, to define the canonical matrix of a star type let us allow only vertex numberings which do not alter the decomposed degree-tuple. For example, the only matrices to be considered in the case of the star type 61 ($c = 5$) shown in Fig 1 are those with the two vertices of degree four labeled 1 and 2, the two vertices of degree three with the incident double edge labeled 3 and 4, and the other two vertices of degree three labeled 5 and 6. Of these eight possible matrices, the one which precedes the others according to the order relation for matrices introduced in Sec. I is called the canonical matrix. As shown in Sec. I the canonical matrices for the star types with the same decomposed degree-tuple can now be completely ordered.

The advantage of using the decomposed degree-tuple restriction on the vertex numbering is now obvious. In the example considered, only eight matrices need be considered to determine the canonical matrix, whereas with no restrictions 720 matrices must be considered. Thus, the pre-ordering criteria involving the degree-tuple and the decomposed degree-tuple facilitate the actual use of the canonical matrix for the final ordering.

Now, it may be shown how one can often write down the canonical matrix of a star type in an easier way than actually constructing all the permissible matrices. Consider again the star type shown in Fig. 1. The a_{12} element in the canonical matrix must be equal to 2. To make $a_{13} = 1$ fixes the numbering of the first four vertices. Of the last two vertices only one is connected to the first vertex and this must be labeled 5 in order to make $a_{15} = 1$. Once the canonical vertex numbering has been obtained, the canonical matrix follows automatically. This rather simple example is fairly typical of most of the star types for $c \leq 5$. The most difficult cases are the star types with all single edges and all vertices of the same degree. For example, consider star type 115 in Heap's list for $c = 5$.[11] Although there are no double edges, there are triangles. Clearly, vertices 1, 2, and 3 must occupy one of the triangles for $a_{12} = a_{13} = a_{23} = 1$. Using the symmetry of the graph, we see that there are only two ways to do this. The discriminating matrix element is a_{45} which is 1 in one case and 0 in the other. The rest of the vertices can then be numbered easily. As another example, star type 117 can be given the canonical vertex numbering immediately when the symmetry is noticed.

Although the list of star types for $c \leq 5$ given by Heap[11] does not show the degree-tuples or the decomposed degree-tuples, it does show the canonical vertex numbering of each star type which is more difficult to supply. To use this list for identifying a star type it is not really essential to have all the canonical matrices written down, since for $c \leq 5$ there are so few star types with the same decomposed degree-tuple and for most of these the first line or two of the canonical matrix suffices for identification.

To identify a graph using a computer is conceptually simple even without the use of the canonical matrix. One simply compares the matrices associated with each of the $v!$ vertex numberings with the matrices in the list until one agrees. However, if one first permutes the vertices to find the canonical

matrix and then compares it to the list of M canonical matrices, the maximum number of comparisons is reduced from $M(v!)$ to $(M + v!)$. Even so, the computer time involved becomes noticeable for $v = 8$ which is also when the manual method becomes difficult.[11] An interesting alternative has been discussed by M. E. Fisher.[14] The idea is to compare the diagonal elements, $(A^n)_{ii}$, of successive powers of the matrices associated with the star types, which amounts to comparing the eigenvalues of the matrices. Unfortunately, there exist pairs of different graphs which have the same eigenvalues, so this is not a conclusive test. However, the smallest known pair of star types with the same eigenvalues has $c = 19$ and $v = 9$.[14] If this could be proved, one would have a valid test for the cyclomatic orders of immediate interest. However, even if the "ultimate arbitrariness" could be disregarded, the ordering scheme based on the canonical matrix seems preferable to one based on the diagonal elements of successive powers because it is easier to use manually. Still, it is pleasant to find that the first disagreement between the two orderings does not occur until one considers star types 42 and 43 in $c = 5$.

Finally, it may be mentioned that the canonical matrix is useful for obtaining star types with a given decomposed degree-tuple. One constructs a matrix row by row using, for each element, the largest number which does not violate the restrictions imposed by the decomposed degree-tuple and such that the matrix is canonical. If such a matrix can be found, one records it, locates the last matrix element where a choice could be made, diminishes that element by one, and attempts to complete the matrix subject to the above restrictions, and so on. A list of stars obtained manually in a trial of the method was correct for $c = 4$ and $c = 5$ except for the last and hardest degree-tuple, (3, 3, 3, 3, 3, 3, 3, 3), as was disclosed by Heap's computer enumeration.[11]

[14] M. E. Fisher, J. Comb. Theory 1, 105 (1966).

SUMMARY

It has been shown that it is possible to completely order undirected graphs systematically while still retaining the natural subdivisions useful for particular problems. In the case of the star types, a particular ordering, based on a sequence of natural subdivisions followed by the canonical matrix ordering, and the accompanying identification procedure have been presented.

ACKNOWLEDGMENTS

I wish to thank M. F. Sykes, M. E. Fisher, and B. R. Heap for valuable discussion and criticism. Also, I am grateful to the National Science Foundation for a NATO Postdoctoral Fellowship and to the host institution, King's College, London, for its hospitality.

Normal Thresholds in Subenergy Variables*

P. V. LANDSHOFF, D. I. OLIVE, AND J. C. POLKINGHORNE
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England
(Received 9 November 1965)

The discontinuity across the normal threshold in a subenergy variable is evaluated by S-matrix methods. The effect on the discontinuity of continuing past higher Landau singularities is discussed.

1. INTRODUCTION

SEVERAL authors[1] have discovered that the structure of the normal thresholds in the subenergy variables of multiparticle amplitudes is complicated. Here we give a derivation of the discontinuities corresponding to these singularities and investigate how their form is affected by the presence of "higher" singularities.

Our method, which is independent of perturbation theory, other approximate models, crossing, and any "unphysical unitarity relations," is similar to that previously used[2,3,4] for the derivation of the discontinuities corresponding to certain other singularities. It involves consideration of the physical unitarity equations E_1 and E_2 operating to either side of the singularity under study. The equation E_1 is continued analytically into the region where E_2 operates, so that now in the latter region there are two different equations. After some manipulation, these yield the desired discontinuity. This is discussed in Sec. 2.

The result involves a simple integration, but it is applicable only for values of the subenergy sufficiently near to the normal threshold under study. When it is continued away from the threshold and across the singularity curves corresponding to higher Landau–Cutkosky diagrams, the integration contours become distorted. This is discussed in Sec. 3. An elementary knowledge of some of the properties of Landau curves provides a simple technique for determining the nature of the distortion for a given continuation. This technique is described in an Appendix.

2. DERIVATION OF THE DISCONTINUITY

Consider a theory in which there are two types of particles. One, of mass m, is denoted by a straight line; the other, of mass M, is denoted by a wavy line. Neither has any intrinsic quantum numbers, and the masses are chosen such that both particles are stable:

$$m < M < 2m.$$

We evaluate the discontinuity associated with the normal threshold $\sigma = (m + M)^2$ in the subenergy variable $\sigma = (p_1 + p_2)^2$ of the amplitude of Fig. 1.

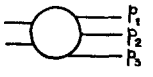
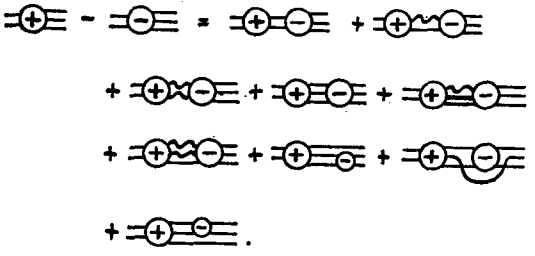
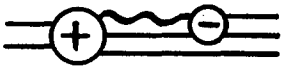


FIG. 1. The amplitude under discussion.

For σ just less than $(m + M)^2$, and suitable values of the other variables, the unitarity equation for this amplitude reads[5]


(1)

For σ just greater than $(m + M)^2$, an extra term


(2)

must be added to the right-hand side. Here, as usual,[5] the labels (+) and (−), respectively, refer to the physical amplitude and its Hermitian conjugate. As far as the variable σ is concerned, the corresponding limits onto the normal-threshold cuts are shown in Fig. 2.

* The research reported in this document has been sponsored in part by the Air Force Office of Scientific Research, OAR, under Grant AF EOAR 63-79 with the European Office of Aerospace Research, United States Air Force.

[1] For example, V. Anisovich, A. Ansel'm, and V. Gribov, Zh. Eksperim. i Teor. Fiz. **42**, 224 (1962) [English transl.: Soviet Phys.—JETP **15**, 159 (1962)]; N. Bonnevay, Nuovo Cimento **30**, 1325 (1963); J. Bronzan and C. Kacser, Phys. Rev. **132**, 2703 (1963); R. Hwa, *ibid.* **134**, B1086 (1964). We are grateful to Dr. I. J. R. Aitchison for a discussion of these papers.

[2] D. I. Olive, Nuovo Cimento **29**, 326 (1963).

[3] P. V. Landshoff and D. I. Olive, J. Math. Phys. **7**, 1464 (1966).

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

[5] An explanation of the diagrammatic notation is found in Ref. 4 or in D. I. Olive, Phys. Rev. **135**, B745 (1964).

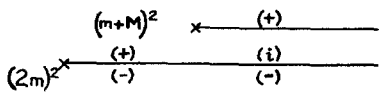
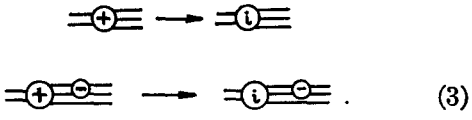


FIG. 2. The complex σ plane. The labels (+) and (−), respectively, denote the physical and Hermitian-conjugate boundary values.

We now continue the whole Eq. (1) analytically in the variable σ, starting at a real value less than $(m+M)^2$, passing *below* the branch point at $(m+M)^2$ to a real value greater than $(m+M)^2$. We see from Fig. 2 that this continuation passes to the correct side of the branch point for preserving the label (−) on those (−) amplitudes in (1) that have σ as an external variable, that is, in all the terms except the first on the left-hand side and the seventh and eighth on the right-hand side. However, for those amplitudes that are labeled with a (+) and have σ as an external variable, namely, in the first term on the left-hand side and the last term on the right-hand side, the continuation is "unnatural" in the sense that it goes into the region labeled (i) in Fig. 2. So we write



(3)

The normal threshold singularity in question appears in (1) not only because of its presence in the bubbles, but also because it can be generated by the integrations in the terms on the right-hand side. We know[4,6] that the singularities of unitarity integrals are generated in an analogous way to those of Feynman integrals. They can be represented by Landau–Cutkosky diagrams and they lie on the corresponding Landau curves. The singularity at $\sigma = (m+M)^2$ can occur in the fourth term of the right-hand side; the singularity here is generated by one of the particles m in the intermediate state together with the single-particle pole M in the right-hand bubble, the other two particles in the intermediate state corresponding to a loop which is contracted out [see Fig. 3(a)]. The singularity can also occur in the fifth and sixth terms [see Figs. 3(b), 3(c)]; in each of these two figures it occurs in two ways. It is only necessary to consider the singularity at $\sigma = (m+M)^2$, since we wish to evaluate the discontinuity across this singularity alone and to make our continuation within a small nearby region through which no other Landau curves pass.

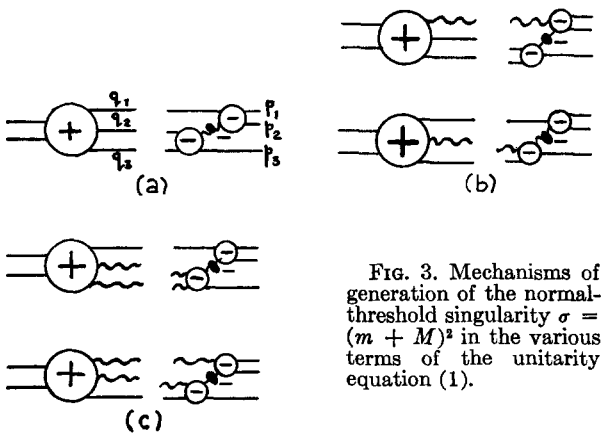
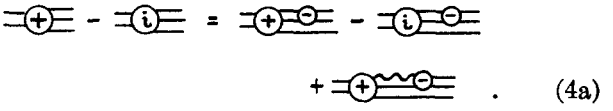


FIG. 3. Mechanisms of generation of the normal-threshold singularity $\sigma = (m+M)^2$ in the various terms of the unitarity equation (1).

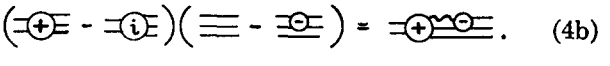
The occurrence of the normal-threshold singularity in this way means that, when we continue (1) across $\sigma = (m+M)^2$, the contours of integration might be distorted from their normal arrangement in the physical unitarity integral. This would mean that terms in the continued unitarity relations and in the physical unitarity integral, which corresponded to the same bubble picture, would in fact relate to integrals with different contours of integration. If this happens, the continuation is called "unnatural" and the difference between the integrals with the different integration contours must be evaluated in calculating the discontinuity. If this does not happen, the continuation is said to be "natural."

We show in the Appendix that the continuation we have chosen is in fact natural in all three of these cases. So, in this continuation of (1) into $\sigma > (m+M)^2$, the only changes that must be made are those shown in (3). If, therefore, we subtract the continued equation from the physical unitarity equation that operates in this region, namely (1) with the extra term (2), we have



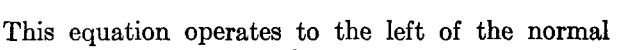
(4a)

Rearrangement gives



(4b)

The first factor on the left is the discontinuity we seek, so the problem now is to find the inverse of the second factor. To this end, consider a unitarity equation for the amplitude $m + m \to m + m$:

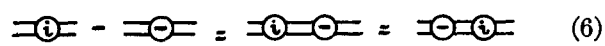


(5)

This equation operates to the left of the normal threshold $\sigma = (m+M)^2$, where σ is now the total-

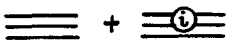
[6] J. C. Polkinghorne, Nuovo Cimento 23, 360 (1962); *ibid.* 25, 901 (1962).

energy variable for this amplitude. If we continue analytically to the right of this threshold, passing below it in the complex plane as before, we obtain

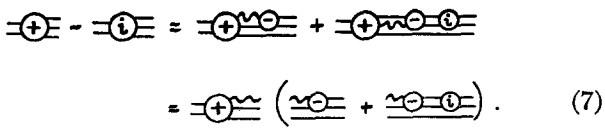


(6)

for $\sigma > (m + M)^2$. Hence "postmultiplication" of (4b) by



gives

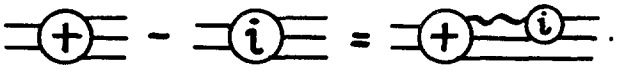


(7)

But the analog of (6) for the amplitude $m + M \to m + m$ is

(8)

Hence, (7) reduces to



(9)

3. DISCUSSION

The result (9) is the discontinuity we have been seeking and we now examine just when it is valid. First, notice that (1) only applies for a certain range of values of the variables other than σ. For other ranges, more or fewer terms will be present, but this will actually not affect the final result (9), because the only terms that mattered in the derivation were those in (3). This is so if σ is sufficiently close to the threshold $(m + M)^2$. We now consider what happens if it is not.

Since the term (2) occurs in a physical unitarity relation, the integration implied is over all physical values of the internal momenta. The integration is two dimensional and we can choose, as one of the integration variables, the subenergy $\sigma' = (q_1 + p_3)^2$ in terms of the numbering of Fig. 4. If the integration over the second variable, whatever it is, is supposed

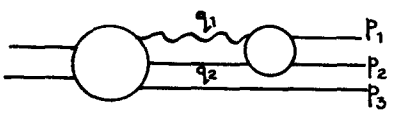


FIG. 4. The labeling of momenta as used in Sec. 3 for the discussion of the integration in the term on the right-hand side of (9).

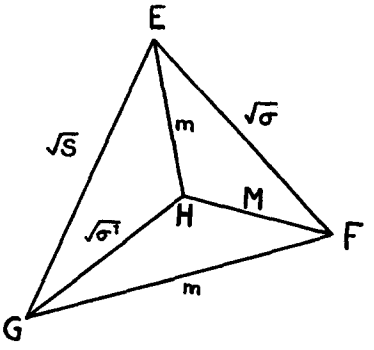


FIG. 5. Vector diagram associated with Fig. 4.

performed, the range of integration of σ' is

$$\sigma'_- \leq \sigma' \leq \sigma'_+,$$

where the end points σ'_- and σ'_+ are found by considering the vector diagram, Fig. 5, in which H is not necessarily coplanar with EFG. In this diagram, lines EF, EG, GF, EH, and HF correspond to the vectors $(p_1 + p_2)$, $(p_1 + p_2 + p_3)$, p_3, q_1, and q_2, respectively with corresponding specified lengths $\sigma^{\frac{1}{2}}$, $s^{\frac{1}{2}}$, m, m, and M [where $s = (p_1 + p_2 + p_3)^2$ is the total-energy variable]. The way they form triangles in Fig. 5 corresponds to the linear relations between the vectors resulting from their definitions and from the conservation law $p_1 + p_2 - q_1 - q_2 = 0$. For example, the latter finds expression in the fact that EF, FG, and GE form a closed loop. Now, the triangle EFG is fixed while the point H is forced to move in a circle with EF as axis. Since the length of GH gives the possible values of $\sigma'^{\frac{1}{2}}$ in the integration region, the extreme values obviously occur when H lies in the plane of the triangle EFG.

If, instead, we take $\sigma'' = (p_3 + q_2)^2$ as residual integration variable, the integration range can be found similarly.

Since it is the term (2) that results in the term on the right-hand side of (9), the integration in the latter will be exactly similar. We have been considering σ not too much greater than $(m + M)^2$, but if it is increased, complications arise from σ'_- or σ'_+ passing through a value that coincides with a normal threshold of the left-hand bubble in the variable σ'.

If, in particular, $\sigma'_\pm$ coincides with the threshold $\sigma' = 4M^2$, the external variables σ and s lie on the Landau curve corresponding to the first triangle diagram in Fig. 3(c), since it is known that the curve corresponds to this geometrical configuration, which in this case is called a "dual diagram." The part L_1 of this Landau curve lying in the physical region is the arc of a hyperbola drawn in solid line

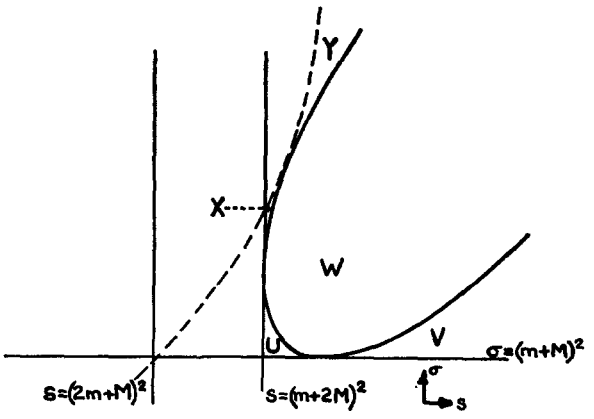


FIG. 6. The real (s, σ) plane. The straight lines are normal thresholds, the solid curve is an arc of the Landau curve L_1 corresponding to the first triangle diagram in Fig. 3(c), and the dotted curve is the "second-type" curve.

in Fig. 6. We expect the integrals considered to be singular all along L_1, since it is generated by an end-point mechanism.

We conclude that, in the regions U and V close to the $\sigma = (m + M)^2$ normal threshold, the σ' integration in (9) takes the straightforward arrangements shown in Figs. 7(a) and 7(b), whereas if we pass from regions U or V across the triangle curve L_1 into W, the contour must depend upon the route we take around L_1 in complex (s, σ) space.

Consider, in particular, possible routes from U to W. Since the piece of singular curve L_1 we have to cross has negative gradient, the plane normal to it in complex (s, σ) space lies in

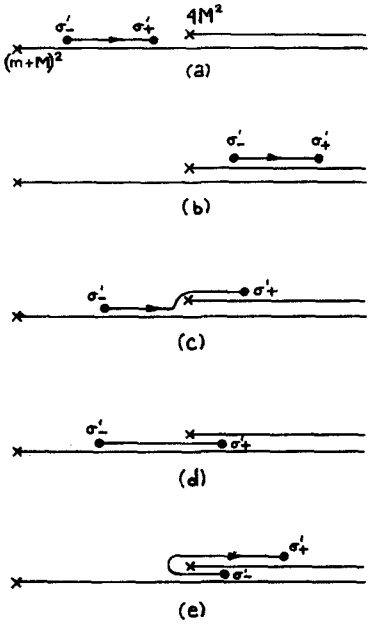


FIG. 7. The σ' integration contour for continuations across L_1 into various regions of Fig. 6.

$$\text{Im } s/\text{Im } \sigma > 0.$$

So we may avoid the singularity either by a distortion Im s, Im $\sigma > 0$ or by a distortion Im s, Im $\sigma < 0$. By the methods of the Appendix, we see that in the former case σ'_+ passes to the upper side of the $4M^2$ branch point, thus giving the contour of Fig. 7(c), while in the latter case it passes to the lower side giving the contour of Fig. 7(d). Part of the latter contour lies in the unphysical limit between the cuts attached to the branch points $\sigma' = 4M^2$, $\sigma' = (m + M)^2$.

Suppose we take the first route from U to W and now across into region V. This time we may cross the arc of L_1 using either of the distortions

$$\text{Im } s > 0, \qquad \text{Im } \sigma < 0,$$

$$\text{Im } s < 0, \qquad \text{Im } \sigma > 0.$$

The techniques of the Appendix tell us that in the first case σ'_- passes above the $4M^2$ branch point and in the second it passes below, thus leading, respectively, to Figs. 7(b) and 7(e).

Now, the derivation of Sec. 2 is valid for region V and leads to Fig. 7(b). We can understand why only one of the continuations from U to V via W leads to this result by considering a path JJ' (lying in the σ plane of the amplitude and running from a point J in the physical limit) round the $\sigma = (m + M)^2$ normal threshold and returning directly back to J'. Initially, J and J' lie in the region U (Fig. 6) and the difference between the amplitude evaluated at the two points is the normal-threshold discontinuity of interest.

As we increase s, J and J' cross L_1 twice and are given positive or negative imaginary parts in σ in accordance with the previous discussion. As J and J' move into W and finally into V, we have the sequence of pictures illustrated in Fig. 8. In drawing these pictures, we have assumed that in the amplitude it is only the part of L_1, separating region U from region W, which is singular in the physical limit (this was proved in Ref. 3). Corresponding to the continuations of the discontinuity integral leading to Figs. 7(b) and 7(e), we have sequences 8(a) and

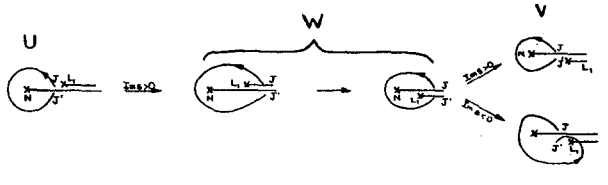


FIG. 8. The complex σ plane for various values of s, showing how different continuations from the region U of Fig. 6 across L_1 into W yield different discontinuities.

8(b) in Fig. 8. Sequence 8(a) does indeed lead to the discontinuity of interest in region V, whereas the discontinuity obtained in sequence 8(b) involves the L_1 cut.

It is not only L_1 that can distort the integration over the subenergy σ'. The Landau singularity L_2 corresponding to the second diagram of Fig. 3(b) plays a part. This Landau curve is a degenerate hyperbola (a pair of coincident straight lines) and can be considered as a limiting case of a curve like L_1 as the external mass varies (in S-matrix theory, this is achieved by adding more external lines[4,5,6]). In the limit the points A, B, and P of Fig. 12 coincide at the intersection of the normal thresholds, and the regions U and X disappear. The authors in Ref. 1 have shown that the total discontinuity across L_2 is zero (so that it is nonsingular) and that crossing it wraps the σ' contour around the $(m + M)^2$ normal threshold as in Figs. 9(a) or 9(b).

We show this by demonstrating the corresponding properties of L_1 which reduce to this in the degenerate limit.

Let us consider the four possible ways of continuing the integral from region V through region W to region Y (see Fig. 6). According to the methods of the Appendix, the final contours are as shown in Fig. 9.

The final contours in cases (c) and (d) are the same so that two different paths on the Riemann surface of the integral lead to the same point, whatever the nature of the σ' branch point. Actually this branch point, being a two-particle threshold, is two sheeted, and so in this case the contours resulting in cases (a) and (b), while looking different, are actually equivalent because the difference between them is a contour encircling the two-sheeted branch point twice. Then the sum of the σ discontinuities across the two branches of L_1 is zero.

It is also found that each branch of L_1 is itself two sheeted (as will be shown in a subsequent paper) and that similar arguments apply to continuations from U to W to X.

We have been discussing the subenergy σ' in some detail, but similar effects are encountered in the integration over the subenergy $\sigma'' = (p_3 + q_2)^2$ and involve the other Landau diagrams of Figs. 3(b) and 3(c). [The Landau diagram of Fig. 3(a) has not been mentioned, because the corresponding singularity curve does not enter the physical region.] Also, the triangle Landau singularities, corresponding to the subenergy integration and being distorted by a normal-threshold singularity, are not the only ones to be involved. A complete discussion would also have to consider higher singularities, corresponding to the integration being entangled with other singularities of the bubbles in the integral.

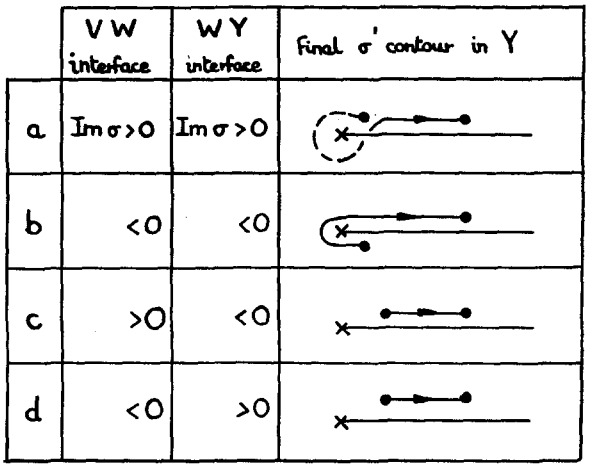


FIG. 9. Integration contours for various continuations from region V of Fig. 6 into Y.

In conclusion, we mention that our analysis has concerned the $(m + M)$ normal threshold, which is a normal threshold *inside* the physical region for the amplitude we were considering. This is so that we could use physical unitarity relations to either side of the threshold. But we expect that very similar results hold for the $(m + m)$ normal threshold, which is on the boundary of the physical region. An analogous proof would require the unphysical unitarity relation[7] operating to the unphysical side of the threshold.

APPENDIX

We now explain a technique for deciding how an integration end point passes a singularity of the integral. We saw that the end points of the σ' integration were determined in terms of the external variables s and σ by Fig. 5, which could be thought of as a dual diagram corresponding to a Feynman diagram with variable internal mass σ'. Since the dual diagram is equivalent to the Landau equations, our problem reduces to understanding the behavior of the Landau curve as an internal mass varies.

The argument to be presented is capable of general application but we just consider[4,8] the Feynman graph of Fig. 10, with internal masses m_1, m_2, and m_3, and external masses m_{23}, m_{31}, and

[7] Unphysical unitarity relations, and their proof, are discussed in Ref. 4 and in the paper by Olive mentioned under Ref. 5.

[8] M. Fowler, P. V. Landshoff, and R. W. Lardner, Nuovo Cimento 17, 956 (1960). There is a printer's error in this paper; Fig. 1(b), which is relevant here, should be rotated clockwise through 90°.

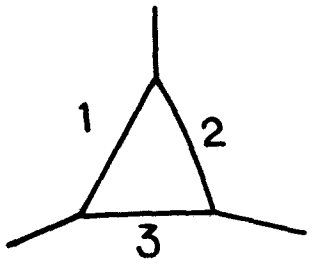


FIG. 10. A Feynman graph discussed in the Appendix.

m_{12}. The equation of the leading Landau singularity of this graph (with all internal lines on the mass shell) can be obtained from the function

$$D = \sum \alpha_i \alpha_j m_{ij}^2 - C \sum \alpha_i m_i^2, \qquad \text{(A1)}$$

where

$$C = \alpha_1 + \alpha_2 + \alpha_3,$$

by eliminating the parameters α by means of the equations

$$\partial D/\partial \alpha_i = 0, \qquad i = 1, 2, 3. \qquad \text{(A2)}$$

Since D is homogeneous in the α, it is evident that (A2) implies $D = 0$.

Suppose now that a set of the m's is chosen, corresponding to a point on the leading singularity, and the corresponding values of the α's are found. If now a small variation is made in the m's, D will change according to

$$dD = \sum \frac{\partial D}{\partial (m_{ij}^2)} d(m_{ij}^2) + \sum \frac{\partial D}{\partial (m_i^2)} d(m_i^2) + \sum \frac{\partial D}{\partial \alpha_i} d\alpha_i.$$

So the variation will correspond to remaining on the singularity if

$$0 = \sum \alpha_i \alpha_j \, d(m_{ij}^2) - C \sum \alpha_i \, d(m_i^2). \qquad \text{(A3)}$$

The Feynman graph with Fig. 5 as a dual diagram is shown in Fig. 11. The dotted line carries mass $\sqrt{\sigma'}$. Comparing with Fig. 10 we identify

$$\begin{aligned} &m_1^2 = M^2, \quad && m_2^2 = m^2, \quad && m_3^2 = \sigma', \\ &m_{12}^2 = \sigma, \quad && m_{31}^2 = s, \quad && m_{23}^2 = M^2. \end{aligned} \qquad \text{(A4)}$$

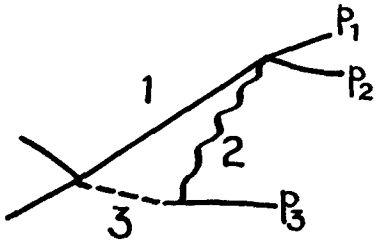


FIG. 11. The Feynman graph having Fig. 5 as dual diagram.

We have said in Sec. 3 that, when one of the end points coincides with $\sigma' = 4M^2$, we have the Landau singularity corresponding to the first diagram in Fig. 3(c), drawn as the curve L_1 in Fig. 6. So, if we take a small detour $(ds, d\sigma)$ around L_1, we see from (A3), with the identification (A4), that the displacement $d\sigma'$ of the end point from $4M^2$ is given by

$$0 = \alpha_1 \alpha_2 \, d\sigma + \alpha_1 \alpha_3 \, ds - C \alpha_3 \, d\sigma'.$$

Hence, to decide whether $d\sigma'$ goes above or below $4M^2$ for a given sign of Im $d\sigma$, Im ds we merely need to know the signs of the α's and of C. The real motion is likewise determined.

The signs of the α's are easily discovered[8,4] and are shown in Fig. 12. This figure shows the leading curve for the general Feynman graph of Fig. 10, drawn in the real (m_{13}^2, m_{12}^2) plane with m_{23} fixed at a value satisfying

$$m_3 > m_2 + m_{23} \qquad \text{(A5)}$$

[since, using (A4), we are interested in $2M > m + M$.] The straight lines in this diagram are the normal thresholds $m_{13}^2 = (m_1 + m_3)^2$, $m_{12}^2 = (m_1 + m_2)^2$. It is known that, at the contacts A and B with these normal thresholds, the signs of the α's are, respectively, $(+0+)$ and $(++0)$; continuity then gives the signs on the other parts of the curve. The point P denotes the contact of the curve with the "second type" curve[9]

$$m_{13} = m_{12} + m_{23}.$$

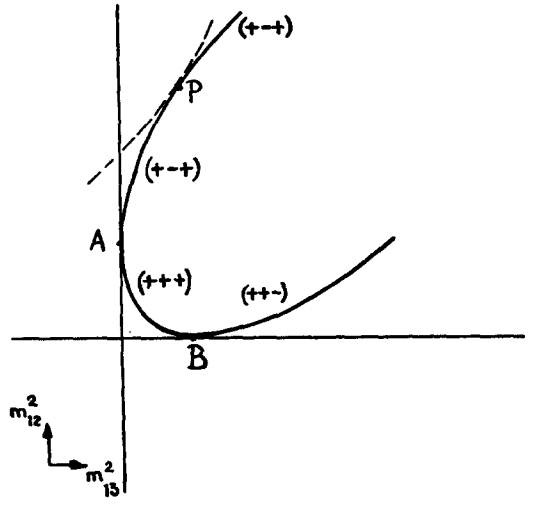


FIG. 12. The signs of the α's for various parts of an arc of Landau curve corresponding to the Feynman graph of Fig. 10. m_{23} is fixed at a value satisfying (A5). The dotted curve is the "second-type" curve.

[9] D. B. Fairlie, P. V. Landshoff, J. Nuttall, and J. C. Polkinghorne, J. Math. Phys. 3, 594 (1962).

[By (A4) this becomes

$$s^{\frac{1}{2}} = \sigma^{\frac{1}{2}} + M.$$

This parabola also bounds the physical region in s and σ and appears in Fig. 6 as a dotted line.]

P is the point at which C vanishes; on the arc PAB∞ C is positive; on ∞P, negative.

In Sec. 3 we also discussed the curve L_2, corresponding to the second diagram of Fig. 3(b). This is an example of a degenerate Landau curve; the inequality (A5) is replaced by an equality. The points P, A, B, of Fig. 12 all coincide at O and the arcs ∞P, B∞ coincide. Hence, to discover what happens when we take a detour $(ds, d\sigma)$ around L_2, one just considers similar successive detours around the arcs ∞P, B∞ in the nondegenerate case, and considers what happens in the limit as m_2 is reduced so as to give degeneracy.

Another similar problem we may solve by these methods is that encountered in Sec. 2. This concerns the generation of the $\sigma = (m + M)^2$ normal threshold in the fourth, fifth, and sixth terms of the right-hand side of (1) by a pole in the right-hand bubble. Consider, for definiteness, the mechanism indicated in Fig. 3(a) for the fourth term. Take as integration variables $u = (p_1 + p_2 - q_1)^2$ and four other variables. Suppose the latter are integrated out. Then the normal threshold is generated by an end point of the u integration coinciding with the pole $u = M^2$ of the right-hand bubble. Now, this end point corresponds to the four-momentum, q_1 being parallel to $(p_1 + p_2)$. But this is precisely the same condition as that for the singularity of the Feynman graph of Fig. 13, in which the dotted line is given mass $u^{\frac{1}{2}}$. The D function for this graph is

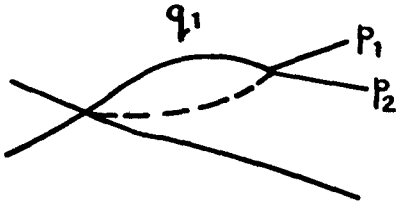


Fig. 13. The Feynman graph used in the discussion at the end of the Appendix.

$$xy\sigma - (x + y)[xm^2 + yu],$$

where x, y are, respectively, the Feynman parameters for the lines q_1 and the other internal line. In the same way as before, a displacement of σ from the singularity being generated corresponds to a displacement du of the end point from the generating pole, given by

$$0 = xy\, d\sigma - (x + y)y\, du.$$

Since x, y for a normal threshold are positive, we see that Im $d\sigma$ and Im du take the same sign, which is the information needed in Sec. 2.

We stress that, although the techniques of this Appendix "borrow" some results of perturbation theory, this is entirely a matter of algebraic convenience and the results of the text are quite independent of perturbation theory.

The Icecream-Cone Singularity in S-Matrix Theory*

P. V. LANDSHOFF, D. I. OLIVE, AND, J. C. POLKINGHORNE
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England
(Received 9 November 1965)

The icecream-cone singularity is analysed by S-matrix theory methods. It is found to appear only in that part of the physical region predicted by finite-order perturbation theory. The corresponding discontinuity is derived, and the nature of the singularity is discussed.

1. INTRODUCTION

IN a previous paper[1] it is shown that, at least as far as the physical region is concerned, the structure of the simple triangle singularities is very similar to that in finite-order perturbation theory. It is hoped that the analysis given there is readily extendable to all the "simple" physical-region singularities of the S matrix, that is those whose Landau–Cutkosky diagram consists only of a network of single lines.[2] But the analysis of the non-simple singularities appears to be more involved. The simplest example of such a singularity is the "icecream cone", for which one of the internal lines of the Landau–Cutkosky diagram is doubled (Fig. 1).

The doubling of the internal line has two effects. First, the extraction of the discontinuity from the physical unitarity equations is rather more complicated than for the simple triangle. Second, the result for the discontinuity cannot be compared with finite-order perturbation theory, because an infinity of nontrivial Feynman graphs, corresponding to iterations of the two-particle loop, possess the singularity as their leading singularity. This is connected with the result we find below, that, unlike the case of the simple triangle, the expression for the discontinuity contains subsidiary amplitudes that are not all evaluated on the physical sheet. (A similar result can be obtained formally by summing an infinite number of perturbation theory discontinuities.) However, we do find that the Riemann sheet properties of the singularity in the physical region are as in finite-order perturbation theory: only the positive-α parts are singular.

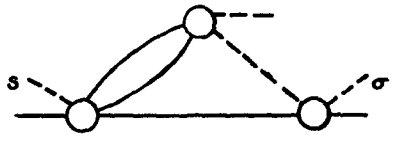


FIG. 1. The icecream-cone diagram.

In writing this paper, we have assumed that the reader has at least some familiarity with our paper on the simple triangle.[1] In particular, we do not here give full details of how we calculate to which side of integration hypercontours singularities of integrands pass when analytic continuation is performed. These calculations may either be done by means of Cayley determinants[1] or, rather more simply, by using some elementary properties of the Feynman denominator functions for the simple triangle graph.[3] Perhaps we should add that we do not expect the reader to have any knowledge of our paper[4] on the icecream cone in perturbation theory; the work here is quite independent of perturbation theory.

In the equal-mass case, the icecream-cone singularity curve is degenerate, becoming a pair of coincident straight lines. So we consider a theory with two types of particle having the quantum numbers of pions and nucleons, but with the mass of the nucleon not much greater than that of the pion:

$$\tfrac{3}{2}\mu > M > \mu. \tag{1}$$

This choice of masses serves to reduce the number of terms in the unitarity relations. As is indicated in Fig. 1, the icecream-cone singularity considered here is contained in the amplitude for the process

$$\pi + N \to \pi + \pi + N \tag{2}$$

and involves an internal nucleon–antinucleon closed loop. The labels s and σ in the figure denote the

* The research reported in this document has been sponsored in part by the Air Force Office of Scientific Research, OAR, under Grant AF EOAR 63-79 with the European Officie of Aerospace Research, United Stated Air Force.

[1] P. V. Landshoff and D. I. Olive, J. Math. Phys. 7, 1464 (1966).

[2] Further cases have been discussed by M. J. W. Bloxham, Nuovo Cimento (to be published), and a general discussion will be given by M. J. W. Bloxham, D. I. Olive, and J. C. Polkinghorne (in preparation).

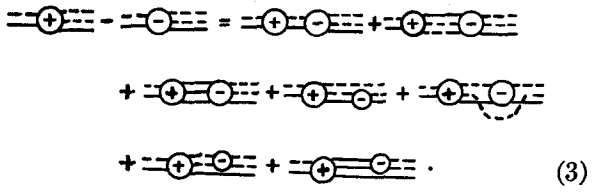
[3] P. V. Landshoff, D. I. Olive, and J. C. Polkinghorne, J. Math. Phys. 7, 1593 (1966).

[4] P. V. Landshoff, D. I. Olive, and J. C. Polkinghorne, Nuovo Cimento 43, 444 (1966).

total-energy variable and the relevant final-state subenergy.

2. METHOD OF ANALYSIS

In the range $(3M)^2 \leq s < (3M + \mu)^2$ of the total energy, the unitarity equation for the process (2) reads



(3)

(The third and last terms on the right-hand side each contain an antinucleon in the intermediate state.) The form (3) of the unitarity equation enables us to derive the discontinuity of the physical amplitude

(4a)

corresponding to the icecream-cone singularity L. There is another form of the unitarity equation, in which the labels (+) and (−) on the right-hand side of (3) are interchanged. With this one could obtain, by exactly similar methods, the corresponding discontinuity of the Hermitian-conjugate amplitude

(4b)

The unitarity equation (3) is valid on either side of L, but is not an analytic continuation of itself across L, because of the presence of the singularity L in the various terms. The singularity L may be found in a term on the right-hand side for either of two reasons (or both): it is *generated* by the presence of lower singularities within the bubbles, or it is *regenerated* by the presence of L itself within a bubble. The generation mechanisms are tabulated

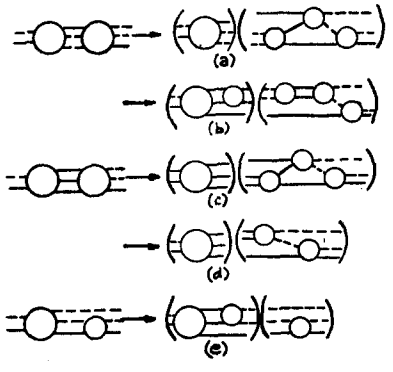


FIG. 2. Generation mechanisms.

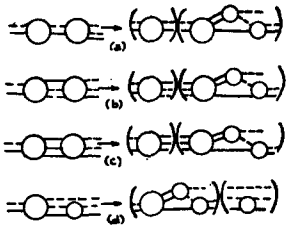


FIG. 3. Regeneration mechanisms.

in Fig. 2 and the regeneration mechanisms in Fig. 3. They are explained more fully below.

The singularity curve L is the same as the leading singularity for the Feynman graph of Fig. 4. The part lying in the physical region for the process (2) is drawn in the real (s, σ) plane in Fig. 5. In this figure we have labeled, for the various arcs of L, the signs of the Feynman parameters α_1, α_2, α_3 corresponding to the internal lines of Fig. 4. The point

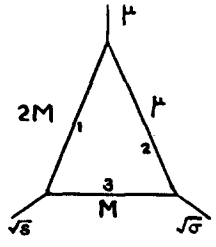


FIG. 4. The equivalent triangle diagram.

X is the contact of L with the second-type singularity curve (which is singular on unphysical sheets[5]) at which the quantity $C = \alpha_1 + \alpha_2 + \alpha_3$ changes sign; on the arc ∞X, C is negative, on XAB∞ it is positive. This information is relevant for the simple method[3] of determining how singularities of integrands lie with respect to boundaries of integration regions.

Our procedure is to continue (3) analytically from outside L to inside L, and compare the result with the (different) Eq. (3) operating inside L. For any of the mechanisms in Figs. 2 and 3 for producing

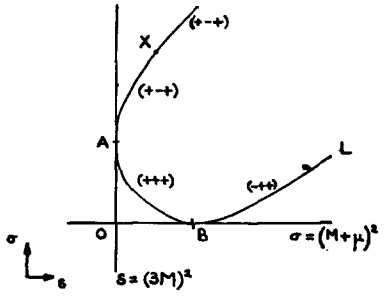


FIG. 5. The Landau curve in the physical region.

[5] D. B. Fairlie, P. V. Landshoff, J. Nuttall, and J. C. Polkinghorne, J. Math. Phys. 3, 594 (1962).

the singularity L in the terms of (3), there is a *natural* continuation[1] across L in the sense that it is this route around L in complex (s, σ) space which preserves the form of the term. Other continuations lead to distortions of the integration contour so that the continued term differs from the corresponding term in the unitarity equation on the other side of L by a discontinuity which must be calculated. (In this paper, we adhere to the convention that internal lines in bubble diagrams correspond to momenta integrated over physical values.) The natural continuations are not the same for each of the mechanisms, which is why Eq. (3) as a whole is not an analytic continuation of itself across L.

We do not want to have to bring in the icecream-cone singularity of the amplitudes for the processes $\pi + \pi + N \to \pi + \pi + N$ and $N + \bar{N} + N \to \pi + \pi + N$ if we can avoid it. We therefore choose to make that continuation which is natural as far as the regeneration mechanisms of Figs. 3(b) and 3(c) [and so also Fig. 3(a)] are concerned. If η is the variable that measures distance along the inward normal to L at any point of L, its explicit definition being that displacements $(ds, d\sigma)$ from L along the normal are given by

$$ds = \alpha_1 \, d\eta, \qquad d\sigma = \alpha_2 \, d\eta, \tag{5}$$

the required continuation is[1,6] the one that avoids L by a route in Im $\eta < 0$.

We now have to determine what is the effect of this continuation on the various generation mechanisms of Fig. 2.

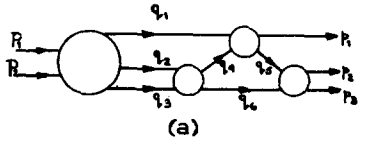
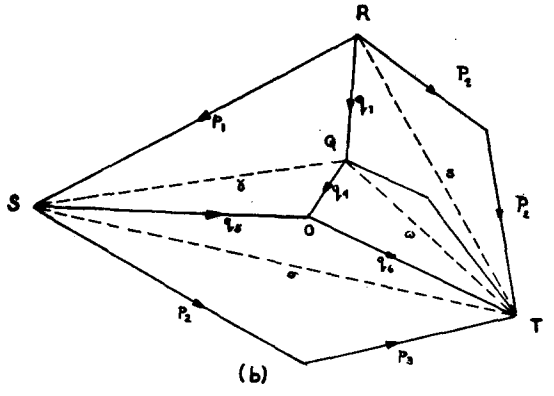


FIG. 6. Momenta for the mechanism of Figs. 2(a) and 2(c).



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

3. GENERATION MECHANISMS

We show that the continuation via Im $\eta < 0$ is natural for all the mechanisms in Fig. 2, with the following exceptions: Fig. 2(b) on the arc XA, Fig. (2d) on the arc XA, Fig. 2(e) on the arcs XA and AB. We also discover how these exceptions differ from the natural continuation, which for these would be via Im $\eta > 0$. That is, we evaluate the discontinuities across L produced on these arcs by these mechanisms.

Figures 2(a) and 2(c)

Here the icecream-cone singularity L is generated by the presence of a simple triangle singularity L' in the right-hand bubble. L' can only generate L when L' itself is singular, and we already know[1] that this demands that the α-parameters for L' all be positive. This means, as we now show, that the mechanism produces a singularity only on the arc AB of L.

The relevant four-vector diagram is drawn in Fig. 6(b), with the momenta labeled as in Fig. 6(a). The critical configuration yielding L' is the coplanarity of q_4, q_5, q_6. The corresponding α's, which will be distinguished by primes from those previously introduced for L, are the coefficients multiplying the three vectors $-q_4$, $-q_5$, $-q_6$ radiating out from O in the linear relation among the vectors that expresses the fact of their coplanarity. The condition that the α' be positive is that O lie inside the triangle QST. In the critical configuration yielding L, there is the additional restriction that q_1, q_4 be parallel, that is RQO is a straight line. The corresponding α are the coefficients multiplying the vectors **OR, OS, OT** in the vanishing linear combination of these vectors, and they are positive if O lies inside the triangle RST. But it is evident that this is the case if and only if O lies inside QST, so the two sets α, α' are positive or not positive together.

Hence on the arcs ∞X, XA, and B∞ of L, either continuation Im $\eta \gtrless 0$ is natural since the singularity is, in fact, absent. It remains to show that Im $\eta < 0$ is natural on AB. Suppose that the integration variables are taken as

$$\gamma = (p_1 - q_1)^2, \qquad \omega = (q_2 + q_3)^2,$$

and two others that have been integrated out. The icecream-cone singularity arises when the boundary $B(\gamma, \omega; s, \sigma) = 0$ of the (γ, ω) integration touches the simple triangle singularity $L'(\gamma, \omega; \sigma) = 0$. The boundary B corresponds to the four-vectors p_1, q_1

and $(p_1 + p_2)$ being coplanar, that is, the points QRST in Fig. 6(b) lying in a plane. This condition we may express[3] by introducing a function D'' and parameters α_1'', α_2'', α_3'' which are the coefficients of the vectors **QR**, **QS**, **QT** in their vanishing linear combination:

$$D'' = \alpha_2''\alpha_3''\sigma + \alpha_3''\alpha_1''s + \alpha_1''\alpha_2''p_1^2 - (\alpha_1'' + \alpha_2'' + \alpha_3'')(\alpha_1''q_1^2 + \alpha_2''\gamma + \alpha_3''\omega),$$

and solving

$$D'' = \partial D''/\partial\alpha_1'' = \partial D''/\partial\alpha_2'' = \partial D''/\partial\alpha_3'' = 0. \quad (6)$$

The equivalence between these equations and the geometrical condition for the boundary is simply the equivalence between the Landau equations and the dual diagram construction for their solution.[6] The equation of L', which corresponds to OQST being coplanar, may be obtained similarly; we have already introduced the parameters α_1', α_2', α_3' for the vectors **OQ**, **OS**, **OT**, and we know that the latter have to be positive in the critical situation in which L is generated (OQR collinear). By inspection[7] of Fig. 6(b), in this case the α'' are also positive. If we make a variation of γ, ω on B, keeping the other variables fixed and using (6), we see that

$$0 = dD'' = -(\alpha_1'' + \alpha_2'' + \alpha_3'')(\alpha_2''\,d\gamma + \alpha_3''\,d\omega),$$

and so

$$\partial\gamma/\partial\omega = -\alpha_3''/\alpha_2''.$$

Similarly, we may calculate that on L'

$$\partial\gamma/\partial\omega = -\alpha_3'/\alpha_2',$$

and so the touching of B and L' requires that

$$\alpha_3''/\alpha_2'' = \alpha_3'/\alpha_2'$$

at the contact. As we have said, the point (s, σ) is then on L. If we make a small displacement $(ds, d\sigma)$ away from L we find using (6) that B moves according to

$$0 = \alpha_2''\alpha_3''\,d\sigma + \alpha_3''\alpha_1''\,ds - (\alpha_1'' + \alpha_2'' + \alpha_3'')(\alpha_2''\,d\gamma + \alpha_3''\,d\omega).$$

If the displacement takes the form (5), with α_1 and α_2 both positive (since we are considering the arc AB of L), and if Im $d\eta < 0$, we see that B moves such that

$$\text{Im}\,(\alpha_2''\,d\gamma + \alpha_3''\,d\omega) < 0.$$

Similarly, we may calculate that L' moves such that

$$\text{Im}\,(\alpha_2'\,d\gamma + \alpha_3'\,d\omega) > 0.$$

Thus, in the plane in complex (γ, ω) space normal to the critical contact between B and L', we have found that the edge B of the integration region passes below the singularity L' of the right-hand bubble. Since the latter carries the label $(-)$, this is therefore a natural continuation.

Figure 2(b)

Label the momenta as in Fig. 7 and take the integration variables as $l = (q_1 + q_2)^2$, $u = (q_1 + q_2 - p_1)^2$, and others which are supposed already to be integrated out. The icecream-cone singularity L arises from the boundary B' of the (l, u) integration passing through the intersection of the pole $u = \mu^2$ of the right-hand bubble and the normal threshold $l = 4M^2$ of either bubble. In the complex l plane the integration contour passes *between* these two normal thresholds, since the bubbles carry labels of opposite sign. This means that it is "anchored" to the point $l = 4M^2$ and cannot be distorted upwards or downwards in the l plane. To determine whether or not the continuation Im $d\eta < 0$ is natural, we must see whether, for $l = 4M^2$, the edge B' of the u integration passes above or below the pole $u = \mu^2$ in the complex u plane. Our standard methods[1,3] show that, except on the arc XA, it passes *below*, so that the continuation is natural.

On the arc XA we must determine how our continuation differs from the natural continuation. Since the continuation from $d\eta < 0$ to >0 via Im $d\eta > 0$ is the natural continuation, the required difference is just the negative of the discontinuity corresponding to an anticlockwise circuit in the complex η plane from the natural configuration at $\eta = 0_+$. To find this, consider the integral as successive u- and then l-integrations. The pole $u = \mu^2$ produces an end-point singularity S in the u-integration. The singularity L arises in the l-integration from S pinching one of the normal thresholds $l = 4M^2$. Standard methods[1,3] again show that, on the arc XA, S is above the contour in the l plane for Im $\eta > 0$, so that it is the branch point $l = 4M^2$ below the contour that participates in the pinch, namely that of the left-hand bubble with

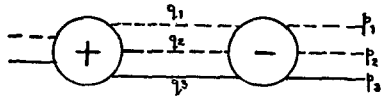


FIG. 7. Momenta for the mechanism of Fig. 2(b).

[7] Actually, Fig. 6(b) is not Euclidean, but topological notions of "inside" and "outside" are not affected by this complication.

the label (+) [Fig. 8(a)]. As we make our anti-

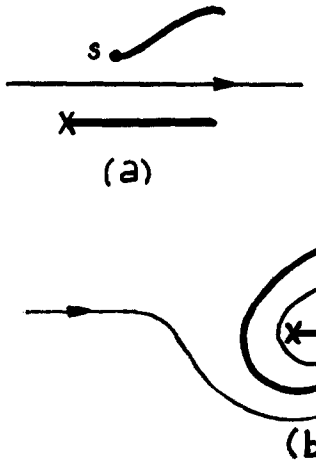
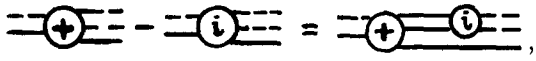


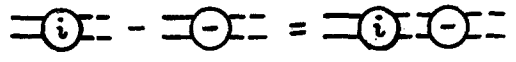
FIG. 8. The distortion of the l-contour on encircling the singularity.

clockwise circuit in the η plane, S encircles the branch point $l = 4M^2$ in an anticlockwise sense, dragging the integration contour underneath the $4M^2$ cut [Fig. 8(b)]. The resulting discontinuity is an integral of the $4M^2$-cut discontinuity of the S-cut discontinuity.[8] The latter discontinuity is the residue of the integrand at the $u = \mu^2$ pole, and the former discontinuity involves the discontinuity of the amplitude $\pi + N \to \pi + \pi + N$ across the $4M^2$ cut. This we write as[9]



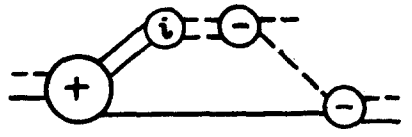
(7)

where

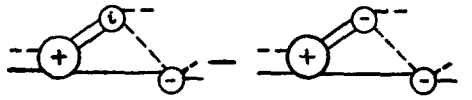


(8)

defines the (i) bubble. The (i) bubble is a convenient construct in terms of physical region amplitudes and (8) need not be thought of as involving an appeal to an unphysical unitarity equation [though in fact (8) is just the continuation of such an equation[10]]. Taking careful account of signs, we find that our continuation demands the addition of the term



to the term under study. By (8), we may write this as



(9)

Figure 2(d)

If we label the three internal momenta in a way similar to the previous case, and define $l = (q_1 + q_2)^2$, $u = (q_1 + q_2 - p_1)^2$ as before, then part of the boundary of the (l, u) integration is the curve B' previously introduced. However, $l = 4M^2$ now plays a different role. It is again a singularity of both the right- and left-hand bubbles, but it is also a singularity of the Jacobian of the transformation of the integration variables from q_1, q_2, q_3 to l, u and other variables that were supposed to be integrated out. Furthermore, it is part of the boundary of integration, together with B',

The icecream-cone singularity L arises from the "corner" u_0 formed by B and $l = 4M^2$ of the integration boundary passing through the pole $u = \mu^2$ of the right-hand bubble (Fig. 9). If we make a displacement (5) away from the critical configuration, we find[3] that u_0 suffers a displacement du along $l = 4M^2$, where

$$du = [\alpha_3(\alpha_1^2 + \alpha_2^2)/\alpha_2(\alpha_1 + \alpha_2 + \alpha_3)]\, d\eta.$$

So, for Im $d\eta < 0$ the corner passes below the pole $u = \mu^2$ except on the arc XA, and the continuation is natural except on this arc.

The integral has the form

$$\int_{u_0} du \int_{4M^2}^{l_+(u)} dl \, \frac{\phi(u, l)}{u - \mu^2 - i\epsilon},$$

where $[l_+(u), u]$ lies on B'. This is

$$\int_{u_0} du \, \frac{\psi(u, l_+)}{u - \mu^2 - i\epsilon},$$

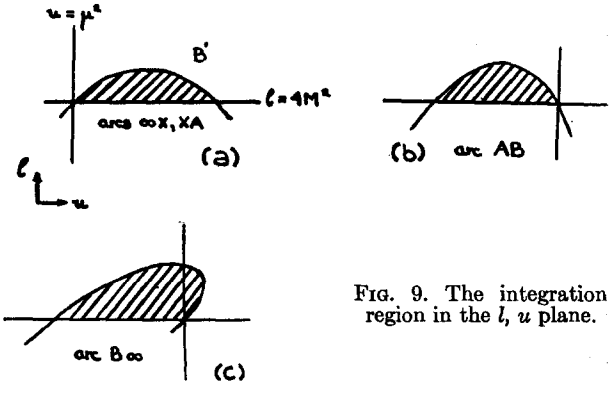


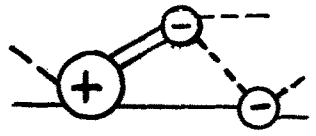
FIG. 9. The integration region in the l, u plane.

[8] Compare a similar discussion in the Appendix of J. C. Polkinghorne, Nuovo Cimento **25**, 901 (1962).

[9] This, we show in Ref. 3, is the correct form of the discontinuity near the $l = 4M^2$ normal threshold. This is appropriate for use in the discontinuity associated with L when we are very near to L; if we continue away from L the form (7) of the l-discontinuity may suffer distortion of its integration contour.

[10] D. I. Olive, Nuovo Cimento **29**, 326 (1963).

where $\psi(u, l) = \int_{4M^2}^{l} \phi(u, l)\, dl$ and has a singularity when the upper end point coincides with the singularity $l = 4M^2$ of ϕ, namely at $u = u_c$. So, in the u plane there is a pole $u = \mu^2$, and a branch point at $u = u_c$ on top of the end point of the integration. When $d\eta$ takes a circuit in the complex η plane, u_c goes round the pole and picks up the residue at the pole. The discontinuity of the bubbles with respect to the $l = 4M^2$ normal thresholds does not come in—the pole is independent of the sheet of these branch points. Again with due care about signs, the effect is that we must add the term



(10)

to the natural continuation.

Figure 2(e)

Labeling the momenta as in Fig. 10, we take the

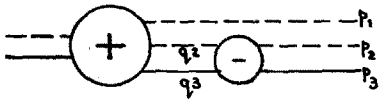
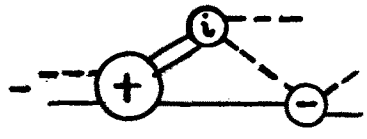


FIG. 10. Momenta for the mechanism of Fig. 2(e).

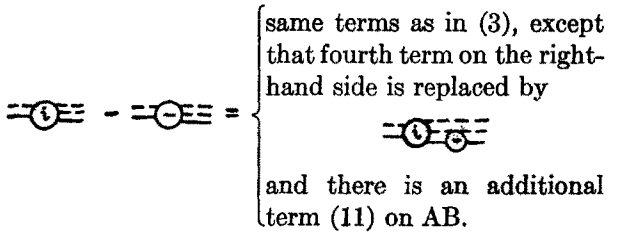
integration variables as $l = (p_1 + q_2)^2$ and others that we suppose have been integrated out. The icecream-cone singularity L arises when the end point of the l integration coincides with the normal thresholds $l = 4M^2$ of the left-hand bubble. In a continuation (5) across L with Im $d\eta < 0$, the end point passes above the normal threshold on the arcs ∞X, B∞ and so this continuation is natural on these arcs. On the other arcs, XA and AB, it differs from the natural continuation by an integral in which the left-hand bubble is replaced by its discontinuity across the $l = 4M^2$ cut. From (7) this integral is



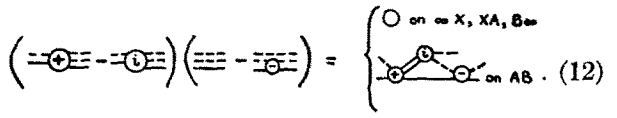
(11)

4. GETTING THE ANSWER

The terms (9), (10), and (11) cancel on the arc XA and are not present on the arcs ∞X, B∞. On the arc AB, only the term (11) is present. Hence the continuation of (3) from outside L to inside L via Im $d\eta < 0$ reads



Subtracting this from (3) and rearranging, we have

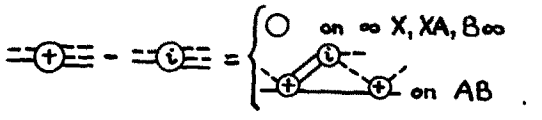


(12)

Postmultiplying by the factor

and using the unitarity relation

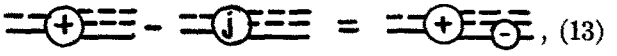
we obtain



So, just as in finite-order perturbation theory, the icecream-cone singularity appears only on the positive-α part of the arc.

5. NATURE OF THE SINGULARITY

The aim of this section is to show that the icecream-cone singularity is two-sheeted. We label by (j) the amplitude obtained by continuation round the normal threshold $\sigma = (M + m)^2$ and we assume the relation



(13)

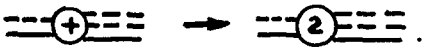
where on the right-hand side the integration is as it would be in an ordinary unitarity integral, provided that σ is sufficiently close to the normal threshold. In particular, this will be the case when σ is at a point P just above the part OB of the normal threshold (Fig. 5). Equation (13) is analogous to Eq. (7) but, unlike the latter, it refers to a singularity on the edge of the physical region. This means that it cannot be proven by the methods of Ref. 3 but must, at present, be taken as an explicit assumption. It can be shown[11] that (13)

[11] W. Zimmerman, Nuovo Cimento 21, 249 (1961); see also Ref. 6.

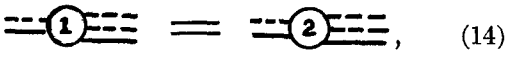
implies that the $\sigma = (M + m)^2$ normal threshold is two-sheeted. In addition to (13), we make use of our result that only the arc AB of L is singular on the physical Riemann sheet. The fact that the arc B∞ is nonsingular on the physical sheet has the important consequence that AB is not singular for the amplitude (j), since the (j) amplitude on AB is obtained from the physical amplitude on B∞ by continuing around the normal threshold, and the disc theorem of complex variable theory implies that the nonsingularity persists on continuation.[6]

We now continue (13) away from P, round the arc AB of L, and back again to P. We do this first in one sense around L, so that as a result of the continuation

say, and then in the other sense so that

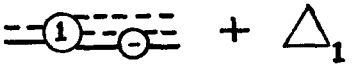


Our aim is to show that

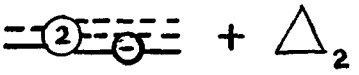

(14)

which will establish the two-sheetedness of L.

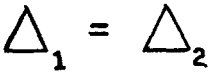
The continuations will have no net effect on the second term on the left of (13), because we have said that the arc AB is not singular for this term. The term on the right is, except for the labels on one of the bubbles, the same as that in Fig. 2(e), which we have already discussed. From the discussion there it is evident that the results of the two continuations are



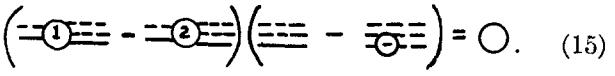
and



In direct analogy with (11), the functions Δ_1 and Δ_2 are integrals of the discontinuity of the left-hand bubble across the $l = 4M^2$ cut. Further, because the latter cut is assumed two-sheeted,



If now we subtract one continuation of (13) from the other, and rearrange, we have


(15)

This equation may be inverted in exactly the same way as (12) and the result is just (14).

Existence of Solutions of N/D Equations*

D. ATKINSON
Department of Physics, University of California, Berkeley, California
(Received 25 January 1966)

The manifold of solutions of certain classes of N/D equations is considered, under the restriction that the scattering amplitude have a uniform bound in all complex directions. Two cases are treated: (i) The N integral equation is Fredholm, but the associated homogeneous equation may have solutions; (ii) the equation has the marginally singular behavior characteristic of an asymptotically constant left-hand cut discontinuity. The existence theorem in the latter case proceeds by the construction of the appropriate resolvent.

1. INTRODUCTION

THE purpose of this paper is to investigate the existence of solutions of N/D equations.[1] In the cases in which the integral equations can be reduced to Fredholm form, the possibility that the associated homogeneous equation has a solution is considered.[2] The "marginally singular" behavior, in which $F(\nu) \sim \log \nu$, $\nu \to \infty$, is investigated by the explicit construction of the resolvent kernel of the resulting integral equation. Some properties of this resolvent, together with those of the solutions it generates, are adduced. This work complements that of Ref. 3, where similar resolvents are applied to certain physical problems.

It is unnecessary to consider equations with multiple subtractions, since it can be shown that an n-times-subtracted equation is equivalent to an unsubtracted equation with n *CDD* poles.[4] Accordingly, since *CDD* poles do not affect the Lebesgue classes of any of the functions or kernels considered in this paper, for brevity, all equations are written without such poles. Again, in an inelastic system, the Lebesgue classes are not altered if $\eta(\nu)$ or $R^{-1}(\nu)$ (functions describing the inelasticity[1,5]) have no finite zeros, and do not tend to zero too rapidly as $\nu \to \infty$.[4] Then an inelastic equation can be treated by the same methods as those given in this paper.

The question of zeros of the D function is not considered explicitly. They must occur neither in the complex plane, nor in the physical region. However, in the absence of essential singularities, there can only be a finite number of zeros, so that the additional constraints on the input have no implications regarding the Lebesgue classes of the permitted solutions.[6]

In Sec. 2, the possibilities are divided into the classes in which (a) the N equation is Fredholm; (b) the kernel is L^2 but the inhomogeneous term is not; (c) the equation is marginally singular. Sections 3 and 4 are devoted to this marginally singular case.

2. LEBESGUE CLASS OF N/D KERNELS

Consider the elastic scattering of two indistinguishable spinless bosons of mass unity; and let the partial-wave amplitude be $A_J(\nu)$, where ν is the momentum squared, and J the angular momentum. Suppose that the argument of $A_J(\nu)$ has limited total fluctuation in the physical region: $0 \leq \nu < \infty$. Then a decomposition

$$A_J(\nu) = \nu^{-J} N_J(\nu)/D_J(\nu) \tag{2.1}$$

exists,[7] in which $N_J(\nu)$ has the unphysical cut $-\infty < \nu \leq -1$, and $D_J(\nu)$ has the unitarity cut $0 \leq \nu < \infty$; moreover, neither N nor D has kinematical singularities. For simplicity, it is assumed that there are no elementary particles or bound states with the quantum numbers of the scattering channel, so that N has no poles and D has no zeros.

The function $N_J(\nu)$ satisfies

$$N_J(\nu) = F_J(\nu) + \frac{1}{\pi}\int_0^\infty d\nu'\, \nu'^J \rho(\nu') \times \frac{F_J(\nu') - F_J(\nu)}{\nu' - \nu} N_J(\nu'), \tag{2.2}$$

* This work was supported in part by the Air Force Office of Scientific Research, Grant No. AF-AFOSR-232-63.

[1] G. F. Chew and S. Mandelstam, Phys. Rev. 119, 467 (1960).

[2] D. Morgan, Nuovo Cimento 36, 813 (1965).

[3] D. Atkinson and A. P. Contogouris, Nuovo Cimento 39, 1082 (1965). A. P. Contogouris and D. Atkinson, Nuovo Cimento 39, 1102 (1965).

[4] See D. Atkinson and D. Morgan, Nuovo Cimento 41, 559 (1966).

[5] G. Frye and R. L. Warnock, Phys. Rev. 130, 478 (1963).

[6] A. P. Contogouris (unpublished); A. P. Contogouris and D. Atkinson, "N/D Equations, Bethe-Salpeter Models and High Energy Scattering" (to be published).

[7] R. Omnes, Nuovo Cimento 8, 316 (1958).

where

$$F_J(\nu) = \frac{1}{\pi}\int_{-\infty}^{-1}\frac{d\nu'}{\nu'-\nu}\,\mathrm{Im}\,[\nu'^{-J}A_J(\nu')] \qquad (2.3)$$

and the phase-space factor is $\rho(\nu) = [\nu/(\nu+1)]^{\frac{1}{2}}$. The denominator function is given in terms of a solution of (2.2) by

$$D_J(\nu) = 1 - \frac{1}{\pi}\int_0^\infty d\nu'\,\nu'^J\rho(\nu')\frac{N_J(\nu')}{\nu'-\nu}. \qquad (2.4)$$

Although the ensuing discussion is limited to integral J values, most of the results can be extended to any real J.

According to the unitarity condition for a physical amplitude,

$$\rho(\nu)\,|A_J(\nu)|^2 = \mathrm{Im}\,A_J(\nu) \le 1, \quad 0 \le \nu < \infty. \qquad (2.5)$$

It is supposed that a similar limit exists on the left-hand cut, namely

$$\lim_{\nu\to\infty}\mathrm{Im}\,A_J(\nu) \le 1. \qquad (2.6)$$

As may be shown from the Phragmén–Lindelöf theorem,[8] the following set of assumptions is sufficient to ensure that (2.5) implies (2.6):

$$A_J(\nu) \sim \nu^\alpha(\log\nu)^\beta(\log\log\nu)^\gamma\cdots \times(\log\log\cdots\log\nu)^\omega, \quad \nu\to\infty, \qquad (2.7)$$

$$A_J(\nu) \sim \nu^{\alpha'}(\log\nu)^{\beta'}(\log\log\nu)^{\gamma'}\cdots \times(\log\log\cdots\log\nu)^{\omega'}, \quad \nu\to-\infty,$$

$$|A_J(\nu)| < C_1 e^{(\pi-\epsilon)\nu}, \quad \epsilon>0, \qquad (2.8)$$

where $\alpha, \beta, \cdots, \omega, \alpha', \beta', \cdots, \omega'$, and C_1 are some constants, and (2.8) is supposed to hold in all complex directions.

Three classes which satisfy (2.6) are distinguished and treated separately:

$$\left.\begin{aligned}&\text{(a)}\quad |A_J(\nu)| < C_2\nu^{-\frac{1}{2}-\epsilon}\\ &\text{(b)}\quad |A_J(\nu)| < C_3\nu^{-\epsilon}\end{aligned}\right\}\ \text{some}\ \ \epsilon<0,$$

$$\text{(c)}\quad |A_J(\nu)| \underset{\nu\to\infty}{\to} \lambda \le 1. \qquad (2.9)$$

In this paper the possibility that $A_J(\nu)$ tends to zero at $\nu=\infty$ more slowly than any power of ν is not considered; although the case $F_J(\nu)\sim 1/\log\nu$ can be solved by an extension of the methods of Sec. 3.

Class (a)

Both the kernel and the inhomogeneous term of (2.2) are square-integrable for any $J \ge 0$, because

$$|F_J(\nu)| < C_4\int_1^\infty\frac{d\nu'}{\nu'+\nu}\nu'^{-\frac{1}{2}-J-\epsilon}, \quad \nu\ge 0$$

$$\le C_5\nu^{-\frac{1}{2}-\epsilon} \quad \text{all} \quad J\ge 0. \qquad (2.10)$$

From this, it follows that (2.2) is Fredholm, so that, unless there is a homogeneous solution, a unique square-integrable solution N_J exists. Moreover, the integral in (2.4) that defines D_J exists.

In the case that a homogeneous solution of (2.2) does exist, i.e., there is a function $n_J(\nu)$ such that

$$n_J(\nu) = \frac{1}{\pi}\int_0^\infty d\nu'\,\nu'^J\rho(\nu')\frac{F_J(\nu')-F_J(\nu)}{\nu'-\nu}n_J(\nu'), \qquad (2.11)$$

there will be, *in general*, no solution of (2.2) (the Fredholm alternative theorem). However, in such a case one can define a function

$$d_J(\nu) = -\frac{1}{\pi}\int_0^\infty d\nu'\,\nu'^J\rho(\nu')\frac{n_J(\nu')}{\nu'-\nu}. \qquad (2.12)$$

This specifies a new decomposition

$$A_J(\nu) = \nu^{-J}n_J(\nu)/d_J(\nu) \qquad (2.13)$$

satisfying the usual requirements, except that $d_J(\nu)\to_{\nu\to\infty}0$. In this case $A_J(\nu)$ is said to belong to a *CDD* class of negative index.[4] The solution n_J of (2.11) need not be unique, although the number of linearly independent solutions must be finite (the multiplicity of the corresponding eigenvalue).

Finally, there is the possibility that there are solutions $\hat{n}_J(\nu)$ of the associated homogeneous equation, but that, for any such,

$$\int_0^\infty d\nu\,F_J(\nu)\hat{n}_J(\nu) = 0. \qquad (2.14)$$

In this case a solution of (2.2) does exist, but it is not unique, since any homogeneous solution $n_J(\nu)$ may be added. This possibility is considered elsewhere[4] in connection with coincident zeros of N_J and D_J.

Class (b)

In this case

$$|F_J(\nu)| < C_6\nu^{-J-\frac{1}{2}\epsilon}, \qquad (2.15)$$

so that (2.2) is Fredholm for $J\ge 1$, and the considerations pertinent to class (a) apply. For $J=0$, $F_J(\nu)$ may not be square integrable, so that (2.2) is not Fredholm. However, the kernel is still L_2, for

[8] E. C. Titchmarsh, *The Theory of Functions* (Oxford University Press, New York, 1950).

$$\left|\frac{F_0(\nu) - F_0(\nu')}{\nu - \nu}\right|$$

$$< C_7 \int_1^\infty d\nu'' \frac{\nu''^{-\epsilon}}{(\nu'' + \nu)(\nu'' + \nu')}, \quad \nu, \nu' \geq 0$$

$$< C_7 \int_1^\infty d\nu'' \frac{\nu''^{-\epsilon}}{(\nu'' + \nu)\nu''^{\frac{1}{2}(1-\epsilon)}\nu^{\frac{1}{2}(1+\epsilon)}}$$

$$= C_8(\nu\nu')^{-\frac{1}{2}(1+\epsilon)}. \tag{2.16}$$

Hence (2.2) has a resolvent kernel; but a solution may not exist if the contraction of this resolvent with $F_0(\nu)$ fails to converge.

However, a solution may be found in this case by subtracting the D equation at, say, $\nu = 0$, and normalizing it to unity at this point. The modified integral equation is

$$N_0(\nu) = F_0(\nu) + \frac{1}{\pi}\int_0^\infty d\nu' \, \rho(\nu') \times \frac{\nu F_0(\nu) - \nu' F_0(\nu')}{\nu - \nu'} \frac{N_0(\nu')}{\nu'}, \tag{2.17}$$

with $F_0(\nu)$ defined as in (2.3), and $D_0(\nu)$ given by

$$D_0(\nu) = 1 - \frac{\nu}{\pi}\int_0^\infty d\nu' \, \rho(\nu') \frac{N_0(\nu')}{\nu'(\nu' - \nu)}. \tag{2.18}$$

The symmetrized form of (2.17) is

$$\frac{N_0(\nu)}{\nu^{\frac{1}{2}}} = \frac{F_0(\nu)}{\nu^{\frac{1}{2}}} + \frac{1}{\pi}\int_0^\infty d\nu' \, \rho(\nu') \times \frac{1}{(\nu\nu')^{\frac{1}{2}}} \frac{\nu F_0(\nu) - \nu' F_0(\nu')}{\nu - \nu'} \frac{N_0(\nu')}{\nu'^{\frac{1}{2}}}. \tag{2.19}$$

The inhomogeneous term is now square integrable, for

$$|F_0(\nu)/\nu^{\frac{1}{2}}| < C_6\nu^{-\frac{1}{2}-\epsilon}, \tag{2.20}$$

while the Lebesgue class of the kernel has not been changed by the subtraction procedure (an invariable property of N/D kernels). Hence a solution exists even for the S wave, involving no *CDD* parameters.

Class (c)

This is the "marginally singular" possibility in which the discontinuity tends to a constant at infinity. This kernel singularity, which can be specified loosely by saying that the norm diverges logarithmically, is the most singular behavior tolerated by unitarity. The additional assumption

$$\text{Im } A_J(\nu) = \lambda_J + r_J(\nu), \tag{2.21}$$

where $0 < \lambda_J \leq 1$ and $r_J(\nu) = O(\nu^{-\epsilon})$ with $\epsilon > 0$, is made.

For Class (c) amplitudes, it is more convenient to replace (2.1) by

$$A_J(\nu) = N_J(\nu)/D_J(\nu) \tag{2.22}$$

for all physical J, and then subtract both N_J and D_J equations at threshold. The resulting equations, which replace (2.2)–(2.4), are

$$\frac{N_J(\nu)}{\nu} = \frac{a_J}{\nu} + \frac{F_J(\nu)}{\nu} + \frac{1}{\pi}\int_0^\infty d\nu' \, \rho(\nu') \frac{F_J(\nu') - F_J(\nu)}{\nu' - \nu} \frac{N_J(\nu')}{\nu'}, \tag{2.23}$$

where

$$F_J(\nu) = \frac{\nu}{\pi}\int_{-\infty}^{-1} \frac{d\nu'}{\nu'(\nu' - \nu)} \text{Im } A_J(\nu') \tag{2.24}$$

and

$$D_J(\nu) = 1 - \frac{\nu}{\pi}\int_0^\infty d\nu' \, \rho(\nu') \frac{N_J(\nu')}{\nu'(\nu' - \nu)}. \tag{2.25}$$

For $J \geq 1$, the subtraction constant

$$a_J = -\frac{1}{\pi}\int_{-\infty}^{-1} \frac{d\nu'}{\nu'^2} \text{Im } A_J(\nu').$$

For $J > 1$ there are additional threshold conditions, while for $J = 0$, a_J constitutes an arbitrary parameter. Substituting (2.21) into (2.24) gives

$$F_J(\nu) = (\lambda_J/\pi) \log(\nu + 1) + C_J + O(\nu^{-\epsilon}), \tag{2.26}$$

where the constant C_J is given by

$$C_J = \frac{1}{\pi}\int_0^\infty \frac{d\nu'}{\nu'} r_J(\nu'). \tag{2.27}$$

If one writes

$$\begin{aligned} x &= \nu + 1, \\ \phi_J(x) &= N_J(\nu)/\nu, \\ \phi_J^0(x) &= \frac{a_J + F_J(\nu)}{\nu} \\ &= \frac{(\lambda_J/\pi)\log x + a_J + C_J + O(x^{-\epsilon})}{x - 1}, \\ \rho(x) &= 1 - \{(\nu + 1)^{\frac{1}{2}}[\nu^{\frac{1}{2}} + (\nu + 1)^{\frac{1}{2}}]\}^{-1}, \end{aligned} \tag{2.28}$$

Eq. (2.23) becomes

$$\phi_J(x) = \phi_J^0(x) + \frac{\lambda_J}{\pi^2}\int_1^\infty dx' \left[\frac{\log x'/x}{x' - x} + K_J(x, x')\right]\phi_J(x'), \tag{2.29}$$

where $K_J(x, x')$ is a square-integrable residual kernel. This singular equation may be written

$$\phi_J = \phi_J^0 + (\lambda_J/\pi^2)S\phi_J + (\lambda_J/\pi^2)K_J\phi_J, \tag{2.30}$$

where

$$S(x, x') = (\log x'/x)\cdot(x' - x)^{-1}$$

is a marginally singular kernel.

In the next section, it is shown that a resolvent kernel R_J exists and satisfies

$$R_J = \frac{1}{\pi^2} S + \frac{\lambda_J}{\pi^2} SR_J = \frac{1}{\pi^2} S + \frac{\lambda_J}{\pi^2} R_J S; \qquad (2.31)$$

moreover, this allows one to define solutions of (2.29) for any Fredholm kernel K_J.

3. CONSTRUCTION OF THE RESOLVENT

As the first stage in the solution of (2.29), set

$$\omega_J(x) = \phi_J^0(x) + \frac{\lambda_J}{\pi^2} \int_1^\infty dx'\, K_J(x, x')\phi_J(x') \qquad (3.1)$$

so that

$$\phi_J(x) = \omega_J(x) + \frac{\lambda_J}{\pi^2} \int_1^\infty dx' \frac{\log x'/x}{x' - x} \phi_J(x'). \qquad (3.2)$$

If one can construct a resolvent kernel R satisfying (2.31), then the general solution of (3.2), if it exists, satisfies

$$\phi_J(x) = \omega_J(x) + \lambda_J \int_1^\infty dx'\, R_J(x, x'; \lambda_J)\omega_J(x') + A\Psi_J(x), \qquad (3.3)$$

where A is an arbitrary constant, and $\Psi_J(x)$ is the general solution of the homogeneous equation

$$\Psi_J(x) = \frac{\lambda_J}{\pi^2} \int_1^\infty dx' \frac{\log x'/x}{x' - x} \Psi_J(x'). \qquad (3.4)$$

As may be verified from the properties of Legendre functions,[9] one solution of (3.4) is

$$\Psi_J(x) = P_{-s_0}(2x - 1), \qquad (3.5)$$

where

$$\lambda_J = \sin^2 \pi s_0$$

and

$$0 \leq \operatorname{Re} s_0 \leq \tfrac{1}{2}.$$

Suppose now that the "shifted Mehler transform" of $\omega_J(x)$ is defined by

$$\tilde{\omega}_J(s) = s \tanh \pi s \int_1^\infty dx\, \omega_J(x) P_{-\frac{1}{2}+is}(2x - 1). \qquad (3.6)$$

This has the unique inverse of

$$\omega_J(x) = \frac{1}{2} \int_{-\infty}^{\infty} ds\, \tilde{\omega}_J(s) P_{-\frac{1}{2}+is}(2x - 1). \qquad (3.7)$$

[9] *Higher Transcendental Functions*, A. Erdelyi, Ed., (McGraw-Hill Book Company, Inc., New York, 1953), Vol. 2.

Then, if $\omega_J(x)$ is such that $\tilde{\omega}_J(s)$ exists in part of the strip $-\frac{1}{2} < \operatorname{Im} s < \frac{1}{2}$, Eq. (3.2) can be diagonalized by Mehler transformation:

$$\tilde{\phi}_J(s) = \tilde{\omega}_J(s) + \lambda_J \operatorname{sech}^2 \pi s \tilde{\phi}_J(s). \qquad (3.8)$$

This equation can be solved for the Mehler transform $\tilde{\phi}_J(s)$. Inversion then yields

$$\phi_J(x) = \int_c ds\, \tilde{\omega}_J(s) \frac{P_{-\frac{1}{2}+is}(2x - 1)}{1 - \lambda_J \operatorname{sech}^2 \pi s}, \qquad (3.9)$$

where it may be verified that any contour C extending from $-\infty$ to $+\infty$ within the strip $-\frac{1}{2} < \operatorname{Im} s < \frac{1}{2}$ yields a solution of (3.2). By substituting the definition of $\tilde{\omega}_J(s)$ into (3.9), one has

$$\phi_J(x) = \omega_J(x) + \lambda_J \int_1^\infty dx'\, R_J(x, x'; \lambda_J)\omega_J(x'), \qquad (3.10)$$

where

$$R_J(x, x'; \lambda_J) = \frac{1}{2} \int_c ds \frac{s \tanh \pi s}{\cosh^2 \pi s - \lambda_J} \times P_{-\frac{1}{2}+is}(2x - 1)P_{-\frac{1}{2}+is}(2x' - 1). \qquad (3.11)$$

The integrand in (3.11) has poles whenever

$$\cosh^2 \pi s = \lambda_J. \qquad (3.12)$$

For $0 < \lambda_J < 1$, the general solution of (3.12) is

$$s = in \pm (i/\pi) \operatorname*{arc\,cos}_{(0, \frac{1}{2}\pi)} \lambda_J^{\frac{1}{2}}, \qquad n = \cdots -1, 0, 1, 2, \cdots. \qquad (3.13)$$

There are two poles inside the strip $-\frac{1}{2} < \operatorname{Im} s < \frac{1}{2}$, obtained by setting $n = 0$ in (3.13).

A particular resolvent $R_J(x, x'; \lambda_J)$ may be picked out by defining the contour C in (3.11) to be the real axis $(-\infty, \infty)$. It is easy to verify that R_J satisfies (2.27). Moreover, homogeneous solutions are obtained by taking C to be a closed contour encircling either of the poles in the strip $-\frac{1}{2} < \operatorname{Im} s < \frac{1}{2}$. These are multiples of $P_{-s_0}(2x - 1)$ or $P_{-1+s_0}(2x - 1)$, where s_0 is defined in (3.5), depending on which pole is taken. However,

$$P_{-s_0}(2x - 1) = P_{-1+s_0}(2x - 1). \qquad (3.14)$$

Hence, there is only one independent homogeneous solution. If C is taken to be a contour surrounding a pole outside the strip $-\frac{1}{2} < \operatorname{Im} s < \frac{1}{2}$, then the result is not a homogeneous solution $\Psi_J(x)$, since the integral in (3.4) would fail to converge.

Finally, the general solution of (3.2) can be written

$$\phi_J(x) = \omega_J(x) + \lambda_J \int_1^\infty dx\, R_J(x, x'; \lambda_J)\omega_J(x') + AP_{-s_0}(2x - 1). \quad (3.15)$$

That (3.15) is in fact a solution can be verified by checking to see if, when using Eqs. (3.1) and (2.28), the integral converges uniformly with respect to x. This justifies the Mehler transformation.

As a function of λ_J, $R_J(x, x'; \lambda_J)$ has a branch point at $\lambda_J = 1$. This follows from the fact that the real contour $C \equiv (-\infty, \infty)$ in (3.11) is pinched at $s = 0$ when $\lambda_J \to 1$. The cut may be defined $(1 \leq \lambda_J < \infty)$. The function $P_{-s_0}(2x - 1)$ is entire in s_0; but s_0 has branch points at $\lambda_J = 0, 1$. However, $P_{-s_0}(2x - 1)$ has a branch point at $\lambda_J = 0$ only: the branch cut $(1 \leq \lambda_J < \infty)$ vanishes identically, as a result of the relation (3.14). According to (3.15), the general solution will have branch cuts $(-\infty < \lambda_J \leq 0)$ and $(1 \leq \lambda_J < \infty)$ (assuming that the arbitrary factor A is chosen to be holomorphic in λ_J). However, one solution, defined by $A = 0$, has just one finite branch point $\lambda_J = 1$: this solution is analytic at $\lambda_J = 0$ and has a power series expansion about that point.

It is instructive to display the sheet structure of $R_J(x, x'; \lambda_J)$ in the λ_J variable. In fact, R_J is uniform in the variable s_0, defined as in (3.5), or equivalently,

$$s_0 = (i/\pi) \log [(1 - \lambda_J)^{\frac{1}{2}} + i\lambda_J^{\frac{1}{2}}]. \quad (3.16)$$

The first sheet of λ_J, cut $(-\infty, 0)$, $(1, \infty)$, is defined to map onto the strip $0 < \text{Im } s_0 < \frac{1}{2}$. In Fig. 1 the complete λ_J structure is shown mapped into the s_0 plane. The discontinuity of R_J across the cut $(1 \leq \lambda_J < \infty)$ is just the difference between the values of R_J at some point $\lambda_J = \frac{1}{2} + it_0$ (say) and at $\lambda_J = \frac{1}{2} - it_0$. The corresponding discontinuity across $(-\infty < \lambda_J \leq 0)$, the difference calculated from $\lambda_J = it_0$ to $\lambda_J = -it_0$ is zero, a consequence of the symmetry of the integrand in (3.11) with respect to s. This means that there is no cut $(-\infty, 0)$ on sheet I, a fact that has already been noted. Sheet II is defined to be the sheet connected to sheet I across the cut $(1 \leq \lambda_J < \infty)$: it maps onto the strip $\frac{1}{2} \leq \text{Im } s_0 \leq 1$. On this sheet there is a branch cut $(-\infty < \lambda_J \leq 0)$ as well as the cut $(1 \leq \lambda_J < \infty)$. This is true of all sheets except the first. As can be seen from Fig. 1, a double circuit of $\lambda = 1$ which does not encircle $\lambda = 0$ brings one back onto sheet I. On the other hand, repeated circuits enclosing $\lambda = 0$ and $\lambda = 1$ plunge one into lower and lower sheets. The branch points may be said to be individually like square roots, but together like a logarithm.

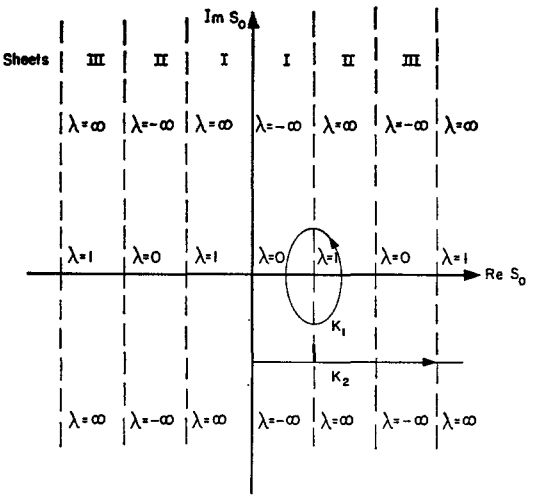


FIG. 1. Map of the infinite-sheeted λ plane into the entire s_0 plane. The contour K_1 is the map of a double circuit of $\lambda = 1$, not enclosing $\lambda = 0$. The line K_2 is the map of an infinite encirclement of both $\lambda = 0$ and $\lambda = 1$.

It should be noted that the continuation of $\phi_J(x)$ onto a higher sheet is not a solution of (3.2) in general, because the integral would diverge. The $N_J(\nu)$ and $D_J(\nu)$ functions that can be defined from the continued $\phi_J(x)$ correspond in fact to N/D solutions of higher CDD classes.[4]

4. COMPLETE SOLUTION

The problem to be investigated in this section is the existence of a solution of Eq. (2.29), corresponding to each solution (3.15) of (3.2). Suppose one writes Eq. (3.15) in operator form:

$$\phi_J = \omega_J + \lambda_J R_J \omega_J + AP_{-s_0}. \quad (4.1)$$

Then, substituting for ω_J according to (3.1),

$$\phi_J = \phi_J^0 + (\lambda_J/\pi^2)K_J\phi_J + \lambda_J R_J \phi_J^0 + (\lambda_J^2/\pi^2)K_J R_J \phi_J + AP_{-s_0}. \quad (4.2)$$

It was shown in Ref. 3 that R_J satisfies the family of majorizations

$$|R_J(x, x'; \lambda_J)| \leq B_1 x^{-\frac{1}{2}-p} x'^{-\frac{1}{2}+p} \quad (4.3)$$

for any $-\frac{1}{2} + s_0 \leq p \leq \frac{1}{2} - s_0$, with B_1 some constant. It follows from (2.26) and (2.28) that

$$|K_J(x, x')| \leq B_2 (xx')^{-\frac{1}{2}-\frac{1}{2}\epsilon}. \quad (4.4)$$

Then, from (4.3), (4.4), and (2.28),

$$\phi_J^0 + \lambda_J R_J \phi_J^0$$

is a square-integrable function and

$$K_J + \lambda_J K_J R_J \tag{4.5}$$

is a square-integrable kernel. Hence, in the case $A = 0$, (4.2) is a Fredholm equation, and a solution almost always exists. However,

$$P_{-s_0}(2x - 1) \sim x^{-s_0}, \qquad 0 < s_0 < \tfrac{1}{2}, \tag{4.6}$$

so that if $A \neq 0$, the inhomogeneous term in (4.2) is not square-integrable, and, in general, a solution will not exist. In fact, suppose that the resolvent of the Fredholm kernel $K_J + \lambda_J K_J R_J$ is $\mathfrak{F}_J$. If the contraction $\mathfrak{F}_J P_{-s_0}$ converges, then a solution of (4.2), and hence of (2.29), exists for any constant A.

If, for a given λ_J, the condition (2.21) can be strengthened to

$$r_J(\nu) = o(\nu^{-s_0-\epsilon}) \quad \text{for some} \quad \epsilon > 0, \tag{4.7}$$

one can show that $\mathfrak{F}_J P_{-s_0}$ always converges, so that there is always a one-parameter infinity of solutions. For in this case

$$|K_J(x, x')| < B_3 x^{-\frac{1}{2}-\eta} x'^{-\frac{1}{2}+\eta-s_0-\epsilon} \tag{4.8}$$

for any $-\frac{1}{2} + s_0 < \eta < \frac{1}{2}$. Then it follows from (4.3) and (4.8) that

$$\left|K_J(x, x') + \lambda_J \int_1^\infty dx'' K_J(x, x'') R_J(x'', x')\right| < B_4 x^{-\frac{1}{2}-p-s_0} x'^{-\frac{1}{2}+p} \tag{4.9}$$

for any $-\frac{1}{2} + s_0 \leq p \leq \frac{1}{2} - s_0$. If it is supposed that the Fredholm resolvent $\mathfrak{F}_J$, which is itself Fredholm, satisfies some power bound,

$$|\mathfrak{F}_J(x, x')| < B_4 x^{-u} x'^{-v} \quad \text{with} \quad u, \; v > \tfrac{1}{2}, \tag{4.10}$$

then (4.9) implies that, in fact, $u = \frac{1}{2} + p + s_0$, $v = \frac{1}{2} - p$, with $-\frac{1}{2} + s_0 \leq p \leq \frac{1}{2} - s_0$, as in (4.9). By choosing $p = -\frac{1}{2} + s_0$, one sees immediately that $\mathfrak{F}_J P_{-s_0}$ converges, so that (4.2) always has a solution.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge helpful discussions with Dr. A. P. Contogouris and with Dr. M. B. Halpern.

This work was done under the auspices of the U. S. Atomic Energy Commission.

The r-Particle Distribution Function in Classical Physics*

F. A. Blood, Jr.†
Brookhaven National Laboratory, Upton, New York
(Received 24 January 1966)

An expression for the r-particle distribution function ($r \geq 3$) for classical gases in terms of the pair potential is converted to an expression in terms of the pair correlation function by graph-theoretical means. The new formula involves a sum over all basic graphs with G-bonds, having r root points, where $G(\mathbf{r}) + 1$ is the pair correlation function.

INTRODUCTION

THE r-particle distribution functions for a classical system are normally expressed as a functional of the pair potential between particles.[1] A quantity which is more readily accessible to direct experimental determination[2] is the pair correlation function, $g(r)$, and it will be shown here how to express the r-particle distribution function as a functional of the radial distribution function.

The systems considered here are classical systems of identical particles, in equilibrium at temperature T, having a symmetric pair potential. If there are N particles in a volume V, then one knows from statistical mechanics that the relative probability of finding particle 1 in the volume element $d\mathbf{r}_1$ about $\mathbf{r}_1$, particle 2 in the volume element $d\mathbf{r}_2$ about $\mathbf{r}_2$, $\cdots$, and particle N in the volume element $d\mathbf{r}_N$ about $\mathbf{r}_N$, is

$$W_N(\mathbf{r}_1, \cdots, \mathbf{r}_N)\, d\mathbf{r}_1 \cdots d\mathbf{r}_N = W_N(1, \cdots, N)\, d\mathbf{r}_1 \cdots d\mathbf{r}_N = \exp[-\beta\Phi_N(1, \cdots, N)]\, d\mathbf{r}_1 \cdots d\mathbf{r}_N, \quad (1)$$

$$\Phi_N(1, \cdots, N) = \sum_{i<j=2}^{N} \varphi(|\mathbf{r}_i - \mathbf{r}_j|), \quad (1a)$$

where $\varphi(|\mathbf{r}_i - \mathbf{r}_j|)$ is the pair potential between particle i located at $\mathbf{r}_i$ and particle j located at $\mathbf{r}_j$, and where $\beta = 1/kT$.

The r-particle distribution functions, n_r, are defined so that

$$n_r(\mathbf{r}_1, \cdots \mathbf{r}_r)\, d\mathbf{r}_1 \cdots d\mathbf{r}_r = n_r(1, \cdots, r)\, d\mathbf{r}_1 \cdots d\mathbf{r}_r$$

is the probability of finding *any* particle in the volume element $d\mathbf{r}_1$ about $\mathbf{r}_1$, a second particle in the volume element $d\mathbf{r}_2$ about $\mathbf{r}_2$, $\cdots$, and an rth particle in the volume element $d\mathbf{r}_r$ about $\mathbf{r}_r$, irrespective of the positions of the $N - r$ other particles. Thus we see that

$$n_r(1, \cdots, r) = \frac{1}{Q_N} \frac{N!}{(N-r)!} \int \cdots \int d\mathbf{r}_{r+1} \cdots d\mathbf{r}_N \times W_N(1, \cdots, N), \quad (2)$$

$$Q_N = \int \cdots \int d\mathbf{r}_1 \cdots d\mathbf{r}_N\, W_N(1, \cdots, N). \quad (2a)$$

If we let N and V go to infinity, keeping the density $\rho = N/V$ constant, the expression (2) for n_r can be put in an entirely different form (while keeping n_r a functional of φ) by the use of graph theory. This form is given at the beginning of Sec. 3, after the definitions necessary for our graph theory work are given in Sec. 1. It involves sums of weights over all $S^{(r)}$ graphs having f-bonds, where

$$f(|\mathbf{r}_i - \mathbf{r}_j|) = f_{ij} = \exp[-\beta\varphi(|\mathbf{r}_i - \mathbf{r}_j|)] - 1. \quad (3)$$

Next in Sec. 3, we give another form for n_r, expressing it as an explicit functional of the pair correlation function. This form involves sums of weights over all $B^{(r)}$ graphs having G-bonds, where

$$G(|\mathbf{r}_i - \mathbf{r}_j|) = G_{ij} = g_{ij} - 1 = \rho^{-2} n_2(|\mathbf{r}_i - \mathbf{r}_j|) - 1. \quad (4)$$

Here, g_{ij} is the pair correlation function.

A similar statement was advanced but not proved by van Leeuwen, Groeneveld, and deBoer[3] in their paper on the hypernetted chain theory, namely that $\exp(-\beta\varphi_{12})$ could be expressed as an explicit functional of the pair correlation function. These statements are proved at the end of Sec. 3 with the help of the lemmas developed in Sec. 2.

* Work supported in part by the U. S. Atomic Energy Commission.

† Most of the work on this paper was done while the author was a National Science Foundation Cooperative Fellow (1964–65) at Case Institute of Technology, Cleveland, Ohio.

[1] G. E. Uhlenbeck and G. W. Ford, *Studies in Statistical Mechanics*, J. De Boer and G. E. Uhlenbeck, Eds. (North-Holland Publishing Company, Amsterdam, 1962), Vol. 1, Part B. The work in this reference is a graph-theoretical restatement of the original work by J. E. Meyer and E. Montroll, J. Chem. Phys. 9, 2 (1941).

[2] M. D. Johnson, P. Hutchinson, and N. H. March, Proc. Roy. Soc. (London) **A282**, 283 (1964).

[3] J. M. J. van Leeuwen, J. Groeneveld, and J. De Boer, Physica 25, 792 (1959).

1. TERMINOLOGY AND PROPERTIES OF GRAPHS

In this section, we introduce some specific terminology associated with the graphs that will be used in establishing the results outlined in the Introduction. This terminology, while in many respects related to that employed in other standard works (which is not always uniform), differs from it in details, and has been specifically tailored to the needs which follow. The following definitions should, therefore, be read with care by anyone interested in the details of the proofs of the succeeding sections.

Abstract graph. An abstract graph with *n points*, *r root points*, and $k = n - r$ *field points*, is an entity defined by three sets of elements:

(a) A set of elements, r in number, called *root points*.

(b) A set of elements, $k = n - r$ in number, called *field points*.

(c) A subset (proper or improper) of the set of all (unordered) pairs of distinct points (whether root or field points) called *links*.

An abstract graph may be given a pictorial representation by indicating root points by open circles, field points by filled circles (dots), and links by lines connecting "linked" pairs of points. An example is given in Fig. 1(a) of an abstract graph with two root points, three field points, and seven links.

In order to deal with a graph symbolically, it is necessary to associate with each element of the graph a symbol (label) to identify it. As in all such cases, it is expedient to distinguish between the abstract element itself and the label which identifies it, the latter corresponding to some "coordinatization" of the graph. This is of particular importance for us, since we shall be involved with processes which permute some of the labels of the abstract elements.

The nature of the symbols which are used to label the points of a graph is of no intrinsic significance, provided only that the labels serve to identify and distinguish points. It is often convenient to select them to be simply the first n integers. In this case, we reserve the first r integers to label the root points and the remaining $n - r$ integers to label the field points. With these remarks we then define a labeled graph

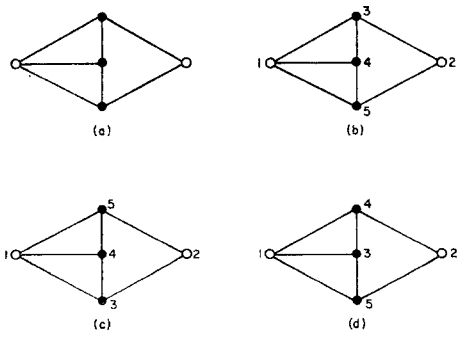


FIG. 1. Abstract and labeled graphs.

Labeled graph. A labeled graph is an abstract graph in which the points carry identifying labels. One may, however, regard a labeled graph as a graph in its own right defined by a set of (labeled) root points, a set of (labeled) field points, and a set of links which can now be explicitly symbolized by giving the labels of the linked pairs of points. Thus the graph of Fig. 1(b) is a labeled graph in which the set of root points is [1, 2], the set of field points is [3, 4, 5], and the set of links is [(1,3), (1,4), (1.5), (2,3), (2,5), (3,4), (4,5)].

Equality of labeled graphs. We define two labeled graphs to be equal if their set of root points, their set of field points, and their set of links, are identical. Thus the graph of Fig. 1(c) is equal to that of Fig. 1(b) as may be seen by writing down these sets for the graph 1(c) and comparing it with the sets written down above for 1(b). On the other hand, the graph in Fig. 1(d) is not equal to that in 1(b) as can be seen by carrying out the same process; in particular, the link (2,4) occurs in 1(d) but not in 1(b).

Generic equivalence of labeled graphs. Two labeled graphs are defined to be *generically equivalent*, if and only if, some permutation of the labels of the *field points alone* of one yields a graph equal to the other. (Note that the labels of the root points are *not* permuted.) Thus in Fig. 1, the labeled graphs (b), (c), and (d) are all generically equivalent, but in this case (b) and (c) are also equal while (b) and (d) are not equal.

Invariance group of a labeled graph: i-permutations; symmetry number. If one considers the totality of permutations of the labels of the field points of a labeled graph, these form a group of order $k!$; namely, the symmetric group on k-objects, where k is the number of field points. This group will contain a subgroup (not necessarily proper) consisting of those permutations which carry the labeled graph into one *equal* to the original; such permutations are called *i-permutations* of the graph. The group of i-permutations of a labeled graph is called its *invariance (or symmetry) group*, and the order of the invariance group is called the *symmetry number*

of the graph. The invariance groups of generically equivalent graphs are clearly isomorphic and have the same symmetry number. Also, $k!$ divided by the symmetry number is the number of *distinct* labeled graphs which can be obtained from a given labeled graph under all $k!$ permutations of the labels of its field points; two graphs are *distinct* if they are *unequal*.

Generic graph. It is often convenient to consider the totality of graphs which are all generically equivalent to constitute a single entity which is called a *generic* graph. A generic graph can be represented by a diagram of the same character as an abstract or labeled graph except that only the root points are labeled. Clearly a labeled graph and a generic graph can be considered *associated* according as the second can be derived from the first by removing the labels from the field points, or the first derived from the second by assigning appropriate labels to the field points.

It will be convenient from this point on to reserve the unqualified word "graph" for a labeled graph, and to refer to generic or abstract graphs by explicit qualification. Furthermore, in succeeding definitions, we shall sometimes define a property of a graph which is purely "topological" in the sense that it is a property of the points and links which is independent of the specific labeling of the points. In such cases, the definitions will be phrased for labeled graphs, but the generalization is usually obvious for a generic or abstract graph.

Chain. A *chain* between the points i_1 and i_s of a graph is an ordered subset $[i_1, i_2, \cdots i_s]$ of *distinct* points of the graph such that there is a link between each pair of adjacent points of the ordered subset. Thus, in Fig. 1(b), the chains between the point 1 and the point 4 are [1, 4], [1, 5, 4], [1, 3, 4], [1, 5, 2, 3, 4], [1, 3, 2, 5, 4]. The point i_1 is called the *origin* of the chain and the point i_s is called its terminus.

Loop. A *loop* is an ordered subset $[i_1, i_2, \cdots, i_s]$ of points of a graph which would be a chain except for the fact that its origin and terminus coincide, i.e., $i_1 = i_s$. In this case any point of a loop can be considered its origin (and terminus).

The points of a chain other than its origin and terminus will be called *interior* points of the chain, and the chain will be said to pass *through* such points. In the case of a loop, we will say that the loop passes *through* all of its points, including the one which is considered to be its origin and terminus. It should also be noted that according to our definitions, if two points i and j are linked, there exists at least one loop passing through them, namely the loop $\{i, j, i\}$.

Since the totality of the chains of a graph or the totality of loops of a graph completely determine the links of a graph, two graphs will be equal, if, and only if, they have the same set of root points, the same set of field points, and the same set of chains or loops.

Connected graph. A connected graph is a graph for which there is a chain connecting all distinct pairs of points of the graph. Thus graph (a) of Fig. 2 is not connected, but graph (b) of the same figure is connected.

Articulation point. An *articulation* point of a *connected* graph is a point with the property that all the remaining points of the graph can be divided into two disjoint nonempty subsets A and B such that every chain from a point of A to a point of B passes through the point a.

Star graph. A *connected* graph with the property that none of its points are articulation points is called a *star* graph. It is readily seen, and shown, that a graph is a star graph if, and only if, there is a loop through every pair of points of the graph.

Primitive graph. A *primitive* graph is a connected graph with at least *two* root points, such that it would become a star graph if links were inserted between every pair of root points which are unlinked in the original graph.

Simple graph. A *simple* graph is a *primitive* graph having no links connecting pairs of root points and with the further property that between any two points there exists at least one chain not passing *through* a root point.

We need to define subclasses of simple graphs in the case that there are two root points. The appropriate definitions follow:

Nodal point; nodal graphs. A field point of a *simple* graph with *two* root points with the property that every chain between the two root points passes through this point is called a *nodal* point. A simple graph with two root points which contains a nodal point is called a *nodal* graph.

Elementary graph. A simple graph with two root points which contains no nodal points is called an *elementary* graph.

There are a number of concepts between this

point and the end of Sec. 2 which apply to simple graphs if there are three or more root points, but only to elementary graphs if there are just two root points. To avoid a number of parenthetical remarks, we make the convention that whenever a simple graph is referred to between here and the end of Sec. 2, it must also be elementary if it has only two root points. For instance, the definition of nonbasic and basic points to follow applies only to elementary and not to nodal graphs.

Nonbasic point; basic point. A *nonbasic* point m is a *field* point of a *simple* graph with the property that there exist two points, p and q (distinct from m and not *both* root points) such that every chain with m as its origin and a root point as its terminus contains either p *or* q as an interior point or as a terminus. A *field* point of a *simple* graph which is not nonbasic is called a basic point. In Fig. 2(g), the point 4 is a basic point while 5, 6, 7, 8, 9 are nonbasic. Clearly a basic point must be linked to at least three other points of the graph.

Basic graph. A *basic* graph is a simple graph all of whose field points are basic points.

In what follows it will often be convenient to designate primitive, simple, and basic graphs with r root points by $P^{(r)}$, $S^{(r)}$, and $B^{(r)}$, respectively, and to designate nodal and elementary graphs by $N^{(2)}$ and $E^{(2)}$, respectively.

Subgraph. A *subgraph* of a graph consists of a proper subset of the points of the original graph together with all links of the original graph which connect pairs of points belonging to the subset.

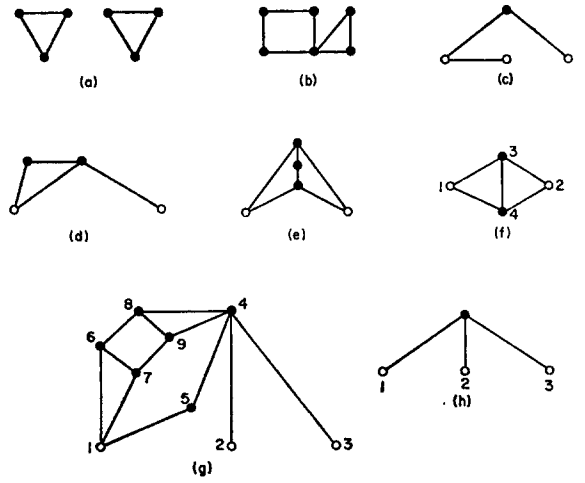
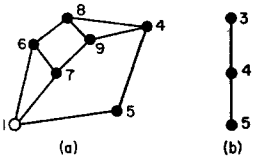


FIG. 2. Examples of types of graphs (a) not connected; (b) connected, not star; (c) primitive, not simple; (d) nodal, not elementary; (e) elementary, not basic; (f) basic; (g) simple, not basic; and (h) basic.

Total $i-j$ subgraph. Let i and j be two points of a simple graph such that either both are basic points, or one is a basic point and the other a root point. Then the *total* $i-j$ *subgraph* of the simple graph is the subgraph consisting of the points i and j together with all points m such that every chain with m as origin and a root point as terminus contains either i or j, but excluding such a graph which consists only of the points i and j with no link between them. Thus the total 1–4 subgraph of the graph 2(g) is the graph 3(a); but the graph 3(b) is not a total 3–5 subgraph of the graph 1(b).

FIG. 3. Subgraphs.



Basic graph derived from a simple graph. We may derive another graph from a *simple* graph by replacing each total i–j subgraph of the simple graph by a single link connecting i and j. The resultant graph will be called the *basic graph derived from the simple graph.* Thus the basic graph derived from the graph 2(g) is the graph 2(h). It can be shown that the basic graph derived from a simple graph is itself a *basic* graph. What this means is that it not only contains as field points only basic points of the simple graph, but that each such field point is a basic point of the *resultant* graph.

Basic equivalence of simple labeled graphs. Two simple graphs are *basically equivalent* if, and only if, a permutation of the labels on *nonbasic* points alone of one yields a graph equal to the other.

Weight associated with a generic graph. The application of graph theory to the considerations to follow lies in the fact that we shall be dealing with certain functionals each of which can be associated with a specific generic graph. The functionals with which we shall deal all have a particular form and we shall confine our remarks to these.

In our considerations, the points of a graph are identified with the positions of the identical particles constituting the physical system under study. Each link of a graph is associated with a specific function of the spatial separation of the two particles whose points are linked. Such a function we call a *bonding function*, and if the bonding function associated with the link (i, j) of the graph is the function $f(\mathbf{r}_{ij}) = f_{ij}$, with $\mathbf{r}_i$ and $\mathbf{r}_j$ the positions of the ith and jth particles and $\mathbf{r}_{ij}$ their separation, then we shall say

that we have an f-bond; if the function is $G(\mathbf{r}_{ij}) = G_{ij}$, we say we have a G-bond.

Now, let $\mathfrak{s}$ be some generic graph with its root points labeled from 1 to r and with $k = n - r$ field points. Consider an associated labeled graph $\mathfrak{s}(L_1)$ where L_1 represents some particular labeling of the field points from $r + 1$ to n. The *weight* that will be associated with such a graph, if the links correspond to f-bonds, will be a function $w_{\mathfrak{s}}(\mathbf{r}_1, \cdots, \mathbf{r}_r; f)$ defined by

$$w_{\mathfrak{s}}(1 \cdots r; f) = [\rho^r/t(\mathfrak{s})] \int \cdots \int d\mathbf{r}_{r+1} \cdots d\mathbf{r}_n \prod f_{ij}, \tag{5}$$

where ρ is a parameter (generally the density of the system in particles per unit volume), $t(\mathfrak{s})$ is a pure number called the weighting number, dependent upon the generic graph $\mathfrak{s}$, the integrals extend over all space, and the product of the bonding functions $\prod f_{ij}$ includes one factor f_{ij} for every (i, j) link which occurs in the generic graph. Since the variables $\mathbf{r}_{r+1}$ to $\mathbf{r}_n$ occurring in the integrand are dummy variables, the weight of a generic graph is independent of the particular labeling (L_1) of the graph that has been selected. The *weighting number* $t(\mathfrak{s})$ will often, but not always, be just the symmetry number $s(\mathfrak{s})$ of the generic graph, and if it is, we will designate the weight by $\omega_{\mathfrak{s}}(1, 2, \cdots, r; f)$, i.e.,

$$\omega_{\mathfrak{s}}(1, \cdots, r; f) = [\rho^r/s(\mathfrak{s})] \int \cdots \int d\mathbf{r}_{r+1} \cdots d\mathbf{r}_n \prod f_{ij}. \tag{5a}$$

Thus, for example, the graph of Fig. 2(f) has associated with it the weight

$$w_{\mathfrak{s}}(1, 2; f) = [\rho^2/t(\mathfrak{s})] \iint d\mathbf{r}_3\, d\mathbf{r}_4\, f_{13}f_{12}f_{23}f_{24}f_{34},$$

and if $w_{\mathfrak{s}} = \omega_{\mathfrak{s}}$, then $t(\mathfrak{s}) = s(\mathfrak{s}) = 2$.

There are some specific results which follow from these definitions which form the foundation of the arguments to follow. It is clear from the definition of a total $i - j$ subgraph that, if we consider the points i and j of such a subgraph to be root points, then such a subgraph is always a $P^{(2)}$ graph. Consider the set of all $S^{(r)}$ graphs. With each there will be associated its derived $B^{(r)}$ graph. In general, many different $S^{(r)}$ graph may give rise to the same derived $B^{(r)}$ graph, but the set of all such derived graphs will be the set of all $B^{(r)}$ graphs. Conversely, beginning with any $B^{(r)}$ graph, we may derive a graph by replacing one or more of its links with $P^{(2)}$ graphs. The set of all $S^{(r)}$ graphs derived in this way from the set of all $B^{(r)}$ graphs will clearly be the set of *all* $S^{(r)}$ graphs. We see, in general, that two or more unequal but generically equivalent $S^{(r)}$ graphs can be derived from the same $B^{(r)}$ graph, but two $S^{(r)}$ graphs derived from generically inequivalent $B^{(r)}$ graphs must be generically inequivalent. Our main interest will be in the number of *basically* inequivalent $S^{(r)}$ graphs derived from the same (labeled) $B^{(r)}$ graph.

The pertinence of these remarks for what follows arises from the fact that an expression which can be written as a sum of weights over the set of all generic $S^{(r)}$ graphs, can be rewritten as a sum of weights associated with generic $B^{(r)}$ graphs in which all possible substitutions of $P^{(2)}$ graphs have been made. In order to show that the two expressions are the same, we need to consider the number of basically inequivalent $S^{(r)}$ graphs derived from the same $B^{(r)}$ graph. In the following section we consider this problem.

2. INVARIANCE GROUPS AND RELATED GROUPS OF GRAPHS

The problem posed at the end of the previous section will be attacked by an exploration of the invariance groups of graphs and of some related groups. The necessary results will be given as lemmas; these are usually sufficiently obvious that only schematic proofs need be given.

Since the difference between basic and nonbasic points of a simple graph are characterized by distinct "topological" properties of the chains connecting them to root points, we have immediately:

Lemma 1. Under an i-permutation of an $S^{(r)}$ graph, basic point labels are carried into basic point labels and nonbasic point labels into nonbasic point labels. Thus, every i-permutation is the direct product of a permutation of basic point labels and a permutation of nonbasic point labels.

From the last property, it follows immediately that the invariance group $G(S^{(r)})$ of a simple graph contains an invariant subgroup $H(S^{(r)})$ consisting of those i-permutations which leave the labels of basic points unchanged. Now, any i-permutation belonging to the invariant subgroup H has by virtue of the "topological" connection of points belonging to a total $i - j$ subgraph of $S^{(r)}$ the property described by the following lemma:

Lemma 2. An i-permutation belonging to H carries the labels of points belonging to a total $i - j$ subgraph into points belonging to the same $i - j$ subgraph. From this follows:

Lemma 3. The group H is the direct product of the invariance groups G_{ij} of the total $i - j$ subgraphs of $S^{(r)}$ if i and j are considered the root points of the total $i - j$ subgraph. The order of the group H, which we denote by $h(S^{(r)})$, is therefore equal to the product πs_{ij} of the symmetry numbers of the total $i - j$ subgraphs of $S^{(r)}$.

The existence of the invariant subgroup $H(S^{(r)})$, allows a separation of the elements of the group $G(S^{(r)})$ into "cosets" which in turn can be regarded as the elements of the "factor group" $G(S^{(r)})/H(S^{(r)})$. Two elements of G belong to the same coset provided the product of one by the inverse of the other is an element belonging to H. The elements of H form one coset (which is the identity element of the factor group G/H) and the membership of every coset is equal to the order of the group H, that is h. The number of cosets f is, therefore, equal to $s(S^{(r)})/h(S^{(r)}) = s(S^{(r)})/\pi s_{ij}$. The important result whose derivation we shall now sketch is that

$$ns(S^{(r)})/\pi s_{ij} = fn = s(B^{(r)}) \tag{6a}$$

or

$$s(S^{(r)}) = s(B^{(r)})\pi s_{ij}/n, \tag{6b}$$

where n is the number of basically inequivalent $S^{(r)}$ graphs obtained from $P^{(2)}$ substitutions in the derived $B^{(r)}$ graph.

Lemma 4. If the derived graph associated with the graph $S^{(r)}(L_1)$ is the (labeled) graph $B^{(r)}(1)$, then the number, n, of basically inequivalent $S^{(r)}$ graphs whose derived graph is $B^{(r)}(1)$ is $s(B^{(r)})/f$, where f is the number of cosets of the symmetry group of $S^{(r)}(L_1)$.

Let $S^{(r)}(L)$ be any labeled $S^{(r)}$ graph whose derived graph is $B^{(r)}(1)$. Then

$$S^{(r)}(L) = N(L)b(L)S^{(r)}(L_1),$$

where $N(L)$ is a permutation of nonbasic points alone and $b(L)$ is a permutation of basic points alone. As L runs through all the graphs whose derived gaph is $B^{(r)}(1)$, the number of distinct $b(L)$ will be n, because two graphs with the same $b(L)$ will be basically equivalent. One can see that the set of all $b(L)$ form a subset of $G(B^{(r)})$, because they leave the derived $B^{(r)}$ graph invariant.

Any i-permutation, $P(j)$, of $S^{(r)}(L_1)$ can also be written as $M(j)a(j)$ where $M(j)$ is a permutation of basic points alone. All those i-permutations $P(j)$ for which the $a(j)$ are equal will be in the same coset of $G(S^{(r)})$, and conversely if $a(j) \neq a(k)$, then $P(j)$ and $P(k)$ will be in different cosets so the number of distinct $a(j)$ is f. The permutations, a, form a subset, A, of $G(B^{(r)})$.

To show that $nf = s(B^{(r)})$, we will show that each element of $G(B^{(r)})$ can be written as ba in one and only one way. Let c be any element of $G(B^{(r)}(1))$. Then $cS^{(r)}(L_1)$ has $B^{(r)}(1)$ as its derived graph and thus $cS^{(r)}(L_1)$ must be basically equivalent to $bS^{(r)}(L_1)$ for some b. Thus, there exists a permutation, M, of nonbasic points alone, such that $McS^{(r)}(L_1) = bS^{(r)}(L_1)$ or $Mb^{-1}cS^{(r)}(L_1) = S^{(r)}(L_1)$. Therefore, there exists an a such that $b^{-1}c = a$ or $c = ba$, so that every element of $G(B^{(r)})$ can be expressed as ba.

Now let us assume $ba = b'a'$ or $b'^{-1}b = a'a^{-1}$. Then $b'^{-1}b$ belongs to A so $b'^{-1}bS^{(r)}(L_1)$ is basically equivalent to $S^{(r)}(L_1)$ which implies that $b'^{-1}b$ is the identity permutation. Thus, $b' = b$, which in turn implies $a' = a$, and the proof is finished.

3. PROOF OF THE RESULT

For our starting point, in proving that $n_r (r \geq 3)$ can be given as an explicit functional of g or equivalently of n_2, we will give a form for n_r which is derived in reference 1:

$$n_r(1, \cdots, r) = \rho^r W_r(1, \cdots, r) \exp\Big[\sum_{\{m\}\leq\{r\}} h_m(\{m\}; f)\Big], \tag{7}$$

where ρ is the density of particles, W is defined in Eq. (1), and the sum indicates that $\{m\}$ runs over all subsets of $\{1, 2, \cdots, r\}$ containing at least two numbers. The subscript m gives the number of labels in the set $\{m\}$. For instance, if $r = 2$,

$$\sum h_m(\{m\}; f) = h_2(1, 2; f)$$

and if $r = 3$, $\sum h_m(\{m\}; f) = h_2(1, 2; f) + h_2(1, 3; f) + h_2(2, 3; f) + h_3(1, 2, 3; f)$. The quantity $h_r(1, \cdots, r; f)$ is defined by

$$h_r(1, \cdots, r; f) = \sum_{S^{(r)}} \omega_{S^{(r)}}(1, \cdots, r; f), \tag{8}$$

where $\omega_{S^{(r)}}$ is defined in Eq. (5a), f is defined in Eq. (3), and the sum is over all generic $S^{(r)}$ graphs.

We will change the form of Eq. (7) slightly in order to get rid of W_r, which depends explicitly on the potential. We see from (7) that $n_2(1, 2) = \rho^2 \exp(-\beta\phi_{12}) \exp[h_2(1, 2; f)]$ so that

$$n_r(1, \cdots, r) = \rho^{-r}\Big(\prod_{i<j=2}^{r} n_2(i, j)\Big) \times \exp\Big[\sum_{\{m\}\leq\{r\}}{}' h_m(\{m\}; f)\Big], \tag{9}$$

where the prime on the sum indicates that $\{m\}$ runs

over all subsets of $\{1, 2, \cdots, r\}$ containing at least *three* points.

We now claim that for $r \geq 3$, $h_r(1, \cdots, r; f)$ can be written as an explicit functional of g. That functional is

$$h_r(1, \cdots, r; f) = B_r(1, \cdots, r; G) = \sum_{B^{(r)}} \omega_{B^{(r)}}(1, \cdots, r; G), \quad (10)$$

where G is defined in Eq. (4) and the sum is over all generic $B^{(r)}$ graphs.

An expression similar to (10) arises when one seeks to express $\exp(-\beta\varphi_{12})$ as an explicit functional of the two-particle distribution function, and an attempt to prove this result was the starting point for this paper. We have

$$\exp(-\beta\phi_{12}) = \rho^{-2} n_2(1, 2) \exp[-h_2(1, 2; f)], \quad (11)$$

where

$$h_2(1, 2; f) = \sum_{N^{(2)}} \omega_{N^{(2)}}(1, 2; f) + \sum_{E^{(2)}} \omega_{E^{(2)}}(1, 2; f) = N(1, 2; f) + E(1, 2; f) \quad (12)$$

in which the sum $\sum_{N^{(2)}}$ is over all generic nodal graphs and the sum $\sum_{E^{(2)}}$ is over all generic elementary graphs. As explained in Ref. 3, the quantity $N(1, 2; f)$ can be obtained from the integral equation

$$N(1, 2; f) = N_{12} = \rho \int d\mathbf{r}_3 \, (G_{13} - N_{13}) G_{32}, \quad (13)$$

which, by iteration, becomes

$$N_{12} = \rho \int d\mathbf{r}_3 \, G_{13} G_{32} - \rho^2 \iint d\mathbf{r}_3 \, d\mathbf{r}_4 \, G_{13} G_{34} G_{42} + \rho^3 \iiint d\mathbf{r}_3 \, d\mathbf{r}_4 \, d\mathbf{r}_4 \, G_{13} G_{34} G_{45} G_{52} + \cdots,$$

so that N_{12} is an explicit functional of g. One can also solve for N when given G in the Fourier transform space.

The authors of the hypernetted chain paper (Ref. 3) then state that $E(1, 2; f)$ can be re-expressed as

$$E(1, 2; f) = B_2(1, 2; G) = \sum_{B^{(2)}} \omega_{B^{(2)}}(1, 2; G), \quad (14)$$

where the sum runs over all generic $B^{(2)}$ graphs. Thus we are led back to an equation of just the form of Eq. (10). The only difference is that $h_2(1, 2; f)$ is replaced by $E(1, 2; f)$, and this is the reason for the anomalous treatment of the $r = 2$ case in Secs. 1 and 2.

We will now sketch the proofs of Eqs. (10) and (14). Again, in order to avoid a number of parenthetical remarks about the $r = 2$ case, we make the convention that when an $S^{(r)}$ graph is referred to, it must also be elementary if $r = 2$. The two main steps in the proof are to show that $B_r(1, \cdots, r; G)$ is a sum of weights over all generic $S^{(r)}$ graphs with *f-bonds*,

$$B_r(1, \cdots, r; G) = \sum_{S^{(r)}} \omega_{S^{(r)}}(1, \cdots, r; f), \quad (15)$$

and to show that the weighting number $t(S^{(r)})$ associated with the weight in the above formula is $s(S^{(r)})$.

The function G can be expressed as

$$G_{12} = \rho^{-2} \sum_{P^{(2)}} \omega_{P^{(2)}}(1, 2; f), \quad (16)$$

where the sum is over all generic $P^{(2)}$ graphs. Thus the set of all graphs with *f-bonds* contributing to B_r must be the set of all generic $S^{(r)}$ graphs, for substitution of $P^{(2)}$ graphs in all possible ways into basic graphs gives the set of all generic $S^{(r)}$ graphs, and so Eq. (15) holds.

If a graph $S^{(r)}$ whose derived basic graph is $B^{(r)}$ were to occur only once in B_r, then the $t(S^{(r)})$ associated with the weight in (15) would be $s(B^{(r)}) \pi s_{ij}$, where s_{ij} is the symmetry number of the ijth subgraph of $S^{(r)}$. In general, however, $S^{(r)}$ occurs not just once but n times so that

$$t(S^{(r)}) = S(B^{(r)}) \pi s_{ij} / n. \quad (17)$$

A little reflection will show that n is just the number of basically inequivalent $S^{(r)}$ graphs associated with $B^{(r)}$ so that Eqs. (6b) and (17) imply $t(S^{(r)}) = s(S^{(r)})$ and the proof is complete.

Since we have proved Eq. (10), we see that n_r can be written as an explicit functional of g. For if we substitute (10) into (9), we obtain

$$n_r = \rho^{-r} \left(\prod_{i<j}^{r} n_2(i, j) \right) \times \exp\left[\sum_{\{m\} \leq \{r\}}' B_r(1, \cdots, r; G) \right].$$

Similarly, by proving Eq. (14), we have also proved that the pair potential can be expressed as an explicit functional of the pair correlation function, as one can see by substituting (12), (13), and (14) into (11).

ACKNOWLEDGMENT

The author would like to thank Dr. L. L. Foldy for his valuable assistance on this paper.

N Degenerate Modes

L. W. BRUCH*
Institute for Exploratory Research, U. S. Army Electronics Command, Fort Monmouth, New Jersey
(Received 13 January 1966)

The effect of a bilinear coupling in a system of degenerate modes is discussed. The normal mode frequencies are obtained for some simple couplings. If each mode is coupled to all the other degenerate modes, the maximum shift in normal mode frequency due to the coupling increases with the number of degenerate modes. Using the same methods, a many-boson problem and a truncated version of it can be compared. The energy eigenvalues have the same relationship as the exact and first-order perturbation excitation energies of the Bogoliubov model Hamiltonian for bosons.

INTRODUCTION

STRAIN field scattering[1] provides an interaction among phonons in an imperfect crystal. Some of the modes may be degenerate in energy due to symmetries of the crystalline lattice. The degeneracy requires special treatment in calculations of normal mode frequencies using perturbation theory.

In this work, the effect of bilinear coupling in a system of N degenerate modes is studied. The problem of determining the normal mode frequencies reduces to a matrix eigenvalue problem. Some couplings among the phonons are assumed for which this eigenvalue problem can be exactly solved. The maximum shift in frequency due to the coupling increases with the number of degenerate levels. There is a qualitative difference in the eigenvalue spectrum for uniform real or complex coupling coefficients.

The solution of this matrix eigenvalue problem is also relevant to a many-boson system. The phonon problem, in position and momentum operators, is transformed to boson creation and annihilation operators. The quasi-particle energies of the resulting problem are the normal mode frequencies of the phonon problem. An approximation, truncating this Hamiltonian by neglect of terms which change the total number of bosons, reduces to a similar matrix eigenvalue problem. The solution of this has the same relation to the exact solution as an "energy gap" solution has to the exact solution of the Bogoliubov Hamiltonian for interacting bosons.[2,3]

Section I presents the Hamiltonians which are discussed and the secular equations determining the normal mode frequencies. In Sec. II, these equations are solved approximately for some models, and, in Sec. III, the relation with a boson system is discussed.

I. HAMILTONIANS AND EQUATIONS OF MOTION

The Hamiltonian for a system of phonons whose interaction is produced by the displacements of a static strain field is assumed to be[1,4]

$$H = \sum_k \omega_k(a_k^+ a_k + \tfrac{1}{2}) + \tfrac{1}{2}\sum_{kl} C_{kl}(a_k + a_{-k}^+)(a_l + a_{-l}^+). \quad (1)$$

This interaction is derived from the first anharmonic term in the lattice potential and is linear in the strain field. ω_k is the energy[5] of a phonon of wave vector $\mathbf{k}$ in the absence of all anharmonic interactions, with $\omega_k = \omega_{-k}$. The summations are over wave vectors in the first Brillouin zone.

C_{kl} is a coefficient depending on the details of the strain field and the anharmonic coefficients. The explicit form is given by Carruthers[1,4]; general properties of C_{kl} which will be used are

$$C_{-kk} = 0, \qquad C_{kl} = C_{lk}, \qquad C_{kl}^* = C_{-k,-l}. \quad (2)$$

a^+ and a are creation and annihilation operators for phonon modes and obey Bose commutation relations,

$$[a_k, a_l^+] = \delta_{k,l}, \qquad [a_k, a_l] = 0. \quad (3)$$

It is convenient to express H in terms of the complex Fourier components of the momentum and position coordinates,[6]

* Present address: Department of Physics, University of Wisconsin, Madison, Wisconsin.

[1] P. Carruthers, Phys. Rev. 114, 995 (1959); P. Carruthers, Rev. Mod. Phys. 33, 92 (1961) and references there.

[2] N. N. Bogoliubov, J. Phys. U.S.S.R. 11, 23 (1947).

[3] K. A. Brueckner and K. Sawada, Phys. Rev. 106, 1117 (1957).

[4] The interaction differs slightly from that of Carruthers (Ref. 1) in keeping terms not conserving the number of phonons. The form used by Carruthers is denoted by H_T here. The C_{kl} used here is Carruthers' $C_{k,-l}$.

[5] $\hbar = 1$, m = atomic mass = 1 in the units used here.

[6] See, for instance, C. Kittel, *Quantum Theory of Solids* (John Wiley & Sons, Inc., New York, 1963), Chap. 2.

$$q_k = (2\omega_k)^{-\frac{1}{2}}(a_k + a_{-k}^+),$$
$$p_k = i^{-1}(\tfrac{1}{2}\omega_k)^{-\frac{1}{2}}(a_{-k} - a_k^+). \tag{4a}$$

These have the properties

$$[q_k, q_l] = 0, \quad [p_k, p_l] = 0, \quad [q_k, p_l] = i\delta_{k,l}. \tag{4b}$$

In terms of these coordinates

$$H = \tfrac{1}{2}\sum_k (\omega_k^2 q_k q_{-k} + p_k p_{-k}) + \sum_{kl} C_{kl}(\omega_k\omega_l)^{\frac{1}{2}} q_k q_l. \tag{5}$$

An equation determining the frequencies of the normal modes is obtained by using the Heisenberg equation of motion twice[7]:

$$i\dot{q}_k = [q_k, H] = ip_{-k},$$
$$-\ddot{q}_k = \omega_k^2 q_k + 2\sum_l C_{-kl}(\omega_k\omega_l)^{\frac{1}{2}} q_l. \tag{6}$$

If the normal model has frequency Ω,

$$0 = (\omega_k^2 - \Omega^2)q_k + 2\sum_l C_{-kl}(\omega_k\omega_l)^{\frac{1}{2}} q_l,$$

and the secular equation for Ω is

$$\det(M - \Omega^2 1) = 0 \tag{7a}$$

with

$$M_{kl} = \omega_k^2\delta_{k,l} + 2C_{-kl}(\omega_k\omega_l)^{\frac{1}{2}}. \tag{7b}$$

This is the same secular equation as would be obtained by applying equation-of-motion methods to Eq. (1) to determine the quasi-particle energies of H.

Since the trace of a matrix is invariant under unitary transformation,

$$\sum_k \omega_k^2 = \sum_i \Omega_i^2, \tag{8}$$

the second sum going over all normal modes.

It is interesting to compare Eq. (7) with the equation for the eigenvalues of a truncated version of the Hamiltonian H. In this truncation the terms in H which change the total number of bosons are dropped; the resulting Hamiltonian H_T is

$$H_T = \sum_k \omega_k(a_k^+ a_k + \tfrac{1}{2}) + \sum_{kl} C_{kl}a_{-k}^+ a_l. \tag{9}$$

The equation-of-motion method[7] here gives

$$i\dot{a}_k = [a_k, H_T] = \omega_k a_k + \sum_l C_{-kl}a_l.$$

Thus the eigenstate energies Ω_T are given by a secular equation similar to Eq. (7):

$$\det(N - \Omega_T 1) = 0;$$
$$N_{kl} = \omega_k\,\delta_{k,l} + C_{-kl}. \tag{10}$$

This will be discussed further in Sec. III.[8]

II. DEGENERATE PERTURBATION THEORY

The detailed solution of the secular equation for a physical strain field coupling is beyond the scope of this work. However, qualitative features of the effect of coupling on the energy spectrum of a system of degenerate modes can be obtained; some models will be solved in the degenerate subspace.

In the absence of interaction, there is assumed to be some degeneracy in energy of the modes. The lowest-order effect, in perturbation theory, of the interaction in the degenerate modes will be treated.[9] This is evaluated by solving the secular equation, Eq. (7), in the space of degenerate modes, neglecting coupling to other modes. These are the N degenerate modes of the title.

In lowest order then, the effect of the strain field is given by the secular equation for the $N \times N$ matrix:

$$\det\{(\omega^2 - \Omega^2)\,\delta_{k,l} + 2C_{-kl}\omega\} = 0. \tag{11}$$

The models for which this will be solved are divided into constant coupling (real and complex) and cyclic coupling among the degenerate modes.

A. Constant Coupling

The general model is that each degenerate mode is coupled to all others with a constant strength. This strength is allowed to be real, imaginary, or complex.

For real constant coupling,

$$C_{-kl} = \begin{cases} C, & k \neq l, \\ 0, & k = l, \end{cases} \tag{12}$$

the secular equation is

$$\begin{vmatrix} \omega^2 - \Omega^2 & 2\omega C & 2\omega C & \cdots & 2\omega C \\ 2\omega C & \omega^2 - \Omega^2 & 2\omega C & \cdots & 2\omega C \\ & \cdots & \cdots & & \\ 2\omega C & 2\omega C & & \cdots & \omega^2 - \Omega^2 \end{vmatrix} = 0.$$

This is the determinant of a cyclic matrix. The matrix can be diagonalized by the unitary trans-

[7] D. Pines, *The Many-Body Problem* (W. A. Benjamin Company, Inc., New York, 1961), p. 44, reviews the method and gives references to the original papers.

[8] See also, A. Bardasis, D. S. Falk, and D. A. Simkin, J. Chem. Phys. Solids 26, 1269 (1965) who have used a uniform method for phonon and interacting boson systems.

[9] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (Addison-Wesley Publishing Company, Inc., Reading, Mass., 1958), Chap. 6.

formation[10,11]

$$U_{jl} = (N)^{-\frac{1}{2}} e^{i2\pi jl/N}, \qquad j, l = 0, 1, \cdots, N-1, \tag{13}$$

where j and l are indices labeling rows and columns of the matrix. The resulting eigenvalues are

$$\Omega^2 = \begin{cases} \omega^2 + (N-1)2\omega C, & 1 \text{ mode}, \\ \omega^2 - 2\omega C, & (N-1) \text{ modes}. \end{cases} \tag{14}$$

This satisfies Eq. (8).

For imaginary constant coupling,

$$C_{-kl} = \begin{cases} -iC, & k > l, \\ iC, & k < l, \\ 0, & k = l, \end{cases} \qquad C \text{ real}; \tag{15}$$

the secular equation is

$$\begin{vmatrix} \omega^2 - \Omega^2 & i2C\omega & i2C\omega & \cdots & i2C\omega \\ -i2C\omega & \omega^2 - \Omega^2 & i2C\omega & \cdots & i2C\omega \\ & \cdots & \cdots & & \\ -i2C\omega & -i2C\omega & & \cdots & \omega^2 - \Omega^2 \end{vmatrix} = 0. \tag{16}$$

The diagonalizing transformation for this matrix is obtained by a generalization of the method for the cyclic matrix.[12] The constant antisymmetric $N \times N$ matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ -1 & 0 & 1 & \cdots & 1 \\ & & \cdots & & \\ -1 & -1 & -1 & \cdots & 0 \end{bmatrix} \tag{17}$$

can be expressed in powers of an $N \times N$ generator matrix Γ

$$\Gamma = \begin{bmatrix} 0 & 1 & 0 & & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & & \cdots & 1 \\ -1 & 0 & 0 & & \cdots & 0 \end{bmatrix}, \tag{18}$$

as

$$A = \Gamma + \Gamma^2 + \cdots + \Gamma^{N-1}. \tag{19}$$

Γ satisfies the operator identity.

$$\Gamma^N = -1. \tag{20}$$

Hence, the eigenvalues of Γ are

$$\gamma_l = \exp[i(\pi/N)(2l+1)], \qquad l = 0, 1, \cdots, N-1, \tag{21a}$$

with associated eigenvectors[13]

$$\psi_l = (N)^{-\frac{1}{2}}(1, \gamma_l, \gamma_l^2, \cdots, \gamma_l^{N-1}). \tag{21b}$$

Then the diagonalizing transformation for A is

$$U_{kl} = (N)^{-\frac{1}{2}} \exp[i(\pi/N)k(2l+1)], \qquad k, l = 0, 1, \cdots, N-1. \tag{22}$$

Since the eigenvectors of Γ are simultaneous eigenvectors of Γ^2, Γ^3, $\cdots$, and Γ^{N-1}, the eigenvalues of A are

$$\lambda_l = \gamma_l + \gamma_l^2 + \cdots + \gamma_l^{N-1} = i \cot[(\pi/2N)(2l+1)]. \tag{23}$$

Hence the roots of Eq. (16) are

$$\Omega_l^2 = \omega^2 - 2\omega C \cot[(\pi/2N)(2l+1)], \qquad l = 0, 1, \cdots, N-1. \tag{24}$$

The complex constant coupling is a generalization of this case:

$$C_{-kl} = \begin{cases} Ce^{-i\varphi}, & k > l, \\ Ce^{i\varphi}, & k < l, \\ 0, & k = l, \end{cases} \tag{25}$$

where C is real and $\varphi \neq 0, \pi, 2\pi, \cdots$. The secular equation is

$$\begin{vmatrix} \omega^2 - \Omega^2 & 2C\omega e^{i\varphi} & \cdots & 2C\omega e^{i\varphi} \\ 2C\omega e^{-i\varphi} & \omega^2 - \Omega^2 & \cdots & 2C\omega e^{i\varphi} \\ & \cdots & & \\ 2C\omega e^{-i\varphi} & 2C\omega e^{-i\varphi} & \cdots & \omega^2 - \Omega^2 \end{vmatrix} = 0. \tag{26}$$

In analogy to Eqs. (17), (18), the $N \times N$ matrix B,

$$B = \begin{bmatrix} 0 & e^{i\varphi} & e^{i\varphi} & \cdots & e^{i\varphi} \\ e^{-i\varphi} & 0 & e^{i\varphi} & \cdots & e^{i\varphi} \\ & & \cdots & & \\ e^{-i\varphi} & e^{-i\varphi} & e^{-i\varphi} & \cdots & 0 \end{bmatrix}, \tag{27a}$$

is expressed in powers of its generator g

$$g = \begin{bmatrix} 0 & e^{i\varphi} & 0 & \cdots & 0 \\ 0 & 0 & e^{i\varphi} & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & & \cdots & e^{i\varphi} \\ e^{-i\varphi} & 0 & & \cdots & 0 \end{bmatrix} \tag{27b}$$

[10] P. O. Löwdin, R. Pauncz, and J. deHeer, J. Math. Phys. 1, 461 (1960).
[11] P. O. Löwdin, J. Chem. Phys. 21, 496 (1953).
[12] G. G. Hall, *Matrices and Tensors* (The Macmillan Company, New York, 1963), Chap. 4.
[13] These are column eigenvectors (right eigenvectors) of the generator matrix, written on one line only for compactness.

as

$$B = g + e^{-i\varphi}g^2 + e^{-2i\varphi}g^3 + \cdots + e^{-i(N-2)\varphi}g^{N-1}. \quad (27c)$$

The operator identity

$$g^N = [\exp i(N-2)\varphi]1 \quad (28)$$

implies the eigenvalues of g are

$$\delta_l = \exp\{i[\varphi + (2/N)(\pi l - \varphi)]\}, \quad l = 0, 1, \cdots, N-1, \quad (29a)$$

with eigenvectors[13]

$$X_l = (N)^{-\frac{1}{2}}(1, e^{-i\varphi}\,\delta_l, e^{-i2\varphi}\,\delta_l^2, \cdots, e^{-i(N-1)\varphi}\,\delta_l^{N-1}). \quad (29b)$$

The diagonalizing transformation for B is

$$U_{kl} = (N)^{-\frac{1}{2}} \exp[i(2k/N)(l\pi - \varphi)], \quad k, l = 0, 1, \cdots, N-1. \quad (30)$$

The eigenvalues of B are

$$\mu_l = \gamma_l + e^{-i\varphi}\gamma_l^2 + \cdots + e^{-i(N-2)\varphi}\gamma_l^{N-1} = \sin[\varphi - N^{-1}(\varphi - \pi l)]/\sin N^{-1}(\varphi - \pi l). \quad (31)$$

Hence the roots of Eq. (26) are

$$\Omega_l^2 = \omega^2 + 2\omega C \sin[\varphi - N^{-1}(\varphi - \pi l)]/\sin N^{-1}(\varphi - \pi l), \quad l = 0, 1, \cdots, N-1. \quad (32)$$

For these three cases, the maximum shift in Ω^2 increases with increasing N.

B. Cyclic Coupling

In this paragraph it is assumed that

$$C_{-kl} = C(\tau_k - \tau_l) \quad (33)$$

and

$$C(\tau \pm N) = C(\tau), \quad (34)$$

where an integer $\tau_k (0 \leq \tau_k \leq N-1)$ is assigned to each wave vector $\mathbf{k}$ in the degenerate subspace. The periodicity condition, Eq. (34), expresses an assumed cyclic nature of the coupling index. The dependence in Eq. (33) might occur if the wave-vector dependence of C is primarily determined by the Fourier components of the strain field displacement and the possible wave-vector differences are multiples of a basic vector.[14]

With these conditions, the secular equation, Eq. (11), reduces to the eigenvalue problem for a cyclic matrix. The diagonalizing transformation,[10,11] Eq. (13), applied to this, gives the eigenvalues

$$\Omega_l^2 = \omega^2 + 2\omega \sum_{k=0}^{N-1} C(k) \exp(-i2\pi kl/N), \quad l = 0, 1, \cdots, N-1. \quad (35)$$

As an example, for a dependence $C(k)$

$$C(k) = \begin{cases} C\cos 2\pi k/N, & k \neq 0 \\ 0, & k = 0 \end{cases}, \quad C \text{ real}, \quad (36a)$$

the eigenvalues are

$$\Omega_l^2 = \begin{cases} \omega^2 + (N-2)\omega C, & l = 1; l = N-1, \\ \omega^2 - 2\omega C, & \text{otherwise.} \end{cases} \quad (36b)$$

An example of complex coupling is

$$C(k) = iB\sin 2\pi k/N, \quad B \text{ real}, \quad (37a)$$

with eigenvalues

$$\Omega_l^2 = \begin{cases} \omega^2 + N\omega B, & l = 1, \\ \omega^2 - N\omega B, & l = N-1, \\ \omega^2, & \text{otherwise.} \end{cases} \quad (37b)$$

The example in Eq. (36) has strongest coupling for phonons of comparable index τ_k and weakest for phonons whose τ's are separated by half the range of τ_k. The qualitative features of the result should remain for a more general cyclic coupling, if the coupling energy has only one oscillation in the range of the index.

C. Discussion

The examples thus far have shown a maximum shift in Ω^2 which, for large N, is proportional to the number of degenerate modes. It is possible to give an example[12] which does not show this behavior, but which indicates the feature of the examples responsible for the effect.

This example is a coupling model in which the degenerate phonons are coupled only to their "nearest neighbors" in the coupling index. This system has a secular equation:

$$\begin{vmatrix} \omega^2 - \Omega^2 & 2C & 0 & 0 & \cdots & 0 \\ 2C & \omega^2 - \Omega^2 & 2C & 0 & \cdots & 0 \\ 0 & 2C & \omega^2 - \Omega^2 & 2C & \cdots & 0 \\ & & \cdots & & & \\ 0 & 0 & \cdots & & 2C & \omega^2 - \Omega^2 \end{vmatrix} = 0, \quad (38a)$$

[14] An indication of the limitations on the coupling implied by these assumptions is that there are now only N coupling coefficients in the problem rather than $\frac{1}{2}N(N-1)$.

where C is real and is the coupling strength. The matrix is of the form known as a pseudocirculant and has the eigenvalues[12]

$$\Omega_l^2 = \omega^2 + 4C \cos [\pi l/(N+1)],$$
$$l = 1, 2, \cdots, N. \quad (38b)$$

Here the maximum shift in Ω^2 is weakly dependent on N for large N.

The basic feature producing the large relative shift, then, is the interaction with $N-1$ other modes. The example of Eq. (36) shows this need not be constant in sign. However, if the coupling is sufficiently limited in range, the large relative shifts do not occur.

The splitting of the degeneracy in the examples was most pronounced in the cases when the coupling was complex. In these cases,

$$C_{kl} \neq C_{-k,-l}.$$

Such an interaction can be obtained when the strain field does not have inversion symmetry. For complex constant coupling, the degeneracy was removed completely.

III. COMPARISON WITH INTERACTING BOSONS

In Sec. I, the phonon Hamiltonian was also written in terms of boson operators, and a truncated version of this was introduced. The eigenvalues of the truncated Hamiltonian H_T are the solutions of a secular equation similar to that for the eigenvalues of H. In the approximation of Sec. II, evaluating the effect of the coupling in degenerate states, the relation between the problems can be made precise.

In the same approximation as Eq. (11), the secular equation for the eigenvalues of H_T in the degenerate subspace becomes the determinant of an $N \times N$ matrix:

$$\det \{(\omega - \Omega_T)\, \delta_{k,l} + C_{-kl}\} = 0. \quad (39)$$

Eq. (11) can be written

$$\det \{[(\omega^2 - \Omega^2)/2\omega]\, \delta_{k,l} + C_{-k,l}\} = 0. \quad (11')$$

Thus there is a correspondence between the eigenvalues of the two problems:

$$(\Omega_T - \omega) = (\Omega^2 - \omega^2)/2\omega \quad (40)$$

and Ω and Ω_T can be represented in terms of a common parameter, α, as

$$\Omega = (\omega^2 + \omega\alpha)^{\frac{1}{2}};$$
$$\Omega_T = \omega + \tfrac{1}{2}\alpha. \quad (41)$$

Ω_T has the leading terms of an expansion of Ω in powers of α. An analogous relation is found in treatments[3] of the Bogoliubov Hamiltonian[2] for interacting bosons. The first-order perturbation treatment of that Hamiltonian gives a single-particle energy

$$\epsilon_1 = \epsilon_0 + \tfrac{1}{2}\beta,$$

while in the exact treatment of the model (including "ring diagrams"),

$$\epsilon = (\epsilon_0^2 + \epsilon_0\beta)^{\frac{1}{2}},$$

where ϵ_0 is the single-particle energy in the absence of interaction and β is a parameter related to the interaction.

The details of C_{kl} have not been required to obtain Eq. (41). The result follows exactly, once the coupling is restricted to the degenerate subspace. Eq. (40) can be used to obtain the energies Ω_T from Ω^2 for the coupling models discussed in Sec. II.

IV. CONCLUSIONS

The effect of coupling on the frequencies of degenerate modes has been calculated in lowest approximation. The coupling models discussed are artificial but suggest a general behavior: if the coupling links each mode with all others in the degenerate space, the maximum shift in frequency due to the coupling will be greater the larger the degeneracy.

A connection has been noted between this system and a system of interacting bosons. Some of the mathematical techniques used are extensions of techniques used in the theory of cyclic hydrocarbons.[11] These connections might be useful in future work on these problems.

Homogeneous Dust-Filled Cosmological Solutions*

DAVID L. FARNSWORTH† AND ROY P. KERR
The University of Texas, Austin, Texas
(Received 8 March 1966)

The Einstein field equations with incoherent matter are discussed for the case of homogeneous space–time, i.e., for metrics allowing a four-parametric, simply transitive group of motions. It is proved that the only universes satisfying the above are those of Einstein, Gödel, and Ozsvath.

I. INTRODUCTION

LET M be a four-dimensional Lorentzian manifold, and let G be a group of smooth (C^∞) isometries, multiplication being written to the left

$$\sigma(m) = \sigma m, \qquad \sigma \in G, \quad m \in M.$$

We assume, following Ozsvath,[1] that G acts simply transitively on M; that is, given any two points, $m, m' \in M$, there exists exactly one transformation $\sigma \in G$ such that $\sigma m = m'$. This implies that the transformation $\bar{m}_0 : G \to M$ defined by

$$\bar{m}_0(\sigma) = \sigma m_0,$$

where m_0 is an arbitrary fixed point of M, is one–one and onto. Since it is a smooth map, it must be a diffeomorphism. This proves that we can (and do) identify M with the homogeneous space G. Under the identification map $\bar{m}_0$ the transformation $\sigma \in G$ becomes $L_\sigma : G \to G$, where

$$\begin{aligned} L_\sigma(\rho) &= \bar{m}_0^{-1}\sigma\bar{m}_0(\rho) \\ &= \bar{m}_0^{-1}\sigma(\rho m_0) \\ &= \sigma\rho; \end{aligned}$$

that is, L_σ is just the usual operation of left multiplication with σ.

$$\begin{array}{ccc} M & \xrightarrow{\sigma} & M \\ \uparrow \bar{m}_0 & & \uparrow \bar{m}_0 \\ G & \xrightarrow{L_\sigma} & G \end{array}$$

To recapitulate, M is identified with the homogeneous space G, with G acting as a group of transformations on itself by left multiplication.

Let G_σ be the tangent plane at $\sigma \in M$ and let dL_σ be the differential of the map L_σ,

$$dL_\sigma : G_\rho \to G_{\sigma\rho}.$$

A vector field $\mathbf{X}$ is said to be left invariant if

$$dL_\sigma \mathbf{X}(\rho) = \mathbf{X}(\sigma\rho).$$

These vector fields are completely determined by their values at the identity e of G,

$$\mathbf{X}(\sigma) = dL_\sigma \mathbf{X}(e)$$

since, if we define $\mathbf{X}$ by the equation

$$dL_\rho \mathbf{X}(\sigma) = dL_\rho\, dL_\sigma \mathbf{X}(e) = dL_{\rho\sigma} \mathbf{X}(e) = \mathbf{X}(\rho\sigma);$$

that is, $\mathbf{X}$ is left invariant. Furthermore, if $\mathbf{X}$, $\mathbf{Y}$ are both left invariant vector fields then their Lie bracket,

$$[\mathbf{X}, \mathbf{Y}] = \mathbf{XY} - \mathbf{YX}$$

is also left invariant,

$$dL_\rho [\mathbf{X}, \mathbf{Y}]_\sigma = [dL_\rho \mathbf{X}, dL_\rho \mathbf{Y}]_{\rho\sigma} = [\mathbf{X}, \mathbf{Y}]_{\rho\sigma}.$$

The algebra of such vector fields is said to be a Lie algebra. Let Φ be the set of all left invariant vector fields. From the above discussion, it is seen that $\dim \Phi = \dim G_\sigma = 4$. Let $\mathbf{e}_a (a = 0, 1, 2, 3)$ be a basis for Φ. The $\mathbf{e}_a{}^{(\sigma)}$ will then naturally give a basis for G_σ, being linearly independent. Let $\epsilon^a(\sigma)$ be the dual basis for $G_\sigma{}^*$; that is, each ϵ^a is a linear differential form (covariant vector) and if we define the usual scalar product between $\alpha \in G_\sigma{}^*$, $\mathbf{X} \in G_\sigma$ by

$$(\alpha, \mathbf{X}) = \alpha_u X^u = \alpha(\mathbf{X}),$$

then

$$(\epsilon^a, \mathbf{e}_b) = \delta_b{}^a. \tag{1.1}$$

Since $[\mathbf{e}_a, \mathbf{e}_b] \in \Phi$, it can be written as a linear combination of the $\mathbf{e}_a$; that is, there exists a set of 24 constants $C^a{}_{bc}$, the constants of structure, such that

$$[\mathbf{e}_a, \mathbf{e}_b] = C^e{}_{ab}\mathbf{e}_e. \tag{1.2}$$

* This work supported in part by the Office of Scientific Research (Grant 454-65), USAF.
† Present address: Department of Mathematics, University of Pittsburgh, Pittsburgh, Pennsylvania.
[1] I. Ozsvath, J. Math. Phys. 6, 590 (1965).

Any two-form β may be considered as a skew-symmetric bilinear map

$$\beta : V \times V \to F,$$

where $F(V)$ is the vector space of all smooth functions (vector fields). We write $\beta(\mathbf{X}, \mathbf{Y})$, where $\beta(\mathbf{X}, \mathbf{Y}) = -\beta(\mathbf{Y}, \mathbf{X})$. This is defined for a simple two-form $\alpha^1 \wedge \alpha^2$ by

$$\alpha^1 \wedge \alpha^2(\mathbf{X}_1, \mathbf{X}_2) = \det(\alpha^i(\mathbf{X}_j)).$$

In particular, if $d\alpha$ is the exterior derivative of α, any one-form, then it may be proved that

$$d\alpha(\mathbf{X}, \mathbf{Y}) = \mathbf{X}\alpha(\mathbf{Y}) - \mathbf{Y}\alpha(\mathbf{X}) - \alpha([\mathbf{X}, \mathbf{Y}]). \tag{1.3}$$

This may be proved by observing that it is fairly trivial if $\alpha = fdg$ and that any one-form may be written as a linear combination of such terms, or equivalently, by using the tensor formalism on both sides of the equation. If we substitute $\alpha = \epsilon^a$, $\mathbf{X} = \mathbf{e}_b$, $\mathbf{Y} = \mathbf{e}_c$ into Eq. (1.3), and use Eqs. (1.1), (1.2) then

$$d\epsilon^a(\mathbf{e}_b, \mathbf{e}_c) = -\epsilon^a(C^d{}_{bc}\mathbf{e}_d) = -C^a{}_{bc} = -\tfrac{1}{2}C^a{}_{uv}\epsilon^u \wedge \epsilon^v(\mathbf{e}_b, \mathbf{e}_c);$$

that is,

$$d\epsilon^a = -\tfrac{1}{2}C^a{}_{uv}\epsilon^u \wedge \epsilon^v, \tag{1.4}$$

the equations of Maurer–Cartan.

Let g be the metric on G; i.e., it is a smooth symmetric bilinear map

$$g : V \times V \to F.$$

As usual, $g(\mathbf{X}, \mathbf{Y})_\sigma$ denotes the function $g(\mathbf{X}, \mathbf{Y})$ evaluated at σ, or the scalar product of the tangent vectors $\mathbf{X}, \mathbf{Y} \in G_\sigma$. The condition for g to be invariant under the group G is that

$$g(\mathbf{X}, \mathbf{Y})_\rho = g(dL_\sigma \mathbf{X}, dL_\sigma \mathbf{Y})_{\sigma\rho}$$

for all $\mathbf{X}, \mathbf{Y} \in V$, $\sigma \in G$. This equation is equivalent to

$$g(\mathbf{X}, \mathbf{Y})_\rho = g(\mathbf{X}, \mathbf{Y})_{\sigma\rho}, \quad \mathbf{X}, \mathbf{Y}, \in \Phi, \tag{1.5}$$

since $\mathbf{X}$, $\mathbf{Y}$ are left invariant vector fields. Defining g_{ab} as the component of g with respect to the basis $\{\mathbf{e}_a\}$,

$$g_{ab} = g(\mathbf{e}_a, \mathbf{e}_b),$$

Eq. (1.5) reduces to $(g_{ab})_\rho = (g_{ab})_{\sigma\rho}$; that is, $dg_{ab} = 0$.

If Γ is the Levi-Civita connection and

$$\omega^a{}_b = \Gamma^a{}_{bu}\, dX^u = \Gamma^a{}_{bc}\epsilon^c,$$

then the covariant differential ∇g is given by

$$\nabla g_{ab} = dg_{ab} - 2\omega^c{}_{(a}g_{b)c} = 0$$

and so

$$\omega_{ab} = g_{ac}\omega^c{}_b = -\omega_{ba}.$$

The Cartan structural equations can be written as

$$d\epsilon^a + \omega^a{}_b \wedge \epsilon^b = 0, \tag{1.6}$$

$$d\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b = -\tfrac{1}{2}R^a{}_{bcd}\epsilon^c \wedge \epsilon^d, \tag{1.7}$$

where $R^a{}_{bcd}$ are the anholonomic components of the Riemann tensor. From Eqs. (1.4) and (1.6)

$$\Gamma^a{}_{[bc]} = -\tfrac{1}{2}C^a{}_{bc}.$$

Using the symmetries of C and Γ, this can be solved for Γ

$$2\Gamma_{abc} = -C_{abc} + C_{bac} + C_{cab}. \tag{1.8}$$

As an immediate consequence, the Γ_{abc} are constants and so

$$d\omega^a{}_b = \Gamma^a{}_{bc}\, d\epsilon^c = -\tfrac{1}{2}\Gamma^a{}_{bc}C^c{}_{uv}\epsilon^u \wedge \epsilon^v.$$

From this and Eq. (1.7)

$$R^a{}_{bcd} = \Gamma^a{}_{bm}C^m{}_{cd} - \Gamma^a{}_{mc}\Gamma^m{}_{bd} + \Gamma^a{}_{md}\Gamma^m{}_{bc}. \tag{1.9}$$

Equation (1.4) cannot be solved for the forms ϵ^a unless the integrability conditions are satisfied,

$$d^2\epsilon^a = -C^a{}_{bc}\, d\epsilon^b \wedge \epsilon^c = 0;$$

that is,

$$C^m{}_{[bc}C^a{}_{d]m} = 0. \tag{1.10}$$

These are the Jacobi equations and are necessary and sufficient conditions for there to exist a group G with the $C^a{}_{bc}$ as structural constants. They are also the cyclic identitites on the Riemann tensor.

$$R^a{}_{[bcd]} \equiv 0.$$

From Eq. (1.9) the Ricci tensor is

$$R_{ab} \equiv -R^m{}_{abm} = \Gamma^u{}_{av}\Gamma^v{}_{bu} - \Gamma^u{}_{(ab)}K_u,$$

where, following Ozsvath, we define

$$K_a = \Gamma^u{}_{au} = -C^u{}_{au}.$$

In this paper we restrict our attention to those spaces which satisfy the Einstein field equations for a dust-filled universe with cosmological constant

$$R_{ab} = -\Lambda g_{ab} - u_a u_b,$$

$$\Lambda = -\tfrac{1}{2}R^a{}_a + \lambda, \tag{1.11}$$

where $\mathbf{u}$ is the momentum vector for the dust and the density ρ is given by

$$\rho = -u_a u^a > 0.$$

It should be noted that $\mathbf{u}$ is not necessarily a unit vector. It is, of course, timelike. We assume throughout this paper that it is nonzero; that is $\rho > 0$.

II. JACOBI IDENTITIES

If Eq. (1.10) is contracted with respect to a and d, we obtain

$$K_a C^a{}_{bc} = 0.$$

The problem of finding all possible homogeneous universes with $\mathbf{K}$ nonzero has been solved completely by Ozsvath[1], so we restrict our attention here to the case where $\mathbf{K}$ is zero.

Assumption:

$$\mathbf{K} = 0. \tag{2.1}$$

Since G is four-dimensional, the Jacobi identities, Eq. (1.10), are equivalent to

$$\epsilon^{abcp} C^d{}_{ab} C^q{}_{cd} = 0,$$

where ϵ^{abcp} is the Levi–Civita tensor density. Also, since antisymmetrization over more than four indicies gives an identically zero expression, we have

$$\epsilon^{[abcd} C^{p]}{}_{ab} C^q{}_{cd} \equiv 0.$$

This can be expanded out to give,

$$\tfrac{1}{2}\epsilon^{abcd} C^p{}_{ab} C^q{}_{cd} \equiv +\epsilon^{abcp} C^d{}_{ab} C^q{}_{cd} - \epsilon^{bcdp} K_b C^q{}_{cd}$$

and so the Jacobi identities are equivalent to

$$\epsilon^{abcd} C^p{}_{ab} C^q{}_{cd} = 0 \tag{2.2}$$

subject to assumption (2.1). If we define the two forms C^p by

$$C^p = C^p{}_{ab} \epsilon^a \wedge \epsilon^b,$$

then Eq. (2.2) can be written as

$$C^p \wedge C^q = 0. \tag{2.3}$$

We now prove the following Lemma.

Lemma 1: Suppose that $\{\theta^A : A = 1, \cdots, N\}$ is a set of N two-forms (in four dimensions), satisfying

$$\theta^A \wedge \theta^B = 0. \tag{2.4}$$

Then one of the following dual conditions must be satisfied:

(i) There exists $(N + 1)$ one-forms, $\{P, \Phi^A\}$, such that

$$\theta^A = \Phi^A \wedge P, \text{ i.e., } \theta^A{}_{bc} = 2\Phi^A{}_{[b} P_{c]}$$

(ii) There exists a vector $\mathbf{l}$ such that

$$\mathbf{l} \perp \theta^A = 0, \text{ i.e., } l^a \theta^A{}_{ab} = 0.$$

When $A = B$, Eq. (2.4) shows that each θ^A is a simple bivector,

$$\theta^A = \alpha \wedge \beta \quad \text{(say)},$$

and therefore each nonzero θ^A defines a unique two plane. When $A \neq B$, Eq. (2.4) shows that the differential forms θ^A and θ^B have a common nontrivial factor,

$$\theta^A = \alpha \wedge \beta, \qquad \theta^B = \alpha \wedge \gamma.$$

Let us suppose that condition (i) of the Lemma is not satisfied, so that there must exist at least three linearly independent bivectors, θ^1, θ^2, θ^3 (say) which do not have a common factor. Otherwise, the θ^A would be linearly dependent on θ^1, θ^2, θ^3 (say) which have a common factor p, and all the θ^A would contain this factor. Let ω^1, ω^2, and ω^3 be the common factors for $(\theta^2, \theta^3)(\theta^3, \theta^1)$ and (θ^1, θ^2), respectively. The ω^i must be distinct, since otherwise the θ^i would contain a common factor. It is trivial that by a suitable normalization of the ω^i we have

$$\theta^1 = \omega^2 \wedge \omega^3, \qquad \theta^2 = \omega^3 \wedge \omega^1, \qquad \theta^3 = \pm\omega^1 \wedge \omega^2.$$

Furthermore, the ω^i are linearly independent since if $\omega^3 = a\omega^1 + b\omega^2$, then the θ^i are all proportional to $\omega^1 \wedge \omega^2$. Now, let $\mathbf{l}$ be a unique vector for which

$$\omega^i(\mathbf{l}) = 0 = \omega^i{}_a l^a \qquad (i = 1, 2, 3)$$

and so

$$\mathbf{l} \perp \theta^i = 0 \qquad (i = 1, 2, 3).$$

We use the $\omega^i (i = 1, 2, 3)$ and an independent one-form ω^4 as a basis, so that any θ^A can be written as

$$\theta^A = \theta^A{}_{bc} \omega^b \wedge \omega^c = 0.$$

The equations $\theta^A \wedge \theta^i = 0$ now become

$$\theta^A{}_{bc} \omega^b \wedge \omega^c \wedge \omega^i \wedge \omega^k = 0;$$

that is,

$$\theta^A{}_{4i} = 0 \quad \text{or} \quad \mathbf{l} \perp \theta^A = 0$$

since $l^a = l^4 \delta^a{}_4$, and so the θ^A satisfy the second possibility of Lemma I.

Comparing Lemma I with Eq. (2.3), we see that if $\mathbf{K}$ is zero then the structure constants must be one of the following types:

$$\begin{aligned} &\text{Type A:} \quad C^a{}_{bc} = \theta^a{}_{[b} p_{c]}, \\ &\text{Type B:} \quad C^a{}_{bc} l^c = 0. \end{aligned} \tag{2.5}$$

These types are considered separately in the remainder of this paper. In particular, we prove in the next section that Type A is incompatible with the field equations.

III. TYPE-A HOMOGENEOUS SOLUTIONS

For these spaces, assumption (2.1) becomes

$$p_m\theta^m{}_a = p_a\theta^u{}_u. \tag{3.1}$$

The field equations (1.11) are written

$$\Gamma^u{}_{va}\Gamma^v{}_{ub} = -\Lambda g_{ab} - u_a u_b. \tag{3.2}$$

On contracting Eq. (3.2) with p^b and using Eq. (1.8), (2.5), and (3.1), we obtain

$$Q\mathbf{p} + g(\mathbf{u}, \mathbf{p})\mathbf{u} = 0,$$

where Q is some function of the $\theta^a{}_b$ and p_a. From this, we obtain

Lemma 2: For Type-A spaces either (i) $\mathbf{p}$ is timelike and parallel to $\mathbf{u}$, or (ii) $\mathbf{p}$ is perpendicular to $\mathbf{u}$ and therefore spacelike.

Since $\mathbf{p}$ can never be null, we replace $\theta^a{}_b$ in Eq. (2.5) by

$$\theta^a{}_b - (\theta^a{}_c p^c/p_d p^d)p_b.$$

This does not change the structure constants, but does give

$$\theta^a{}_b p^b = 0. \tag{3.3}$$

Multiplying Eq. (3.1) by p^a gives

$$\theta^u{}_u = 0, \qquad p_u\theta^u{}_a = 0 \tag{3.4}$$

so that the tensor θ_{ab} is completely orthogonal to $\mathbf{p}$, and is traceless. The field equations now become

$$(S_{rs}S^{rs})p_a p_b - p^2(A_{ar}S^r{}_b + A_{br}S^r{}_a) + \Lambda g_{ab} + u_a u_b = 0, \tag{3.5}$$

where A and S are the antisymmetric and symmetric parts of θ,

$$A_{ab} = \theta_{[ab]}, \qquad S_{ab} = \theta_{(ab)},$$

and are clearly both completely orthogonal to $\mathbf{p}$.

Let us first suppose that $\mathbf{u}$ and $\mathbf{p}$ are parallel, so that they can be taken equal ($\mathbf{u}$ being nonzero). Then, since $A_{a\cdot b}p^b = 0$ and A is antisymmetric, it must be of rank two at most; that is, there exists a spacelike vector $\mathbf{c}$ which satisfies

$$A_{ab}c^b = 0, \quad c^2 = +1, \quad c^a p_a = c^a u_a = 0.$$

On contracting with $c^a c^b$, Eq. (3.5) gives $\Lambda = 0$, and contracting with $u^a u^b$ then gives

$$(1 + S_{rs}S^{rs})\rho^2 = 0. \tag{3.6}$$

If we choose an orthogonal base with $\rho^{\frac{1}{2}}\mathbf{e}_0 = \mathbf{u}$, and

$$g(\mathbf{X}, \mathbf{Y}) = -X_0 Y_0 + \sum_{i=1}^{3} X_i Y_i,$$

then

$$S_{ab}S^{ab} = \sum_i S_{ii}{}^2 + 2\sum_{i<j} S_{ij}{}^2 \geq 0$$

and so, from Eq. (3.6) the density ρ is zero which contradicts our original assumption.

The other alternative was for $\mathbf{u}$ to be perpendicular to $\mathbf{p}$. We choose a coordinate system in which

$$u^a = u^0\delta^a{}_0, \qquad p^a = \delta^a{}_3,$$

$$g_{ab} = \text{diag}\,(-1\ +1\ +1\ +1), \qquad S_{12} = 0.$$

The last condition can be accomplished by a rotation in the $(\mathbf{e}_1, \mathbf{e}_2)$ plane. If Eq. (3.5) is multiplied by S^{ab} and Eq. (3.3, 4) used, it follows that $S_{00} = 0$. Collecting together the restrictions on A and S,

$$S_{3a} = A_{3a} = S_{00} = S_{12} = 0,$$

$$S_{11} = -S_{22} = S \quad \text{(say)}.$$

The four equations corresponding to R_{01}, R_{02}, R_{12}, and $R_{11} - R_{22}$ give

$$E_1 \equiv A_{10}S - A_{12}S_{20} = 0,$$

$$E_2 \equiv -A_{20}S + A_{12}S_{10} = 0,$$

$$E_3 \equiv 2A_{12}S + A_{20}S_{10} + A_{10}S_{20} = 0,$$

$$E_4 \equiv A_{10}S_{10} - A_{20}S_{20} = 0,$$

since $p^2 \neq 0$. It is trivial that

$$2(E_1{}^2 + E_2{}^2) + E_3{}^2 + E_4{}^2 \equiv [2S^2 + (S_{10})^2 + (S_{20})^2] \times [2(A_{12})^2 + (A_{10})^2 + (A_{20})^2]$$

and so either A_{ab} or S_{ab} is zero. The field equations now become

$$-\Lambda g_{ab} = u_a u_b + (S_{rs}S^{rs})p_a p_b,$$

which immediately gives that $\Lambda = 0$ since the right-hand side is of rank ≤ 2, and therefore, $\mathbf{u}$ is parallel to $\mathbf{p}$, which is a contradiction. This completes the proof of the following.

Lemma 3: There are no dust-filled homogeneous universes with Type-A structure contants.

IV. TYPE-B HOMOGENEOUS SOLUTIONS

The Type-B structure constants can be written as follows:

$$C^a{}_{bc} = \sqrt{2}\,\epsilon_{bcu}C^{au}l^v$$

where the Levi-Civita density is chosen so that

$$\epsilon_{0123} = +1.$$

Since $\mathbf{K}$ is zero, C^{au} must satisfy

$$\epsilon_{bcus}C^{cu}l^s = 0,$$

and so

$$C^{[cu]} = m^{[c}l^{u]},$$

where $\mathbf{m}$ is some vector. If the symmetric tensor S^{ab} is defined by

$$S^{ab} = C^{ab} - m^a l^b,$$

then the structure constants can be rewritten as

$$C^a{}_{bc} = \sqrt{2}\,\epsilon_{bcus}S^{au}l^s, \qquad S^{ab} = S^{ba}. \tag{4.1}$$

A straightforward computation of R_{ab} gives

$$\begin{aligned}2S_{ab}S_{**} - 3S_{a*}S_{b*} &+ l^2[4S_{am}S_b{}^m - 2S_{ab}S^u{}_u] + l_aT_b + l_bT_a \\ &+ (l_al_b)[2S_{rs}S^{rs} - (S^u{}_u)^2] + \bar{\Lambda}g_{ab} + u_au_b = 0,\end{aligned}$$

where an asterisk denotes contraction with the vector $\mathbf{l}$

$$S_{b*} = S_{bu}l^u,$$

and $\bar{\Lambda}$ and T_a are defined by

$$\bar{\Lambda} = \Lambda + [(S^d{}_d)^2l^2 + 3S^f{}_{*}S_{f*} - 2S^{us}S_{us}l^2 - 2S^u{}_uS_{**}],$$

$$T_a = 2S^u{}_uS_{a*} - 3S_{u*}S^u{}_a.$$

We consider the three separate cases where $\mathbf{l}$ is null, timelike, or spacelike.

Lemma 4: Any spaces of Type B with $\mathbf{l}$ null must be either the Einstein static universe or the Gödel cosmos. This will follow from Gödel's theorem[2] (proved in Ozsvath[1]) provided we can show that the velocity vector $\mathbf{u}$ is geodesic, expansion free, and shear free, i.e.,

$$u^a{}_{;b}u^b \propto u_a,\ u^a{}_{;a} = 0,\ u_{(a;b)} = 0.$$

The first of these expressions follows from the field equations, for any dust-filled universe, whether homogeneous or not, and the second from $\mathbf{K} = 0$. In order to complete the proof of Lemma 4 we must prove that

$$u_{(a;b)} = -\Gamma^c{}_{(ab)}u_c = +C_{(ab)c}u^c = 0.$$

A coordinate system is chosen in which

$$\mathbf{l} = \mathbf{e}_4 \qquad \mathbf{u} = u^3\mathbf{e}_3 + u^4\mathbf{e}_4$$

and the metric tensor is given by

$$g(\mathbf{X}, \mathbf{X}) = (X^1)^2 + (X^2)^2 + 2X^3X^4$$

so that $\mathbf{e}_3$ and $\mathbf{e}_4$ are null vectors. Since $\rho = -2u_3u_4$, neither u^3 nor u^4 can be zero. If the indices A, B range over 1, 2, we obtain from the R_{44} field equation (remembering $l^2 = 0$)

$$u_4{}^2 = S_{44}{}^2 = S_{**}{}^2 > 0,$$

from the R_{4A} equations $S_{4A} = 0$, from the R_{3A} equations $S_{3A} = 0$, and from the R_{AB} equations $S_{AB} = S_{11}\delta_{AB}$. From Eq. (4.1) and the above, the shear may be written as

$$u_{(a;b)} = \sqrt{2}\,S^A{}_{(a}\epsilon_{b)A34}u^3 = 0.$$

The remaining two cases, where $\mathbf{l}$ is timelike or $\mathbf{l}$ is spacelike, are treated separately. In both, since $l^2 \neq 0$, a multiple of l_al_b can be added to S_{ab}, without affecting Eq. (4.1), giving

$$S_{**} = 0.$$

$\mathbf{l}$ *timelike:* Without loss of generality $\mathbf{l}$ can be taken as a unit timelike vector $l^2 = -1$. We use a basis so that

$$\mathbf{e}_0 = \mathbf{l}.$$

In the following the indices *ijklm* range over 1, 2, 3. Since S_{ij} is a symmetric 3×3 matrix and g_{ij} is positive definite, the basis can be chosen so that

$$S_{ij} = \text{diag}\,(\lambda_1\lambda_2\lambda_3),$$

$$g_{ab} = \text{diag}\,(-1, +1, +1, +1),$$

$$S_{i*} = S_{i0} = A_i \quad \text{(say)}.$$

Since it is known that Type-A spaces cannot exist, it follows that the $C^a{}_{bc}\epsilon^b \wedge \epsilon^c$ cannot have a common factor. This easily gives the condition that

Condition: (i) no two eigenvalues $\{\lambda_i\}$ are zero; (ii) if $A_i = 0$, $\lambda_i \neq 0$.

The field equations become

$$-4S_{im}S^m{}_j + 2S_{ij}S^m{}_m + A_iA_j + u_iu_j + \bar{\Lambda}g_{ij} = 0, \tag{4.2a}$$

$$S_{im}A^m = u_0u_i, \tag{4.2b}$$

$$2S_{ij}S^{ij} - (S^i{}_i)^2 - 2A_iA^i + u_0{}^2 - \bar{\Lambda} = 0. \tag{4.2c}$$

The first set gives

$$A_iA_j = -u_iu_j \qquad (i \neq j).$$

When these are multiplied together, we have

$$(A_1A_2A_3)^2 + (u_1u_2u_3)^2 = 0$$

[2] Gödel's theorem: The only dust-filled homogeneous universes which are expansion free and shear free are those of Einstein and Gödel.

and so one of the A_i is zero, say $A_1 = 0$. Then Eq. (4.2b) with $i = 1$ gives $u_1 = 0$, since $u_0 \neq 0$,

$$u_1 = A_1 = 0.$$

The remaining off-diagonal equations are

$$A_2A_3 = -u_2u_3, \quad \lambda_2A_2 = u_0u_2, \quad \lambda_3A_3 = u_0u_3 \tag{4.3}$$

from which is deduced that

$$[u_0^2 + \lambda_2\lambda_3]u_2u_3 = 0. \tag{4.4}$$

For the present, we assume that the term in brackets is zero, so that Eq. (4.3) yields

$$(u_0)^2 = -\lambda_2\lambda_3,$$
$$(u_2)^2 = -(\lambda_2/\lambda_3)A_2^2,$$
$$(u_3)^2 = -(\lambda_2/\lambda_3)A_3^2,$$

where, by an interchange of $\mathbf{e}_2$ and $\mathbf{e}_3$ if necessary,

$$\lambda_2 > 0 > \lambda_3. \tag{4.5}$$

The remaining field equations give

$$A_2^2(1 - \lambda_2/\lambda_3) = 2(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_1 - \lambda_2) \geq 0,$$
$$A_3^2(1 - \lambda_3/\lambda_2)$$
$$= 2(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3 - \lambda_1) \geq 0, \tag{4.6}$$

$$-u_mu^m = 2\lambda_1^2 - 2\lambda_2^2 - 2\lambda_3^2 - \lambda_2\lambda_3 > 0. \tag{4.7}$$

From which the A_i and u_a can be calculated as functions of the λ_i, provided that the inequalities in Eqs. (4.5)–(4.6), and (4.7) are satisfied, as well as the following equation on the λ_i

$$3\lambda_1^2 + \lambda_2^2 + \lambda_3^2$$
$$- 4\lambda_1(\lambda_2 + \lambda_3) + 5\lambda_2\lambda_3 = 0. \tag{4.8}$$

Since the inequalities, and Eq. (4.8) are all homogeneous, they may be plotted in two dimensions and may be shown to be inconsistent. This proves that Eq. (4.4) has the root $u_2u_3 = 0$. From Eq. (4.3) it will be seen that $A_2A_3 = 0$, and so $A_2 = 0$ (say). From Eq. (4.3) $u_2 = 0$, so that

$$u_1 = u_2 = A_1 = A_2 = 0.$$

$\bar{\Lambda}$ may now be eliminated from the field equations R_{11} and R_{22} giving

$$(\lambda_1 - \lambda_2)(\lambda_1 + \lambda_2 - \lambda_3) = 0. \tag{4.9}$$

The condition for $\mathbf{u}$ to be shear free and therefore for G to be Gödel or Einstein is easily calculated to be

$$u_{[a}A_{b]} = 0,$$
$$u_i(\lambda_j - \lambda_k) = 0, \qquad \{ijk\} = \{1, 2, 3\},$$

which reduces here to $u_3(\lambda_1 - \lambda_2) = 0$. Consequently, we do not consider the root $\lambda_1 = \lambda_2$ in Eq. (4.9) but only

$$\lambda_1 + \lambda_2 = \lambda_3, \qquad \lambda_1 \neq \lambda_2. \tag{4.10a}$$

The field equations reduce to

$$u_0^2 = 2A_3^2,$$
$$u_3^2 = -A_3^2 + 4\lambda_1\lambda_2, \tag{4.10b}$$
$$u_0u_3 = A_3(\lambda_1 + \lambda_2).$$

Eliminating u_a from these equations, a quadratic equation is obtained. One root is $A_3 = 0$, which is inconsistent with $u_0^2 > 0$. The other gives

$$A_3^2 = \tfrac{1}{2}(6\lambda_1\lambda_2 - \lambda_1^2 - \lambda_2^2). \tag{4.10c}$$

The conditions that A_3^2, ρ^2, u_0^2, and u_3^2 be positive (or zero) give

$$1 < \lambda_1/\lambda_2 < 3, \tag{4.10d}$$

where we have chosen the axes so that $\lambda_1 > \lambda_2$.

1 *spacelike:* We assume in the following that $l^2 = +1$. The coordinate system is chosen so that

$$\mathbf{l} = \mathbf{e}_3, \qquad g_{a3} = g(\mathbf{e}_a, \mathbf{e}_3) = \delta^3{}_a.$$

We let the indices $(ij \cdots n)$ range over (0 1 2) here. This gives a natural decomposition of S_{ab} into a 3×3 symmetric matrix S_{ij} and a 3-vector $A_i = S_{i3}$ (remember $S_{33} = S_{**} = 0$). The useful field equations reduce to

$$R_{ij} : 4S_{im}S^m{}_j - 2S_{ij}S^m{}_m$$
$$+ A_iA_j + u_iu_j + \bar{\Lambda}g_{ij} = 0, \tag{4.11a}$$

$$R_{i3} : S_{im}A^m = -u_3u_i. \tag{4.11b}$$

Lemma 5: If S_{ij} is a symmetric matrix satisfying Eq. (4.11), then S_{ij} and g_{ij} can be simultaneously diagonalized.

$$S_{ij} = \text{diag}\,(\lambda_0\lambda_1\lambda_2),$$
$$g_{ij} = \text{diag}\,(-1 + 1 + 1).$$

Proof: Consider the eigenvalue problem

$$S_{ij}X^i = \lambda g_{ij}X^i. \tag{4.12}$$

If λ is complex, then

$$S_{ij}\bar{X}^i = \bar{\lambda}\bar{X}^i, \qquad X_i\bar{X}^i = 0$$

and so contracting Eq. (4.11a) with $\bar{X}^iX^j$ gives

$$|A_iX^i|^2 + |u_iX^i|^2 = 0$$

and so $\mathbf{u}$ is perpendicular to $X^i(=\alpha^i + i\beta^i)$, so that both α^i, β^i are spacelike. This is impossible,

since $\mathbf{X}$ is null and therefore $\alpha^2 + \beta^2 = 0$, and so S cannot have a complex eigenvalue.

Now suppose that the eigenvector $\mathbf{X}$ is real and null. Contracting Eq. (4.11a) with X^iX^j gives, as before,

$$(A_iX^i)^2 + (u_iX^i)^2 = 0.$$

As before, this implies that $\mathbf{u} \perp \mathbf{X}$ which is impossible for $\mathbf{u}$ timelike, $\mathbf{X}$ null. Hence all eigenvectors are real and nonnull.

Finally, if λ is a repeated root for Eq. (4.12) then either there are three independent eigenvalues, or else there exists a vector $\mathbf{Y}$ such that[3]

$$S_i{}^jY_j = \lambda Y_i + X_i, \qquad S_i{}^jX_j = \lambda X_i.$$

Multiplying these by X^iY^i, respectively, subtracting one from the other, and using the symmetries of S_{ij} and g_{ij} gives $X_iX^i = 0$, and so $\mathbf{X}$ is a null eigenvector, which is inconsistant with the field equations. This proves that there are three mutually orthogonal nonnull eigenvectors, which are used as a basis, giving Lemma 5.

It is clear that the field equations are now similar to the previously treated case where $\mathbf{1}$ is timelike. The analysis is almost exactly the same, so we only give the results. Apart from a shear free solution, there are only two possible solutions

$$\begin{aligned}
&\text{(i)} \quad u_1 = u_2 = A_1 = A_2 = 0,\\
&A_0{}^2 = \tfrac{1}{2}(6\lambda_1\lambda_2 - \lambda_1{}^2 - \lambda_2{}^2),\\
&\lambda_0 = -\lambda_1 - \lambda_2, \quad \rho = \tfrac{3}{2}(\lambda_1{}^2 + \lambda_2{}^2) - 5\lambda_1\lambda_2,\\
&3\lambda_1 < \lambda_2 < (3 + 2\sqrt{2})\lambda_1,
\end{aligned} \tag{4.13}$$

$$\begin{aligned}
&\text{(ii)} \quad \mathbf{u} = u^0\mathbf{e}_0, \qquad A_i = 0,\\
&\rho = u_0{}^2 = 4\lambda_1\lambda_2, \qquad \lambda_0 = -(\lambda_1 + \lambda_2).
\end{aligned} \tag{4.14}$$

Theorem: The cosmological spaces given by Ozsvath[1] are a complete set of possible homogeneous dust-filled universes.

Apart from the Einstein and Gödel universes, Ozsvath gives three distinct classes of solutions, each of which depend on two constants. This may be seen to agree numerically with our results. The proof of this theorem consists of transforming the Ozsvath solutions into the canonical forms of this paper, and then comparing with our results. We do not do so here since the analysis is trivial, but messy.

V. METRICS

We reduce all the cosmological solutions to the following type:

$$(ds)^2 = -a_0\,\omega_0{}^2 + a_1\,\omega_1{}^2 + a_2\,\omega_2{}^2 + a_3\,\omega_3{}^2,$$

where the ω_a are left invariant vector fields with $a_b^{\frac{1}{2}}\omega_b = \mathbf{e}_b$ (not summed) and the a_b are chosen so that the structure constants for the ω_a are ± 2 or 0.

$\mathbf{1}$ *timelike:*

When the cosmos is that of Eq. (4.10), the a_b are given by

$$\lambda_2\lambda_3a_1 = \lambda_3\lambda_1a_2 = \lambda_1\lambda_2a_3 = 2,$$
$$a_0 = \tfrac{1}{2}a_1a_2A_3{}^2,$$

so that

$$d\omega_0 = -\omega_1 \wedge \omega_2, \qquad d\omega_1 = \omega_2 \wedge \omega_3,$$
$$d\omega_2 = \omega_3 \wedge \omega_1, \qquad d\omega_3 = \omega_1 \wedge \omega_2.$$

These equations give

$$\omega_0 = dt - \omega_3,\ \omega_1 = -\sin z\,dx + \sin x\cos z\,dy,$$
$$\omega_2 = \cos z\,dx + \sin x\sin z\,dy,\ \omega_3 = \cos x\,dx + dz.$$

This is Ozsvath Type I.[4]

$\mathbf{1}$ *spacelike:*

For the cosmos of Eq. (4.13) or Eq. (4.14), the a_b are given by

$$\lambda_1\lambda_2a_0 = -\lambda_0\lambda_2a_1 = -\lambda_0\lambda_1a_2 = 2,\ a_3 = \tfrac{1}{2}a_1a_2A_0{}^2,$$

so that

$$d\omega_0 = \omega_1 \wedge \omega_2, \qquad d\omega_1 = -\omega_2 \wedge \omega_0,$$
$$d\omega_2 = -\omega_0 \wedge \omega_1, \qquad d\omega_3 = \epsilon\omega_1 \wedge \omega_2,$$

where $\epsilon = 1$ for Ozsvath type II[4] and is zero for Ozsvath III and Gödel. These equations give

$$\omega_0 = dt + e^y\,dx, \qquad \omega_1 + i\omega_2 = e^{-it}[dy + ie^y\,dx],$$
$$\omega_3 = dz + \epsilon\omega_0.$$

The field equation and inequalities of Eqs. (4.10), (4.13), and (4.14) give

$$\begin{aligned}
&\text{(I)} \quad a_3 = a_1 + a_2,\\
&2a_0a_3 = 6a_1a_2 - a_1{}^2 - a_2{}^2,\\
&a_1 < a_2 < 3a_1,\\
&\rho = (4a_1a_2a_3)^{-1}(10a_1a_2 - 3a_1{}^2 - 3a_2{}^2),\\
&\text{(E)} \quad a_b = a,\\
&\text{(II)} \quad a_0 = a_1 + a_2,\\
&2a_3a_0 = 6a_1a_2 - a_1{}^2 - a_2{}^2,\\
&3a_2 < a_1 < (3 + \sqrt{2})a_2,\\
&\rho = (4a_0a_1a_2)^{-1}(10a_1a_2 - 3a_1{}^2 - 3a_2{}^2),\\
&\text{(III)} \quad a_3 = a_0 = a_1 + a_2,\\
&a_1 \neq a_2,\\
&\rho = 2a_0{}^{-1},
\end{aligned}$$

[3] This follows from the standard analysis of $S_i{}^jX_j = \lambda X_i$.

[4] We shall refer to the Ozsvath solution of Eqs. (4.10), (4.13), and (4.14) as (I), (II), and (III), respectively. Also Einstein and Gödel are referred to as (E) and (G).

(G) $a_3 = a_0 = 2a_1 = 2a_2 = a,$

$\rho = a^{-1}.$

The maximal groups of isometries, Φ, of these metrics are as follows: For Ozsvath Type I Φ is 4 parametric, and is generated by the following Killing vectors

(I) $\mathbf{K}_4 = \partial t, \quad \mathbf{K}_3 = \partial z,$

$\mathbf{K}_1 + i\mathbf{K}_2 = e^{-iz}[\cot y\, \partial z + \csc y\, \partial x + i\, \partial y].$

The corresponding Lie brackets are $[\mathbf{K}_4, \mathbf{K}_i] = 0$ and $[\mathbf{K}_i, \mathbf{K}_j] = \epsilon_{ijk}\mathbf{K}_k$, $i, j, k \neq 4$. When M is the Einstein cosmos, Φ is seven-dimensional and the Killing vectors are the above $\mathbf{K}_a$ and in addition

$$\mathbf{L}_3 = \partial x,$$

$$\mathbf{L}_1 + i\mathbf{L}_2 = e^{ix}[\cot y\, \partial x + \csc y\, \partial z + i\, \partial y].$$

The additional Lie brackets are

$$[\mathbf{L}_i, \mathbf{L}_j] = \epsilon_{ijk}\mathbf{L}_k, \quad [\mathbf{L}_i, \mathbf{K}_a] = 0.$$

For the group of spaces with 1 spacelike, the isometry group Φ is generated by (II, III)

$$\mathbf{M}_1 = \partial z, \quad \mathbf{M}_2 = \partial x,$$

$$\mathbf{M}_3 = \partial y - x\, \partial x,$$

$$\mathbf{M}_4 = e^{-y}\, \partial t - x\, \partial y + \tfrac{1}{2}(x^2 - e^{-2y})\, \partial x.$$

Gödel has one more Killing vector,

(G) $\mathbf{M}_5 = \partial t.$

The Lie brackets are

$$[\mathbf{M}_1, \mathbf{M}_a] = 0, \quad [\mathbf{M}_2, \mathbf{M}_3] = -\mathbf{M}_2,$$

$$[\mathbf{M}_4, \mathbf{M}_2] = \mathbf{M}_3, \quad [\mathbf{M}_4, \mathbf{M}_3] = \mathbf{M}_4,$$

$$[\mathbf{M}_5, \mathbf{M}_a] = 0.$$

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 7, NUMBER 9 SEPTEMBER 1966

On Quantization in Space with Torsion*

R. FINKELSTEIN
Department of Physics, University of California, Los Angeles, California
(Received 3 January 1966)

Some elementary quantization questions are discussed under the hypothesis that physical space is a homogeneous space of constant torsion. In such a space, a freely propagating point particle appears to have extension in the sense that it obeys the Schrödinger equation of a symmetric top. Modifications of commutators and scattering amplitudes are given. The geometrical methods are based on the correspondence with the associated group-space.

INTRODUCTION

IT has often been suggested that microscopic physics depends on the astronomical neighborhood. It is at the same time generally believed that the large-scale distribution of energy–momentum is not important at the current level of microscopic physics, since only the metrical properties of local space–time are so determined. On the other hand, different features of the local environment, such as the distribution of the baryonic charge and chirality, may influence the vacuum state of elementary-particle physics in ways which are more difficult to dismiss.[1] If these possibilities should eventually prove important, then the vacuum state would be characterized by certain nonvanishing expectation values, such as the torsion of space, which would not only provide a mechanism for breaking the symmetry group but would also provide a mass scale for microscopic physics in terms of astronomically determined quantities.[2]

In order to discuss some of these possibilities, we may consider a space described by a nonsymmetric affinity. If, in addition, we postulate the existence of a gauge group, the affinity may be separated into four parts,

$$U^{\mu}_{\alpha\beta} = C^{\mu}_{(\alpha\beta)} + C^{\mu}_{[\alpha\beta]} + G^{\mu}_{\alpha\beta} + G^{\mu}_{(\alpha\beta)},$$

* This work was supported by the National Science Foundation.

[1] See for example, T. D. Lee, Phys. Rev. 137, B1621 (1965).

[2] R. Finkelstein, J. Math. Phys. 1, 440 (1960). For example, if our universe is supposed to be a closed space of uniform torsion, and the following simple relation is assumed to determine the torsion, φ^{μ}:

$$|\varphi^{\mu}| \cong r_0^{-2} \langle \bar{\psi}\gamma^5\gamma^{\mu}\psi \rangle,$$

where the expectation value on the right side is chosen for the ground state, and r_0 has the dimensions of a length, then assuming $\langle \bar{\psi}\psi \rangle$ measures the astronomically observed particle density, one finds that r_0 has possibly the order of magnitude $\sim 10^{-12}$ cm.

(G) $a_3 = a_0 = 2a_1 = 2a_2 = a,$

$\rho = a^{-1}.$

The maximal groups of isometries, Φ, of these metrics are as follows: For Ozsvath Type I Φ is 4 parametric, and is generated by the following Killing vectors

(I) $\mathbf{K}_4 = \partial t, \quad \mathbf{K}_3 = \partial z,$

$\mathbf{K}_1 + i\mathbf{K}_2 = e^{-iz}[\cot y\, \partial z + \csc y\, \partial x + i\, \partial y].$

The corresponding Lie brackets are $[\mathbf{K}_4, \mathbf{K}_i] = 0$ and $[\mathbf{K}_i, \mathbf{K}_j] = \epsilon_{ijk}\mathbf{K}_k$, $i, j, k \neq 4$. When M is the Einstein cosmos, Φ is seven-dimensional and the Killing vectors are the above $\mathbf{K}_a$ and in addition

$$\mathbf{L}_3 = \partial x,$$

$$\mathbf{L}_1 + i\mathbf{L}_2 = e^{ix}[\cot y\, \partial x + \csc y\, \partial z + i\, \partial y].$$

The additional Lie brackets are

$$[\mathbf{L}_i, \mathbf{L}_j] = \epsilon_{ijk}\mathbf{L}_k, \quad [\mathbf{L}_i, \mathbf{K}_a] = 0.$$

For the group of spaces with 1 spacelike, the isometry group Φ is generated by (II, III)

$$\mathbf{M}_1 = \partial z, \quad \mathbf{M}_2 = \partial x,$$

$$\mathbf{M}_3 = \partial y - x\, \partial x,$$

$$\mathbf{M}_4 = e^{-y}\, \partial t - x\, \partial y + \tfrac{1}{2}(x^2 - e^{-2y})\, \partial x.$$

Gödel has one more Killing vector,

(G) $\mathbf{M}_5 = \partial t.$

The Lie brackets are

$$[\mathbf{M}_1, \mathbf{M}_a] = 0, \quad [\mathbf{M}_2, \mathbf{M}_3] = -\mathbf{M}_2,$$

$$[\mathbf{M}_4, \mathbf{M}_2] = \mathbf{M}_3, \quad [\mathbf{M}_4, \mathbf{M}_3] = \mathbf{M}_4,$$

$$[\mathbf{M}_5, \mathbf{M}_a] = 0.$$

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 7, NUMBER 9 SEPTEMBER 1966

On Quantization in Space with Torsion*

R. FINKELSTEIN
Department of Physics, University of California, Los Angeles, California
(Received 3 January 1966)

Some elementary quantization questions are discussed under the hypothesis that physical space is a homogeneous space of constant torsion. In such a space, a freely propagating point particle appears to have extension in the sense that it obeys the Schrödinger equation of a symmetric top. Modifications of commutators and scattering amplitudes are given. The geometrical methods are based on the correspondence with the associated group-space.

INTRODUCTION

IT has often been suggested that microscopic physics depends on the astronomical neighborhood. It is at the same time generally believed that the large-scale distribution of energy–momentum is not important at the current level of microscopic physics, since only the metrical properties of local space–time are so determined. On the other hand, different features of the local environment, such as the distribution of the baryonic charge and chirality, may influence the vacuum state of elementary-particle physics in ways which are more difficult to dismiss.[1] If these possibilities should eventually prove important, then the vacuum state would be characterized by certain nonvanishing expectation values, such as the torsion of space, which would not only provide a mechanism for breaking the symmetry group but would also provide a mass scale for microscopic physics in terms of astronomically determined quantities.[2]

In order to discuss some of these possibilities, we may consider a space described by a nonsymmetric affinity. If, in addition, we postulate the existence of a gauge group, the affinity may be separated into four parts,

$$U^{\mu}_{\alpha\beta} = C^{\mu}_{(\alpha\beta)} + C^{\mu}_{[\alpha\beta]} + G^{\mu}_{\alpha\beta} + G^{\mu}_{(\alpha\beta)},$$

* This work was supported by the National Science Foundation.

[1] See for example, T. D. Lee, Phys. Rev. 137, B1621 (1965).

[2] R. Finkelstein, J. Math. Phys. 1, 440 (1960). For example, if our universe is supposed to be a closed space of uniform torsion, and the following simple relation is assumed to determine the torsion, φ^{μ}:

$$|\varphi^{\mu}| \cong r_0^{-2} \langle \bar{\psi}\gamma^5\gamma^{\mu}\psi \rangle,$$

where the expectation value on the right side is chosen for the ground state, and r_0 has the dimensions of a length, then assuming $\langle \bar{\psi}\psi \rangle$ measures the astronomically observed particle density, one finds that r_0 has possibly the order of magnitude $\sim 10^{-12}$ cm.

where ( ) and [ ] mean symmetric and antisymmetric, while C and G mean Hermitian and anti-Hermitian. The antisymmetric part of the affinity is known as the torsion. C represents the connection of the local Lorentz group, while G is the connection of the local gauge group. General relativity corresponds to the situation in which there is only a symmetric Hermitian part, and in which there is in fact no gauge group in the present sense. If one makes the simplest generalization of relativity and puts

$$U^{\mu}_{\alpha\beta} = (U^{\mu}_{\beta\alpha})^{+},$$

$$g_{\alpha\beta} = (g_{\beta\alpha})^{+},$$

then only the symmetric Hermitian and antisymmetric anti-Hermitian parts survive, and one obtains a simple generalization of the theory of the Maxwell field interacting with the gravitational field.[3] The anti-Hermitian torsion, arising from the existence of a gauge group and associated with the possible existence of vector mesons, is probably the most interesting part of the affinity, although with it one runs into still unsolved quantization problems.

In the present paper, we are not concerned with these gauge fields and the associated vector particles. Instead, we postulate the simplest possible space with nonvanishing torsion by making the idealizations that there is no gauge group, that the connection is real, and that the metric is real and symmetric. In addition, we do not quantize the connection but regard it as given in the same way as the flat connection of Euclidean space is related to the usual problems of quantum theory. Our aim is to discuss field quantization when the three-dimensional sections are closed spaces of uniform torsion. Since these sections are also spaces of absolute parallelism, they represent a minimal departure from the ordinary Euclidean case, and certainly can not be excluded on observational grounds.

1. SPACE OF UNIFORM TORSION

The space under consideration has three-dimensional sections which are homogeneous and isotropic and are carried into themselves by a six-parameter group of motions; in these respects, it resembles Minkowski space. We shall formally characterize it as follows[4]:

$$L^{\mu}_{\alpha\beta\gamma}(\pm) = 0, \tag{1.1}$$

$$T^{\mu}_{\alpha\beta|\gamma\pm} = 0, \tag{1.2}$$

$$g_{\alpha\beta|\gamma\pm} = 0, \tag{1.3}$$

where $L^{\mu}_{\alpha\beta\gamma}$, $T^{\mu}_{\alpha\beta}$, and $g_{\alpha\beta}$ are tensors representing the curvature, torsion, and metric. The $\pm$ sign appearing in these equations means, in every case, the covariant derivative with respect to the connection (+) and the transposed connection (−). Euclidean space is the particular solution for which $T^{\mu}_{\alpha\beta} = 0$.

From (1.1)–(1.3) it follows that $T^{\alpha\beta\gamma}$ is completely antisymmetric and therefore has the representation

$$T^{\alpha\beta\gamma} = \epsilon^{\alpha\beta\gamma\mu}\varphi_{\mu}, \tag{1.4}$$

where φ_{μ} is an axial vector which is also constant in the covariant sense. Let φ^2 be the weightless norm of φ_{μ}. Then

$$\varphi_{\alpha|\beta} = 0, \tag{1.5}$$

$$\partial_{\alpha}\varphi^2 = 0, \tag{1.6}$$

and φ_{α} is derivable from a potential:

$$\varphi_{\alpha} = g^{-\frac{1}{2}}\tilde{\varphi}_{\alpha}, \tag{1.7}$$

$$\tilde{\varphi}_{\alpha} = \partial_{\alpha}\theta, \tag{1.7a}$$

where θ is a scalar. The Ricci tensor is entirely determined by $\tilde{\varphi}_{\alpha}$ according to the relation:

$$R_{\alpha\beta} = 2(\tilde{\varphi}^2 g_{\alpha\beta} - \tilde{\varphi}_{\alpha}\tilde{\varphi}_{\beta}) \tag{1.8}$$

$$= 2(\varphi^2 g_{\alpha\beta} - g\varphi_{\alpha}\varphi_{\beta}), \tag{1.8a}$$

where $\varphi^2 = \tilde{\varphi}^2$.

The vector field φ_{α} defines a family of three-dimensional hypersurfaces which intersect φ_{α} orthogonally. Choose an orthogonal coordinate system in which x_4 lies along φ_{α} at every point, and in which the other three axes lie in the orthogonal hypersurfaces. Then

$$\varphi_k = 0, \qquad k = 1, 2, 3, \tag{1.9a}$$

$$\varphi_4 = \varphi, \tag{1.9b}$$

$$R_{kl} = 2\varphi^2 g_{kl}. \tag{1.10}$$

We shall assume that $\varphi^2 = -K$ where $K > 0$. Then hypersurfaces of constant x_4 are Einstein spaces of constant curvature and of line element

$$ds^2 = (dx^2 + dy^2 + dz^2)/(1 + \tfrac{1}{4}Kr^2)^2, \tag{1.11}$$

where the radius of curvature is $R_0 = K^{-\frac{1}{2}}$.

2. ABSOLUTE PARALLELISM

The given space is characterized by two kinds of absolute parallelism. That is, the following equations are integrable for the two tetrad fields $\lambda^{\mu}_{i}(\pm)$:

$$\partial_{\alpha}\lambda^{\mu}_{i}(\pm) + \lambda^{\sigma}_{i}(\pm)L^{\mu}_{\sigma\alpha}(\pm) = 0, \tag{2.1}$$

[3] R. Finkelstein, Rev. Mod. Phys. **36**, 632 (1964).
[4] R. Finkelstein, Ann. Phys. (N. Y.) **12**, 200 (1961).

where

$$L^{\mu}_{\alpha\beta}(+) = L^{\mu}_{\beta\alpha}(-). \tag{2.2}$$

The $\lambda^{\mu}_{i}(+)$ and $\lambda^{\mu}_{i}(-)$ tetrads may be termed (+)- and (−)-parallel, respectively.

Let vectors reciprocal to λ^{μ}_{i} be denoted by λ^{i}_{μ}. Then

$$\lambda^{\mu}_{i}\lambda^{i}_{\mu} = \delta^{i}_{i} \tag{2.3a}$$

and

$$\lambda^{\mu}_{i}\lambda^{i}_{\nu} = \delta^{\mu}_{\nu}. \tag{2.3b}$$

It is possible to describe such a space by assigning these tetrad fields.[2] That is, the connection and metric are simply calculated as follows:

$$L^{\mu}_{\alpha\beta}(\pm) = \lambda^{\mu}_{i}(\pm)\,\partial\lambda^{i}_{\alpha}(\pm)/\partial x^{\beta}, \tag{2.4}$$

$$g_{\alpha\beta} = g_{ij}\lambda^{i}_{\alpha}(\pm)\lambda^{j}_{\beta}(\pm), \tag{2.5}$$

and

$$g_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

This metric satisfies (1.3) and the tetrads are orthogonal with respect to it. The metric is symmetric, but the connection is not, and the torsion is

$$T^{\mu}_{\alpha\beta}(\pm) = \frac{1}{2}\lambda^{\mu}_{i}(\pm)\left(\frac{\partial\lambda^{i}_{\alpha}(\pm)}{\partial x^{\beta}} - \frac{\partial\lambda^{i}_{\beta}(\pm)}{\partial x^{\alpha}}\right). \tag{2.6}$$

3. GROUPS OF MOTIONS

Since a three-dimensional section is a space of constant curvature, it is a hypersphere in four dimensions and is therefore carried into itself by the six-parameter rotation group 0_4. We may factor the six-parameter group 0_4 into $0^{+}_{3} \times 0^{-}_{3}$, where 0^{+}_{3} describes right-handed screw motions which carry the $\lambda^{\mu}_{i}(+)$ field into itself, and, similarly, 0^{-}_{3} describes left-handed screw motions which carry the $\lambda^{\mu}_{i}(-)$ field into itself. It is known that the three-dimensional closed space of constant torsion is itself the group manifold of 0_3, and that the two groups of motion, 0^{+}_{3} and 0^{-}_{3}, are the left and right parameter groups of 0_3. Our method will depend essentially on the fact that the proposed three-dimensional physical space model may be identified with the group manifold of 0_3.

4. EQUATIONS OF THE ROTATION GROUP

We shall write the equations of the rotation group in the familiar form:

$$X' = UXU^{-1}, \tag{4.1}$$

where

$$X = \mathbf{\sigma x} \quad \text{and} \quad U = e^{\frac{1}{2}i\sigma w}. \tag{4.1a,b}$$

The σ are the Pauli matrices, and $\mathbf{w}$ is the rotation of amount w about the axis $\hat{w}$. Then, according to (4.1), U induces a rotation from $\mathbf{x}$ to $\mathbf{x}'$.

The parameters describing the rotation may of course be chosen differently, and we shall denote an arbitrary parameterization of the rotation by $U(a)$. Let the composition law be

$$U(a_3) = U(a_2)U(a_1). \tag{4.2}$$

Equations (4.1) and (4.2) may be expressed in differential form,

$$\partial X'/\partial a^{\alpha} = i(X', \Lambda_{\alpha}), \quad \partial U/\partial a^{\alpha} = i\Lambda_{\alpha}U, \tag{4.3a, b}$$

where

$$\Lambda_{\alpha} = -i(\partial U/\partial a^{\alpha})U^{-1} \tag{4.3c}$$

is subject to the following conditions of integrability:

$$\partial\Lambda_{\alpha}/\partial a^{\beta} - \partial\Lambda_{\beta}/\partial a^{\alpha} + i(\Lambda_{\alpha}, \Lambda_{\beta}) = 0. \tag{4.4}$$

Let us introduce the complete two-dimensional basis $\sigma_i = (1, \sigma)$ such that

$$\sigma_i\sigma_k = a^{m}_{ik}\sigma_m, \tag{4.5a}$$

$$a_{ikm} = i\epsilon_{ikm0} + d_{ikm0}, \tag{4.5b}$$

where

$$d_{ikm0} = \delta_{ik}\delta_{m0} + \delta_{km}\delta_{i0} + \delta_{mi}\delta_{k0} - 2\delta_{i0}\delta_{k0}\delta_{m0}. \tag{4.5c}$$

Then

$$\mathrm{Tr}\ \sigma_i\sigma_k = \delta_{ik}, \qquad i, k = 0, 1, 2, 3, \tag{4.5d}$$

where Tr means one half of the trace, and $\sigma^{i} = \sigma_i$. Expand in this basis:

$$X = \sum_i x^i\sigma_i, \tag{4.6}$$

$$\Lambda_{\alpha} = R_0^{-1}\sum_i \lambda^{i}_{\alpha}\sigma_i, \tag{4.7}$$

$$U = \sum_i u^i\sigma_i, \tag{4.8}$$

where R_0 is a scale factor which will prove convenient later. In describing only the rotation group, the complete basis is not needed since we take

$$x^0 = \mathrm{Tr}\ X = 0, \tag{4.6a}$$

$$\lambda^{0}_{\alpha} = R_0\,\mathrm{Tr}\ \Lambda_{\alpha} = 0, \tag{4.7a}$$

but the four-dimensional basis will be useful later. Now, (4.3a) becomes

$$\partial x^{k'}/\partial a^{\alpha} = (R_0^{-1})\lambda^{\nu}_{\alpha}\mu^{k}_{\nu}, \tag{4.9}$$

where

$$\mu_p^k = 2\epsilon_{ps}^k x^{s'}. \tag{4.9a}$$

Similarly, (4.3b) becomes

$$\partial u^k/\partial a^\alpha = (R_0^{-1})\lambda_\alpha^p \mu_p^k, \tag{4.10}$$

where now

$$\mu_p^k = 2\epsilon_{ps}^k u^s. \tag{4.10a}$$

Finally, the integrability condition becomes

$$\frac{\partial \lambda_\alpha^p}{\partial a^\beta} - \frac{\partial \lambda_\beta^p}{\partial a^\alpha} - \frac{2}{R_0}\epsilon_{st}^p \lambda_\alpha^s \lambda_\beta^t = 0. \tag{4.11}$$

These are the Maurer equations where the coefficients are the structure constants.

The connection of the group manifold is defined in such a way [see (5.5)] that the above triad fields, defined by (4.9) and (4.10), are absolutely parallel and determine the torsion according to (2.6) and (4.11) as

$$T_{\alpha\beta}^\mu = (R_0^{-1})\epsilon_{st}^p \lambda_\alpha^s \lambda_\beta^t \lambda_p^\mu \tag{4.12}$$

in terms of the structure constants.

5. THE GROUP SPACE

The group manifold is a three-dimensional continuum which may be chosen to be a sphere of radius π. Then each point in the sphere is labeled by a vector $\mathbf{w}$ whose length is the magnitude of the rotation, and whose direction is the axis of rotation. Let U_a be some element of the group. The transformation

$$U_b = UU_aU^{-1} \tag{5.1}$$

maps the group space into itself by carrying the group element a into b. This mapping preserves the trace and therefore the magnitude of the rotation. Hence, if a point in the group space is distinguished by $\mathbf{w}$ where w measures the rotation, (5.1) will take a spherical shell into itself.

On the other hand, (4.2) describes a mapping of the group space into itself which may not preserve the length of $\mathbf{w}$. For example $U(a_2)$ carries a_1 into a_3 or $U(a_1)$ carries a_2 into a_3. These mappings are elements of the left and right parameter groups and also describe motions of the group space into itself.

The group manifold is a space of uniform torsion or absolute parallelism with the right (left) parallel fields $\lambda_\mu^i(+)$, $(\lambda_\mu^i(-))$ which appear in the basic equation (4.9) or (4.10), and from which one computes the metric and connection as follows:

$$g_{\alpha\beta} = \delta_{ij}\lambda_\alpha^i \lambda_\beta^j, \tag{5.2}$$

$$L_{\alpha\beta}^\mu = \lambda_i^\mu \partial_\beta \lambda_\alpha^i, \tag{5.3}$$

or

$$g_{\alpha\beta} = R_0^2 \operatorname{Tr} \Lambda_\alpha \Lambda_\beta \tag{5.4}$$

$$= R_0^2 \operatorname{Tr} U_\alpha^+ U_\beta, \tag{5.5}$$

where

$$U_\alpha = \partial_\alpha U \tag{5.6}$$

and

$$L_{\alpha\beta}^\mu = R_0^2 \operatorname{Tr} \Lambda^\mu \partial_\beta \Lambda_\alpha. \tag{5.7}$$

It is therefore possible to characterize the geometry of the group manifold in terms of the vector matrix field Λ^α. We may also recover the scalar matrix function U as follows:

$$U = N_+\left\{\exp\left[i\int_0^x \Lambda_\mu(\pm, a)\, da^\mu\right]\right\}, \tag{5.8}$$

where $N_\pm$ means an ordering in which integrands nearer to the origin are written to right (left). Then by (5.8)

$$\partial_\mu U = i\Lambda_\mu(+)U, \tag{5.9}$$

$$\partial_\mu U = iU\Lambda_\mu(-), \tag{5.10}$$

as required, where $\Lambda_\mu(+)$ and $\Lambda_\mu(-)$ belong to left and right parameter groups, respectively. According to (5.9), however, U may depend on the path connecting the origin to x. However, it may be shown that

$$U(P') = U(P)\sum(P, P'), \tag{5.11}$$

where

$$\sum(P, P') = \exp\left[i\int_0^x R_{[\mu\nu]}\, dS^{\mu\nu}\right], \tag{5.12a}$$

$$R_{\mu\nu} = \partial_\mu \Lambda_\nu - \partial_\nu \Lambda_\mu + i(\Lambda_\mu, \Lambda_\nu), \tag{5.12b}$$

and $\sum(P, P')$ is taken over a surface bounded by a closed path passing through the origin and x. But the vanishing of $R_{[\mu\nu]}$ is just the integrability condition, and therefore $U(P)$ is the same as $U(P')$ when this condition is satisfied.

The preceding equations hold for both $\lambda_\alpha^i(+)$ and $\lambda_\alpha^i(-)$, and therefore $\Lambda_\alpha(+)$ and $\Lambda_\alpha(-)$. From (5.10)

$$\Lambda_\alpha(+) = -i(\partial_\alpha U)U^{-1}, \tag{5.13+}$$

$$\Lambda_\alpha(-) = -iU^{-1}(\partial_\alpha U). \tag{5.13-}$$

Then,

$$\begin{aligned} g_{\alpha\beta}(-) &= -R_0^2 \operatorname{Tr}(U^{-1}\partial_\alpha U)(U^{-1}\partial_\beta U) \\ &= R_0^2 \operatorname{Tr}(\partial_\alpha U^+)(\partial_\beta U), \end{aligned} \tag{5.14-}$$

$$g_{\alpha\beta}(+) = -R_0^2 \operatorname{Tr}(\partial_\alpha U)U^{-1}(\partial_\beta U)U^{-1} = R_0^2 \operatorname{Tr}(\partial_\alpha U)(\partial_\beta U^+). \tag{5.14+}$$

By construction, $g_{\alpha\beta}(\pm)$ are symmetric and therefore by (5.14)

$$g_{\alpha\beta}(+) = g_{\alpha\beta}(-). \tag{5.15}$$

The symmetric part of the connection may be expressed in terms of the metric tensor in the usual way with the aid of the Christoffel formula. The torsion may be computed as follows:

$$\begin{aligned} T^\mu_{\alpha\beta} &= \tfrac{1}{2}\lambda^\mu_i(\partial\lambda^i_\alpha/\partial x^\beta - \partial\lambda^i_\beta/\partial x^\alpha) \\ &= \tfrac{1}{2}R_0^2 \operatorname{Tr} \Lambda^\mu(\partial_\beta\Lambda_\alpha - \partial_\alpha\Lambda_\beta) \\ &= -\tfrac{1}{2}iR_0^2 \operatorname{Tr} \Lambda^\mu(\Lambda_\alpha, \Lambda_\beta). \end{aligned} \tag{5.16}$$

6. DIFFERENTIAL OPERATORS AND IRREDUCIBLE REPRESENTATIONS

We define the displacement operators

$$X_i(\pm) = i^{-1}\lambda^\mu_i(\pm)\,\partial_\mu. \tag{6.1}$$

Then

$$\begin{aligned} (X_s(\pm), X_t(\pm)) &= -2\lambda^\sigma_s\lambda^\nu_t T^\mu_{\sigma\nu}\,\partial_\mu \\ &= \mp(2i/R_0)\epsilon^m_{st}X_m(\pm). \end{aligned} \tag{6.2}$$

Furthermore

$$[X_s(\pm), X_t(\mp)] = 0. \tag{6.3}$$

Therefore the six displacement operators $X_s(\pm)$ decompose into two triplets which satisfy the commutation rules of a three-dimensional angular momentum; the six displacement operators $X_s(\pm)$ may be regarded as a choice of the generators of 0_4 which factors 0_4 into $0_3 \times 0_3$.

We also find

$$\delta^{ij}X_iX_j = -(g)^{-\frac{1}{2}}\,\partial_\mu g^{\frac{1}{2}}g^{\mu\lambda}\,\partial_\lambda \tag{6.4}$$

or

$$\delta^{ij}X_iX_j = -g^{\mu\nu}\psi_{|\mu\nu}, \tag{6.4a}$$

where $\psi_{|\mu\nu}$ is the second covariant derivative. Therefore the Casimir operators of the (+)- and (−)-groups are both

$$\sum_1^3 X_i(\pm)^2 = -\Delta, \tag{6.4b}$$

where Δ is the Laplace–Beltrami operator.

The matrix elements of the irreducible representations of SU_2 are the eigenfunctions of the following differential equations:

$$\Delta\, D^i_{mm'}(a) = -[j(j+1)/R_0^2]\, D^i_{mm'}(a), \tag{6.5}$$

$$X_3(+)\, D^i_{mm'}(a) = (m/R_0)\, D^i_{mm'}(a), \tag{6.6}$$

$$X_3(-)\, D^i_{mm'}(a) = (m'/R_0)\, D^i_{mm'}(a). \tag{6.7}$$

The two operators $X_3(+)$ and $X_3(-)$ are chosen from the left and right parameter groups. These eigenfunctions satisfy the orthogonality relations:

$$\iiint \bar{D}^i_{mn}(a)\, D^{i'}_{m'n'}(a)g^{\frac{1}{2}}\,da_1\,da_2\,da_3 = \delta^{ii'}\,\delta_{mm'}\,\delta_{nn'}\left(\frac{V}{d_i}\right), \tag{6.8}$$

where $d_i = 2j + 1$ is the dimensionality of the representation. In (6.8) $g = |g_{\mu\nu}|$ and $(g)^{\frac{1}{2}}da_1\,da_2\,da_3$ is the invariant volume element for the integration over the group space. Choosing $j = 0$, one sees that $d_i = 1$ and that

$$\iiint g^{\frac{1}{2}}\,da_1\,da_2\,da_3 = V. \tag{6.9}$$

Therefore, the constant V measures the total invariant volume of the group space.

7. GREEN'S FUNCTION AND INTEGRAL EQUATION

Define the following two-point function

$$\rho^i(a, a') = \sum_{m,n} \bar{D}^i_{mn}(a)\, D^i_{mn}(a'), \tag{7.1}$$

where a and a' are two points in the group space corresponding to rotations R and R'. Then

$$\rho^i(a, a') = \chi^i(R^{-1}R') = \sin(j + \tfrac{1}{2})w/\sin\tfrac{1}{2}w, \tag{7.2}$$

where $\chi^j(w)$ is the character of the jth irreducible representation, and where w is the magnitude of $R^{-1}R'$. Then

$$\begin{aligned} &\sum_{m,n,j} s^{2j}\,\bar{D}^i_{mn}(a)\, D^i_{mn}(a') \\ &= \sum s^{2j}[\sin(j + \tfrac{1}{2})w/\sin\tfrac{1}{2}w] = \rho(s, a, a'), \end{aligned} \tag{7.3}$$

where

$$\rho(s, a, a') = [1 - 2s\cos\tfrac{1}{2}w + s^2]^{-1}. \tag{7.3a}$$

In particular,

$$\rho(1, a, a') = (4\sin^2\tfrac{1}{4}w)^{-1}. \tag{7.3b}$$

We now find from (7.3) and (6.8)

$$D^i_{mn}(a') = (d_i/V)\int \rho(1, a, a')\, D^i_{mn}(a)g^{\frac{1}{2}}\,da. \tag{7.4}$$

The eigenvalues of this equation are $d_i = 2j + 1$. The $D^i_{mn}(a)$ may be normalized as follows:

$$u^i_{mn}(a) = (d_i/V)^{\frac{1}{2}}\, D^i_{mn}(a). \tag{7.5}$$

Then by (6.8)

$$\int (u^i_{mn}(a))^2 g^{\frac{1}{2}}\, da = 1. \tag{7.6}$$

In terms of these normalized functions, we write (7.4) as

$$u^i_{mn}(a') = d_i \int K(a', a) u^i_{mn}(a)\, d\tau_a, \tag{7.7}$$

where

$$K(a', a) = V^{-1}\rho(a', a), \tag{7.7a}$$

$$d\tau_a = g^{\frac{1}{2}}\, da. \tag{7.7b}$$

Then the kernel is

$$K(a', a) = \sum_{imn} \bar{u}^i_{mn}(a') u^i_{mn}(a)/d_i, \tag{7.8}$$

and the delta function is

$$\delta(a', a) = \sum_{imn} \bar{u}^i_{mn}(a') u^i_{mn}(a) \tag{7.9}$$

$$= (V)^{-1} \sum_{imn} d_i\, \bar{D}^i_{mn}(a')\, D^i_{mn}(a). \tag{7.9a}$$

We may evaluate (7.9a) with the aid of the series (7.3) as follows:

$$\delta(a', a) = \frac{1}{V} \lim_{s\to 1} \sum (2j+1) s^{2j} \left[\frac{\sin (2j+1)\frac{1}{2}w}{\sin \frac{1}{2}w}\right]$$

$$= \frac{1}{V} \lim_{s\to 1} \frac{d}{ds} [s\rho(s, a, a')]$$

$$= \frac{1}{V} \lim_{s\to 1} \left[\frac{(1-s^2)}{(1+s^2-2s\cos\frac{1}{2}w)^2}\right].$$

8. COORDINATE SYSTEMS

We have been using the familiar representation:

$$U = e^{\frac{1}{2}i\mathfrak{d}w} \tag{8.1a}$$

$$= \cos\tfrac{1}{2}w + i\mathfrak{d}\hat{w}\sin\tfrac{1}{2}w. \tag{8.1b}$$

This choice $(w_1 w_2 w_3)$ corresponds to the use of normal coordinates in differential geometry. The group manifold is then the interior of a three-dimensional sphere.

To put the group space into correspondence with the surface of a four-dimensional sphere write (8.1a) as

$$U = R_0^{-1}(y_0 + i\mathbf{y}\mathfrak{d}). \tag{8.2}$$

Then the unitary condition requires

$$y_0^2 + \sum_1^3 y_k^2 = R_0^2, \tag{8.2a}$$

and the group manifold becomes a closed three-dimensional hypersurface. The four-dimensional rotation group which carries this hypersphere into itself may be factored and gives rise to the two triads of generators $X_i(\pm)$.

In this paper, we shall work mainly with the representation

$$U = (1 + \tfrac{1}{2}iR)/(1 - \tfrac{1}{2}iR), \tag{8.3}$$

where

$$R = \mathfrak{d}\mathbf{r}/R_0. \tag{8.3a}$$

The coordinates are now $\mathbf{r}$ and arise from the stereographic projection of the hypersphere (8.2a) onto a tangent hyperplane.

From (8.2) and (8.3) we have

$$y_0 = \tfrac{1}{2}R_0(U + U^+)$$

or

$$y_0 = \left[\frac{1 - \frac{1}{4}(r^2/R_0^2)}{1 + \frac{1}{4}(r^2/R_0^2)}\right] R_0, \tag{8.4a}$$

and

$$i\mathbf{y}\mathfrak{d} = \tfrac{1}{2}R_0(U - U^+)$$

or

$$\mathbf{y} = \mathbf{r}[1 + \tfrac{1}{4}(r^2/R_0^2)]^{-1}. \tag{8.4b}$$

The connection between the geodesic distance w and the distance from the origin in stereographic coordinates is, of course, given by

$$e^{\frac{1}{2}i\mathfrak{d}w} = (1 + \tfrac{1}{2}iR)(1 - \tfrac{1}{2}iR)^{-1} \tag{8.5}$$

or

$$r/R_0 = 2\tan\tfrac{1}{4}w. \tag{8.5a}$$

9. STEREOGRAPHIC COORDINATES

Let us define a new matrix V in terms of the R matrix which appears in the preceding paragraph,

$$V = 1 - \tfrac{1}{2}iR. \tag{9.1}$$

Then

$$U = V^{-1}V^+ = V^+V^{-1}, \tag{9.2}$$

and

$$\Lambda_\alpha(+) = -i(\partial_\alpha U)U^{-1}, \quad \Lambda_\alpha(+) = R_0^{-1}V^{-1}\sigma_\alpha(V^{-1})^+, \tag{9.3}$$

while

$$\Lambda_\alpha(-) = R_0^{-1}(V^{-1})^+\sigma_\alpha V^{-1}. \tag{9.4}$$

According to (5.7) the associated triads are

$$\lambda^i_\alpha(\pm) = R_0 \operatorname{Tr} \Lambda_\alpha(\pm)\sigma^i. \tag{9.5}$$

One finds the following useful representations of $\lambda_\alpha^i(\pm)$,

$$\lambda_\alpha^i(+) = \mathrm{Tr}\ \sigma^i(V^{-1})\sigma_\alpha(V^{-1})^+, \qquad (9.6+)$$

$$\lambda_\alpha^i(-) = \mathrm{Tr}\ \sigma^i(V^{-1})^+\sigma_\alpha(V^{-1}). \qquad (9.6-)$$

The triads may also be expressed directly in terms of U,

$$\lambda_\alpha^i(+) = \tfrac{1}{4}\,\mathrm{Tr}\ \sigma^i(U+1)\sigma_\alpha(U^+ + 1), \qquad (9.7+)$$

$$\lambda_\alpha^i(-) = \tfrac{1}{4}\,\mathrm{Tr}\ \sigma^i(U^+ + 1)\sigma_\alpha(U+1). \qquad (9.7-)$$

In these formulas the Latin index may be written in either the covariant or contravariant position, but the Greek index is so far only a covariant index.

The preceding general formulas simplify for SU_2, since

$$VV^+ = 1 + \tfrac{1}{4}(r/R_0)^2 = G^{-1}(r). \qquad (9.8a)$$

We see that V is itself unitary except for $G(r)$, the scale factor:

$$V^{-1} = G(r)V^+ \qquad (9.8b)$$

and

$$\Lambda_\alpha(+) = (G/R_0)(V^{-1}\sigma_\alpha V), \qquad (9.9+)$$

$$\Lambda_\alpha(-) = (G/R_0)(V\sigma_\alpha V^{-1}). \qquad (9.9-)$$

The metric may now be computed from either $\Lambda_\alpha(+)$ or $\Lambda_\alpha(-)$,

$$\begin{aligned} g_{\alpha\beta} &= R_0^2\,\mathrm{Tr}\ \Lambda_\alpha(\pm)\Lambda_\beta(\pm) \\ &= G^2\,\mathrm{Tr}\ (V^{-1}\sigma_\alpha V)(V^{-1}\sigma_\beta V) \\ &= G^2\ \delta_{\alpha\beta}. \end{aligned} \qquad (9.10)$$

In stereographic coordinates, therefore, the metric differs from its Euclidean form by only the scale factor $G(r)^2$. We find

$$g^{\frac{1}{2}} = G^3, \qquad (9.11)$$

$$g^{\alpha\beta} = G^{-2}\ \delta_{\alpha\beta}. \qquad (9.12)$$

We now introduce

$$\tau_\alpha = G\sigma_\alpha \qquad (9.13a)$$

and

$$\tau^\alpha = g^{\alpha\beta}\tau_\beta = G^{-1}\sigma_\alpha. \qquad (9.13b)$$

Then

$$\Lambda_\alpha(+) = R_0^{-1}(V^{-1}\tau_\alpha V), \qquad (9.14+)$$

$$\Lambda_\alpha(-) = R_0^{-1}(V\tau_\alpha V^{-1}) \qquad (9.14-)$$

with

$$\{\tau_\alpha, \tau_\beta\} = 2g_{\alpha\beta}, \qquad (9.15)$$

and

$$R_0^2\{\Lambda_\alpha(\pm), \Lambda_\beta(\pm)\} = 2g_{\alpha\beta}. \qquad (9.16)$$

The Greek index appearing on τ may, of course, be raised with the aid of the metric tensor.

We are also interested in carrying out the computation of λ_α^i according to the following equations:

$$\lambda_\alpha^i(+) = \mathrm{Tr}\ \sigma^i V^{-1}\tau_\alpha V, \qquad (9.17+)$$

$$\lambda_\alpha^i(-) = \mathrm{Tr}\ \sigma^i V\tau_\alpha V^{-1}, \qquad (9.17-)$$

and, with the aid of (9.1), find

$$\begin{aligned} \lambda_\alpha^i(\pm) = {} & +G^2[(1-u)\ \delta_\alpha^i \\ & + \tfrac{1}{2}(R_0^{-2})r^i r_\alpha \mp \epsilon_{\alpha t}^{\ i} r^t (R_0^{-1})], \end{aligned} \qquad (9.18)$$

where

$$u = G^{-1} - 1.$$

This result for $\lambda_\alpha^i(\pm)$ holds for stereographic coordinates only. In normal coordinates we have[5]

$$\begin{aligned} \lambda_\alpha^i(\pm) = {} & \delta_\alpha^i(\sin w/w) + w_\alpha w^i(w - \sin w)/w^3 \\ & \pm w^s/w^2(1-\cos w)\epsilon_{\alpha s}^{\ i}. \end{aligned} \qquad (9.19)$$

Finally, the symmetric and antisymmetric parts of the connection are

$$\begin{aligned} L^\mu_{(\alpha\beta)} &= \tfrac{1}{2}g^{\mu\sigma}\left[\frac{\partial g_{\beta\sigma}}{\partial x^\alpha} + \frac{\partial g_{\alpha\sigma}}{\partial x^\beta} - \frac{\partial g_{\alpha\beta}}{\partial x^\sigma}\right] \\ &= [\delta_{\mu\alpha}\partial_\beta + \delta_{\mu\beta}\partial_\alpha - \delta_{\alpha\beta}\partial_\mu]\ \ln G, \end{aligned}$$

where $g_{\alpha\beta}$ is taken from (9.10); and

$$\begin{aligned} L^\mu_{[\alpha\beta]} &= -iR_0^2\,\mathrm{Tr}\ \Lambda^\mu(\Lambda_\alpha, \Lambda_\beta) \\ &= -iR_0^{-1}\,\mathrm{Tr}\ \tau^\mu(\tau_\alpha, \tau_\beta), \end{aligned}$$

with the aid of (5.16) and (9.14); Therefore,

$$\begin{aligned} T_{\mu\alpha\beta} &= -iG^3R_0^{-1}\,\mathrm{Tr}\ \sigma_\mu(\sigma_\alpha, \sigma_\beta) \\ &= (2G^3/R_0)\epsilon_{\alpha\beta\mu}. \end{aligned}$$

Comparing with (4.15), one also finds

$$\epsilon_{stp}\lambda_\alpha^s\lambda_\beta^t\lambda_\gamma^p = G^3\epsilon_{\alpha\beta\gamma} = g^{\frac{1}{2}}\epsilon_{\alpha\beta\gamma}.$$

10. INVARIANT VOLUME

The scalar $g^{\frac{1}{2}}da(= d\tau)$ in stereographic coordinates is

$$d\tau = G^3r^2\ dr\ d\Omega, \qquad (10.1)$$

where $d\Omega$ is the element of solid angle.

[5] D. Wenger, J. Math Phys. (to be published).

We may also express $d\tau$ in geodesic coordinates as follows. By (9.8) and (9.2),

$$G = \tfrac{1}{4}(1 + U)(1 + U^{-1}) = \tfrac{1}{4}(2 + U + U^{-1}). \tag{10.2}$$

But

$$U^2 - \chi U + 1 = 0,$$

where χ is the character of the two-dimensional representation. Therefore,

$$G = \tfrac{1}{4}(2 + \chi) = \tfrac{1}{2}(1 + \cos \tfrac{1}{2}w), \tag{10.3}$$

$$G = \cos^2 \tfrac{1}{4}w. \tag{10.4}$$

By computing dG from (9.8a) and (10.3), we find

$$2G^2 r\, dr = R_0^2 \sin \tfrac{1}{2}w\, dw. \tag{10.5}$$

Similarly, by computing G from these two equations, we obtain

$$Gr = R_0 \sin \tfrac{1}{2}w. \tag{10.6}$$

By (10.5), (10.6), and (10.2),

$$d\tau = \tfrac{1}{2}R_0^3 \sin^2 \tfrac{1}{2}w\, dw\, d\Omega. \tag{10.7}$$

The total invariant volume is then

$$\int d\tau = (4\pi R_0^3)(\tfrac{1}{2}) \int_0^{2\pi} \sin^2 \tfrac{1}{2}w\, dw \tag{10.8}$$

$$= 2\pi^2 R_0^3, \tag{10.9}$$

which is the volume of an Einstein space of radius R_0.

The limit of integration is here taken as 2π rather than π in order that the radial distance in (8.5a) may become infinite. We are then associating different points in group space with the angles w and $w + 2\pi$, and are, therefore, working with SU_2 rather than 0_3.

11. KERNEL OF INTEGRAL EQUATION

The kernel of (7.4) is expressed in geodesic coordinates by (7.3b), but w measures $R^{-1}R'$ and (7.3b) is therefore not given directly in terms of a and a'. Let us now express $\rho(1, a, a')$ explicitly in terms of a and a'.

We have, by definition of w,

$$\mathrm{Tr}\, U'U^{+} = \cos \tfrac{1}{2}w. \tag{11.1}$$

The preceding equation can be expressed in the following forms:

$$\mathrm{Tr}\,(U - U')(U - U')^{+} = 4 \sin^2 \tfrac{1}{4}w \tag{11.2}$$

and

$$G(R - R')^2 G' = 4 \sin^2 \tfrac{1}{4}w, \tag{11.3}$$

where U and R are the two-dimensional matrices of (8.3). Therefore,

$$\rho(1, a, a')^{-1} = \mathrm{Tr}\,(U - U')(U - U')^{+} \tag{11.2a}$$

in normal coordinates and

$$\rho(1, a, a')^{-1} = G(R - R')^2 G' \tag{11.3a}$$

in stereographic coordinates. With the aid of (11.3a) one may also write (7.4) in the stereographic form.

$$D^i_{mn}(\mathbf{r}') = (d_i/V) \int \rho(|\mathbf{r}' - \mathbf{r}|)\, D^i_{mn}(\mathbf{r}) g^{\frac{1}{2}}\, d\mathbf{r}, \tag{11.4}$$

where

$$\rho(|\mathbf{r}' - \mathbf{r}|) = [G(r')\, |\mathbf{r}' - \mathbf{r}|^2\, G(r)]^{-1} \tag{11.4a}$$

$$g^{\frac{1}{2}} = G^3(r). \tag{11.4b}$$

[Equation (11.4) is formally identical with the equation of the Coulomb field in momentum space, where the momentum variables are just the sterographic coordinates. It is then possible to take over the extensive work on the Coulomb problem to construct wavefunctions in the Clifford space, and conversely to gain some geometrical insight into the Coulomb problem by discussing its formulation in Clifford space.]

12. FOUR-DIMENSIONAL SPACE

The equations of Secs. 1, 2, and 9, describe a four-dimensional space having three-dimensional sections, which may be identified with the group manifold of the rotation group. In particular, the metric and connection are given in (2.5) and (2.4) in terms of the tetrads $\lambda^i_\mu(\pm)$. The explicit form of the three dimensional triads is given in (9.18) and (9.19), and we shall now check to see if the corresponding tetrads are also given by the formulas (4.7) or

$$\lambda^i_\alpha = R_0\, \mathrm{Tr}\, \Lambda_\alpha \sigma^i, \tag{12.1}$$

where

$$\sigma_0 = \sigma^0 = 1, \tag{12.1a}$$

and

$$\Lambda_0 = i/R_0. \tag{12.1b}$$

One finds

$$\lambda^0_\mu = \mathrm{Tr}\, \Lambda_\mu = 0, \qquad \mu = 1, 2, 3, \tag{12.2a}$$

$$\lambda^0_0 = i\, \mathrm{Tr}\, 1 = i, \tag{12.2b}$$

and

$$\lambda^k_0 = \mathrm{Tr}\, \sigma^k = 0, \qquad k = 1, 2, 3. \tag{12.2c}$$

The additional components of the metric, still given by (5.2), are

$$g_{00} = -1, \qquad g_{0k} = 0. \qquad (12.3a, b)$$

Similarly, the additional components of the connection, still given by (5.3), are

$$L^0_{\alpha\beta} = 0, \qquad L^\mu_{0\beta} = 0. \qquad (12.4a, b)$$

Thus the complete four-dimensional space is obtained by rigid displacement of the three-dimensional sections in the time direction.

13. PHYSICAL INTERPRETATION OF DISPLACEMENT OPERATORS

The fundamental physical operators are determined by the group of motions of the space. The corresponding displacement operators are

$$X_i(\pm) = i^{-1}\lambda^\mu_i \partial_\mu, \qquad (13.1)$$

where the fundamental triads are given in (9.18). Therefore,

$$X_4 = -i^{-1}\,\partial/\partial t \qquad (13.2)$$

and

$$X_k = p_k - (\tfrac{1}{4}R_0^{-2})A_k \pm R_0^{-1}L_k, \qquad (13.3)$$

where

$$p_k = (i^{-1})\,\partial/\partial r_k, \qquad (13.4)$$

$$L_i = r_i p_k - r_k p_i, \qquad (13.5)$$

$$A_k = r^2 p_k - 2r_k\theta, \qquad (13.6)$$

and

$$\theta = r^m p_m; \qquad (13.7)$$

the r_k are again stereographic coordinates. In the nonrelativistic approximation to the Minkowski case, the group generators are $-(i^{-1})\partial/\partial t$, p_k, and L_k, where the six generators representing linear and angular momentum are p_k and L_k. Here, to a first approximation, the $X_k(\pm)$ are the usual displacement operators p_k; to terms of order R_0^{-1} they are helicity operators

$$p_k \pm R_0^{-1}L_k,$$

and to terms of order $1/R_0^2$ they are corrected by an acceleration term (A_k), because the space is curved. The A_k and θ are identical with the acceleration and dilatation operators appearing in conformal relativity theories. These operators generate an algebra with the commutation table shown.

	p_k	A_k	L_k	θ
p_i	0	$2i\ \delta_{ik} + 2i\epsilon_{ikm}L^m$	$i\epsilon_{ikm}p^m$	$-ip_k$
A_i	$-2i\ \delta_{ik} - 2i\epsilon_{ikm}L^m$	0	$-i\epsilon_{ikm}A^m$	iA_k
L_i	$-i\epsilon_{ikm}p^m$	$i\epsilon_{ikm}A^m$	$i\epsilon_{ikm}L^m$	0
θ	ip_k	$-iA_k$	0	0

In the Minkowski case one may label states by eigenvalues of the four operators p_i and $-(i^{-1})\partial/\partial t$ which commute. Here one may use $X^2(\pm)$, $X_3(+)$, $X_3(-)$, and $-(i^{-1})\partial/\partial t$ which also commute.

14. SCALAR WAVE EQUATIONS OF FREE PARTICLES

Let the eigenfunctions of the commuting set $[X^2, X_3(+), X_3(-), -(i^{-1})\partial/\partial t]$ be u^j_{mn}. Then

$$X^2(\pm)u^j_{mm'} = [j(j+1)/R_0^2]u^j_{mm'}, \qquad (14.1)$$

$$X_3(+)u^j_{mm'} = (m/R_0)u^j_{mm'}, \qquad (14.2)$$

$$X_3(-)u^j_{mm'} = (m'/R_0)u^j_{mm'}, \qquad (14.3)$$

$$-\hbar(i^{-1})\,\partial/\partial t\, u^j_{mm'} = \epsilon_j u^j_{mm'}. \qquad (14.4)$$

We may satisfy these equations by choosing

$$u^j_{mm'}(\mathbf{r}, t) = A^j_{mm'}\, D^j_{mm'}(\mathbf{r}) \exp(-iE_j t/\hbar) \qquad (14.5)$$

according to (6.5), (6.6), and (6.7). Then the functions $u^j_{mm'}(\mathbf{r}, t)$ take over the role of usual plane waves which are eigenfunctions of p_i and $-(i^{-1})\partial/\partial t$. Unlike the plane waves, however, these functions are normalizable. In fact

$$\int |u^j_{mm'}(\mathbf{r}, t)|^2\, g^{\frac{1}{2}}\, d\mathbf{r} = 1, \qquad (14.6)$$

if

$$A^j_{mm'} = (d_j/V)^{\frac{1}{2}} = [(2j+1)/(2\pi^2 R_0^3)]^{\frac{1}{2}} \qquad (14.6a)$$

according to (6.8).

The differential Eqs. (14.1)–(14.3) are formally the same as the equations of the symmetric top, which is characterized by two angular momenta, one referred to fixed axes and the other to moving, or body-fixed, axes. The $X_k(\pm)$ satisfy the commutation laws of an angular momentum and may be put into correspondence with the fixed and moving axes in either way. These commute as they do for the top and are related by a rotation:

$$X^k(-) = \rho^k_l X^l(+). \qquad (14.7)$$

Even a classical point particle, without spin, traveling along a geodesic in a space with torsion,

has a turning angle associated with its motion, namely the turning of $\lambda(-)$ with respect to $\lambda(+)$. Quantum mechanically, a point particle without spin simulates in a space of constant torsion, the motion of an extended body with the shape of a symmetrical top in the way described by Eqs. (14.1)–(14.4). The two quantum numbers, m and m', refer to the right- and left-handed helicity operators, or to the corresponding $\lambda(+)$ and $\lambda(-)$ triad fields.

Equations (14.1)–(14.4) are incomplete since they provide no connection between the energy and the eigenvalues of the helicity operators. Since the present picture is not Lorentz-invariant, the usual wave equation does not hold. However, we shall require approximate Lorentz invariance in the sense that the wave equation holds with the neglect of terms in $1/R_0$. The simplest assumption for the scalar wave equation is

$$\Delta u - (c^{-2})\,\partial^2 u/\partial t^2 = (m_0^2c^2/\hbar^2)u, \tag{14.8a}$$

which is the same as

$$-X^2(\pm)u - (c^{-2})\,\partial^2 u/\partial t^2 = (m_0^2c^2/\hbar^2)u. \tag{14.8b}$$

It is possible to satisfy (14.8) and (14.5) provided that

$$\epsilon_j^2 = m_0^2c^4 + j(j+1)(\hbar c^2/R_0). \tag{14.9}$$

The energy is quantized, since the space is closed; but the spectrum is, of course, effectively a continuous one, since the spacing is determined by $\hbar c/R_0$, where R_0 is the radius of the space.

15. QUANTIZATION OF SCALAR FIELD

Since the volume of space is now finite, it is unnecessary to introduce the usual wavepackets to secure finite integrals.

Let us follow the familiar procedure of expanding the interpolating field

$$A(x) = \sum \{f_\alpha(x)A^\alpha(t) + \bar{f}_\alpha(x)\bar{A}^\alpha(t)\}, \tag{15.1}$$

where the asymptotic operators satisfy the time-independent commutation relations

$$\begin{aligned}(A^\alpha_{\text{in}}, \bar{A}^\beta_{\text{in}}) &= (A^\alpha_{\text{out}}, \bar{A}^\beta_{\text{out}}) = \delta^{\alpha\beta},\\ (A^\alpha_{\text{in}}, A^\beta_{\text{in}}) &= (A^\alpha_{\text{out}}, A^\beta_{\text{out}}) = 0.\end{aligned} \tag{15.2}$$

Then

$$[A_{\text{in}}(x), A_{\text{in}}(y)] = i\,\Delta(x, y), \tag{15.3}$$

where the time-independent commutators are

$$\Delta(x, y) = \Delta^+(x, y) - \Delta^+(y, x), \tag{15.3a}$$

with

$$i\,\Delta^+(x, y) = \sum f_\alpha(x)\bar{f}_\alpha(y). \tag{15.3b}$$

Normalize the $f_\alpha(x)$ as follows:

$$f_\alpha(x) = (2E_\alpha)^{-\frac{1}{2}}u^i_{mm'}(\mathbf{x}, t), \tag{15.4}$$

where $u^i_{mm'}(\mathbf{x}, t)$ is given by (14.5) and satisfies (14.6). Then

$$\begin{aligned}&-i\int g^{\frac{1}{2}}\,d\mathbf{x}\, f_\alpha(x)\overleftrightarrow{\frac{\partial}{\partial x_0}} f_\beta(x)\\ &\quad\equiv -i\int g^{\frac{1}{2}}\,d\mathbf{x}\left\{f_\alpha(x)\frac{\partial}{\partial x_0}\bar{f}_\beta(x) - \bar{f}_\beta(x)\frac{\partial}{\partial x_0}f_\alpha(x)\right\}\\ &\quad= \delta_{\alpha\beta}\end{aligned} \tag{15.5}$$

and

$$A^\alpha(t) = -i\int g^{\frac{1}{2}}\,d\mathbf{x}\, A(x)\overleftrightarrow{\frac{\partial}{\partial x_0}}\bar{f}_\alpha(x_0), \tag{15.6}$$

which differ from the usual expressions only by the substitution of the invariant volume element $g^{\frac{1}{2}}\,d\mathbf{x}$ for $d\mathbf{x}$. However, the f_α do not have their usual meaning, since they are solutions of the new wave equation

$$(\Delta - \partial^2/\partial t^2)f_\alpha = m^2 f_\alpha \tag{15.7}$$

with

$$E_\alpha^2 = m^2C^4 + j(j+1)(\hbar c/R_0)^2. \tag{15.7a}$$

By (15.3b) and (15.4), the Green's function $\Delta^+(x, y)$ is

$$i\,\Delta^+(x, y) = \sum (2E)_i^{-1}u^i_{nm}(x)\bar{u}^i_{nm}(y) \tag{15.8}$$

and

$$i(\partial/\partial x_0 - \partial/\partial y_0)\,\Delta^+(x, y) = \sum u^i_{nm}(x)\bar{u}^i_{nm}(y). \tag{15.9}$$

By (15.8) $\Delta^+(x, y)$ satisfies (15.7) for both x and y.

By (15.9)

$$\begin{aligned}\lim_{x_0\to y_0}\left[i\left(\frac{\partial}{\partial x_0} - \frac{\partial}{\partial y_0}\right)\Delta^+(x, y)\right]&\\ = \sum u^i_{nm}(\mathbf{x}, 0)\bar{u}^i_{nm}(\mathbf{y}, 0)&\\ = \delta(w),&\end{aligned} \tag{15.10}$$

according to (7.9) where w has the same meaning as in (7.2).

Notice also that (15.8) may be rewritten [see (7.21)] as

$$\begin{aligned}i\,\Delta^+(x, y) = \frac{1}{V}\sum \frac{2j+1}{E_j}\frac{\sin(2j+1)\frac{1}{2}w}{\sin\frac{1}{2}w}&\\ \cdot\exp\left[-iE_j(t_x - t_y)\right].&\end{aligned} \tag{15.11}$$

Therefore,

$$\Delta^+(\mathbf{x}, t; \mathbf{y}, t) = \Delta^+(\mathbf{y}, t; \mathbf{x}, t) \tag{15.12a}$$

and

$$\Delta(\mathbf{x}, t; \mathbf{y}, t) = 0 \qquad (15.12b)$$

by (15.3a). It follows from (15.10) and (15.12) that $\Delta(x, y)$ satisfies the same boundary conditions as the corresponding function in flat space; on the other hand, it satisfies the wave equation (15.7) instead of the Klein–Gordon equation.

16. REDUCTION FORMULAS AND SCATTERING AMPLITUDES

The scattering formulas holding now are trivially different from those valid in Minkowski space.

The definition of R-products is unchanged

$$R(x \mid x_1 \cdots x_n) = (-i)^n \sum \theta(x - x_1) \cdots \theta(x_{n-1} - x_n) \cdot [\cdots [A(x), A(x_1)] \cdots A(x_n)]. \qquad (16.1)$$

One obtains in the usual way[6] but with the aid of (15.6),

$$(\Phi, [R(x \mid x_1 \cdots x_n), \bar{A}_m^\alpha]\Psi) = i \int g^{\frac{1}{2}}\, dz \frac{\partial}{\partial z_0}\left(M \frac{\partial f_\alpha}{\partial z_0} - f_\alpha \frac{\partial M}{\partial z_0}\right), \qquad (16.2)$$

where

$$M = (\Phi, R(x \mid x_1 \cdots x_n z), \Psi) \qquad (16.2a)$$

is the matrix element of an R-product between two arbitrary normalizable states. The evaluation of the right side (16.2) continues in the usual way, but with the use of the new wave equation (15.7), so that

$$i \int g^{\frac{1}{2}}\, dz \frac{\partial}{\partial z_0}\left[M \frac{\partial f_\alpha}{\partial z_0} - f_\alpha \frac{\partial M}{\partial z_0}\right] = i \int g^{\frac{1}{2}}\, dz\{M\, \Delta f_\alpha - f_\alpha\, \Delta M\} + i \int g^{\frac{1}{2}}\, dz\, f_\alpha(K_z M)$$

with

$$\Delta = (g^{-\frac{1}{2}})\, \partial/\partial x^k\, g^{\frac{1}{2}} g^{kl} (\partial/\partial x^l)$$

and

$$K_z = \Delta - \partial^2/\partial t^2 - m^2.$$

The first integral vanishes, since Green's theorem holds for compact spaces:

$$\int g^{\frac{1}{2}}\, dz\, (M\, \Delta f_\alpha - f_\alpha\, \Delta M) = 0,$$

and therefore one has the usual results, e.g.,

$$[R(x \mid y), A_{\text{in}}^\alpha] = i \int f_\alpha(z) K_z R(x \mid yz) g^{\frac{1}{2}}\, dz.$$

Similarly, if one starts from

$$A_{\text{out}}(x) = A_{\text{in}}(x) + \int \Delta(x - y) j(y) g^{\frac{1}{2}}\, dy,$$

where

$$K_x A(x) = j(x),$$

one obtains the scattering relations

$$[A_{\text{out}}(x), A_{\text{in}}(y)] - [A_{\text{in}}(x), A_{\text{in}}(y)] = \iint \Delta(x, x')\, \Delta(y, y') K_{x'} K_{y'} R(x', y') \cdot [g(x')]^{\frac{1}{2}} [g(y')]^{\frac{1}{2}}\, dx'\, dy',$$

from which the usual dispersion formulas are obtained.

[6] H. Lehmann, K. Symanzik, and W. Zimmerman, Nuovo Cimento 6, 319 (1957).

Analytic Properties of Resolvents and Completeness*

L. FONDA, G. C. GHIRARDI, AND T. WEBER
Istituto di Fisica Teorica dell'Università, Trieste, Italy
AND
A. RIMINI
Istituto di Fisica Teorica and Istituto di Meccanica dell' Università, Trieste, Italy
(Received 9 February 1966)

The main purpose of this paper is to prove, by using a simple formal procedure, the completeness of the set of eigenstates of a non-Hermitian operator whose resolvent satisfies certain physically plausible analytic properties. It is also shown how the obtained completeness relation can be related to the scattering solutions of the eigenvalue equation with extension to the multichannel case. All proofs are heuristic only.

1. PRELIMINARY CONSIDERATIONS

NON-HERMITIAN operators occur rather frequently in the theory of nuclear reactions. However, their properties are scarcely investigated. In particular, very little is known about the completeness of the set of their eigenvectors. Nevertheless, in certain cases, it is important to know whether completeness holds true or does not.

It is the main purpose of this paper to show how one can prove the completeness of the set of eigenvectors of a non-Hermitian operator whose resolvent has suitable analytic properties. We shall also show how to relate the obtained completeness relation to the solutions of the eigenvalue equation of the given linear operator belonging to the continuous spectrum and satisfying outgoing- or incoming-wave boundary conditions. In particular, the method used is shown to be applicable in the multichannel case.

The non-Hermitian operators we have in mind are the effective Hamiltonians which occur in the theory of nuclear reactions. Such operators, though non-Hermitian, still possess features of a Hamilton operator, e.g., they contain a kinetic energy term. In conjecturing the analytic properties of the resolvent, we have been guided by the requirement that the considered operator be reasonable, on physical grounds, as an effective Hamiltonian.

All the proofs in this paper use physicists' language and have heuristic value only. Obviously this is a drawback, particularly in the present field, but it is not unreasonable to hope that similar results can be proved rigorously.

We briefly review some properties of the resolvent of a self-adjoint operator.[1] Let H be a self-adjoint linear transformation in Hilbert space, and let $H = \int_{-\infty}^{+\infty} \lambda dE_\lambda$. When z is in the resolvent set, we take into account the resolvent operator

$$G(z) = (zI - H)^{-1} = \int_{-\infty}^{+\infty} \frac{1}{z - \lambda} dE_\lambda, \qquad (1)$$

which is bounded and defined everywhere in the Hilbert space. The function $(\varphi, G(z)\psi)$, φ and ψ being arbitrary elements of the Hilbert space, is analytic in the resolvent set. If we write

$$G(z) = 1/z[I + \epsilon(z)],$$

$\epsilon(z)$ converges strongly to zero for $\mathrm{Im}\, z \neq 0$ in the limit $z \to \infty$. For any real λ_0, μ_1, and μ_2, one has

$$\lim_{z \to \lambda_0} (z - \lambda_0) G(z) = E_{\lambda_0} - E_{\lambda_0 - 0} \qquad (2)$$

and

$$-\lim_{\delta \to 0} \lim_{\epsilon \to 0} \frac{1}{2\pi i} \int_{\mu_1 + \delta}^{\mu_2 + \delta} d\mu \, [G(\mu + i\epsilon) - G(\mu - i\epsilon)] = E_{\mu_2 + 0} - E_{\mu_1 + 0}, \qquad (3)$$

where the limits mean strong convergence.[2]

In order to investigate the connection between the properties of the resolvent and the completeness of the set of eigenstates of a linear operator, not necessarily Hermitian, let us confine ourselves to a finite-dimensional linear vector space. In such a case, it is easily shown that completeness follows from assumptions which are an obvious generaliza-

* Work supported in part by the Istituto Nazionale di Fisica Nucleare.

[1] All the properties listed below for a self-adjoint operator can either be found in the literature or can be easily deduced. See, for example, M. H. Stone, *Linear Transformations in Hilbert Space* (American Mathematical Society, New York, 1964); F. Riesz and B. Sz. -Nagy, *Functional Analysis* (Frederick Ungar Publishing Company, New York, 1955); N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space* (Frederick Ungar Publishing Company, New York, 1961).

[2] The δ device is necessary in order to make the operators E_λ continuous to the right if there are proper values of H falling on the continuous spectrum.

tion of some of the properties of the resolvent of a self-adjoint transformation.

To be precise, the condition that the elements of the resolvent matrix have at most simple poles is necessary and sufficient for the completeness of the set of the eigenvectors.

The necessity of the condition is immediate. To show sufficiency let us write

$$G_{ij}(z) = [(zI - H)^{-1}]_{ij} \qquad (i, j = 1, 2, \cdots n).$$

These functions, by hypothesis, are analytic in the complex z-plane, exception made for a finite number of simple poles[3] in general not real, at the positions E_k. We define

$$P_{ij}^{(k)} = \lim_{z \to E_k} (z - E_k) G_{ij}(z). \tag{4}$$

Writing

$$G_{ij}(z) = 1/z[\delta_{ij} + \epsilon_{ij}(z)],$$

it follows from the definition that $\epsilon_{ij}(z)$ vanishes in the limit as $|z|$ goes to infinity. Let us consider a counterclockwise contour Γ which includes all poles of $G_{ij}(z)$. Integrating $G_{ij}(z)$ over this contour, the Cauchy integral formula yields

$$\int_\Gamma G_{ij}(z)\, dz = 2\pi i \sum_k P_{ij}^{(k)}.$$

On the other hand, the same integral can be evaluated by letting Γ go to infinity. Using the asymptotic behavior of $G_{ij}(z)$, we obtain

$$\int_\Gamma G_{ij}(z)\, dz = 2\pi i\, \delta_{ij}.$$

Therefore,

$$\sum_k P_{ij}^{(k)} = \delta_{ij}. \tag{5}$$

The operators $P^{(k)}$ are idempotent and mutually orthogonal. In fact,

$$\sum_j P_{ij}^{(k')} P_{jl}^{(k)} = \lim_{\substack{z' \to E_{k'} \\ z \to E_k}} (z' - E_{k'})(z - E_k) \sum_j G_{ij}(z') G_{jl}(z)$$

$$= \lim_{\substack{z' \to E_{k'} \\ z \to E_k}} (z' - E_{k'})(z - E_k) \frac{G_{il}(z) - G_{il}(z')}{z' - z}$$

$$= \lim_{z' \to E_{k'}} \left[\frac{z' - E_{k'}}{z' - E_k} P_{il}^{(k)} \right] = \delta_{kk'} P_{il}^{(k)}.$$

Besides, the operators $P^{(k)}$ satisfy

$$HP^{(k)} = E_k P^{(k)}. \tag{6}$$

[3] There are, of course, at most n poles.

In fact,

$$\sum_j H_{ij} P_{jl}^{(k)} = \lim_{z \to E_k} \{(z - E_k) \sum_j (H_{ij} - z\, \delta_{ij}) G_{jl}(z) + (z - E_k) z G_{il}(z)\} = E_k P_{il}^{(k)}.$$

This result, together with Eq. (5), shows that

$$H = \sum_k E_k P^{(k)}. \tag{7}$$

2. THE GENERAL CASE

We now turn our attention to a not necessarily Hermitian operator H, defined in a set everywhere dense in the Hilbert space, assuming certain analytic properties for its resolvent $G(z) = (zI - H)^{-1}$. The proof will be an extension of that given for finite matrices, but now it ceases to be rigorous. We make the following assumptions:

(1) There exists a set everywhere dense in the Hilbert space such that if φ and ψ are members of it, the function $(\varphi, G(z)\psi)$ is an analytic function of z with simple poles at the points E_b, not necessarily real, and a branch cut stretching on the real axis from $\bar{E}$ to $+\infty$, so that

$$\lim_{z \to E_b} (z - E_b)(\varphi, G(z)\psi) = (\varphi, P_b \psi), \tag{8}$$

$$\lim_{\varepsilon \to E + i0} (\varphi, G(z)\psi) - \lim_{\varepsilon \to E - i0} (\varphi, G(z)\psi) = -2\pi i (\varphi, P(E)\psi), \tag{9}$$

for $\qquad \bar{E} \leq E < +\infty.$

The properties of the defined operators P_b and $P(E)$ will be discussed below. We have assumed the isolated singularities to be at most simple poles since, for finite matrices, such assumption is necessary and sufficient for the completeness of the eigenstates of H. The cut has been assumed to stretch along the real axis, since, in physical problems, the part of the Hamiltonian which survives when the particles are separated is self-adjoint and describes the motion of free particles.

(2) There is no accumulation point for the poles E_b and no pole falls on the branch cut. Thus, in particular, $\lim_{z \to \bar{E}} (z - \bar{E})(\varphi, G(z)\psi) = 0$. The number of poles is then finite.

(3) The function $(\varphi, \epsilon(z)\psi)$, defined through the equality

$$(\varphi, \epsilon(z)\psi) = (\varphi, [zG(z) - I]\psi),$$

vanishes in the limit as $|z| \to \infty$, $\operatorname{Im} z \neq 0$.

We shall now show that, under conditions (1)–(3),

there exists a complete set of states belonging to H. Consider the integral

$$I = \int_c dz\ (\varphi, G(z)\psi),$$

where the contour $C = C_1 + C_2 + C_3 + C_4$ shown in Fig. 1 includes all the poles of $(\varphi, G(z)\psi)$. The integral I can be immediately evaluated, by the use of Cauchy theorem, giving

$$I = 2\pi i \sum_b (\varphi, P_b\psi).$$

On the other hand, we can compute I on the single paths C_1, C_2, C_3, and C_4. In fact,

$$\int_{C_1+C_2+C_3} dz\ (\varphi, G(z)\psi) = \int_{E+i\epsilon}^{\infty+i\epsilon} dz\ (\varphi, G(z)\psi) - \int_{E-i\epsilon}^{\infty-i\epsilon} dz\ (\varphi, G(z)\psi) + \int_{C_3} dz\ (\varphi, G(z)\psi).$$

We go to the limit $\epsilon \to 0$. In this limit, the contribution from C_2 vanishes while contours C_1 and C_3 run on the real axis, one on the lower rim and the other on the upper rim of the cut. We then obtain

$$\int_{C_1+C_2+C_3} dz\ (\varphi, G(z)\psi) = -2\pi i \int_E^\infty dE\ (\varphi, P(E)\psi).$$

The integral over C_4 is easily evaluated in the limit as the radius of the circle C_4 tends to infinity by using assumption (3), giving

$$\int_{C_4} dz\ (\varphi, G(z)\psi) = 2\pi i(\varphi, \psi).$$

Collecting the results we have

$$\sum_b (\varphi, P_b\psi) + \int_E^\infty dE\ (\varphi, P(E)\psi) = (\varphi, \psi).$$

Now, φ and ψ being members of a set everywhere dense in Hilbert space, we get

$$\sum_b P_b + \int_E^\infty dE\ P(E) = I. \tag{10}$$

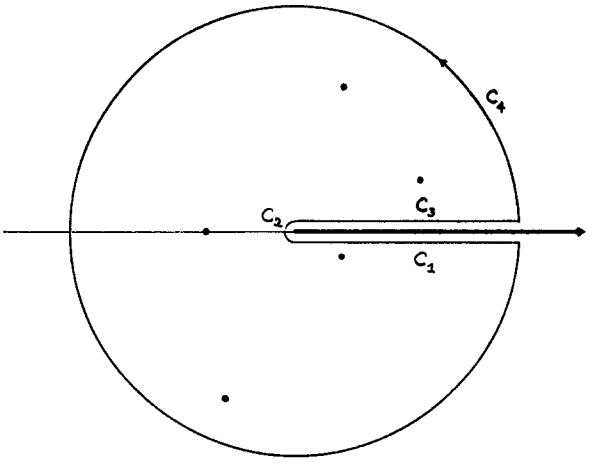


FIG. 1. The path C of integration. The dots indicate the poles E_b.

Let us now show that

$$\begin{aligned} P_{b_1}P_{b_2} &= \delta_{b_1 b_2}P_{b_1}, \\ P_bP(E) &= P(E)P_b = 0, \\ P(E)P(E') &= P(E)\ \delta(E - E'). \end{aligned} \tag{11}$$

The first of Eqs. (11) follows by using a procedure analogous to the one employed in the case of finite matrices.

For the second equation we have

$$\begin{aligned} &-2\pi i P(E)P_b \\ &= -2\pi i P_b P(E) \\ &= \lim_{\epsilon\to 0^+} \lim_{z\to E_b} (z - E_b)\frac{G(E + i\epsilon) - G(z)}{z - E - i\epsilon} \\ &\quad - \lim_{\eta\to 0^+} \lim_{z\to E_b} (z - E_b)\frac{G(E - i\eta) - G(z)}{z - E + i\eta} \\ &= \lim_{\epsilon\to 0^+} P_b\frac{1}{E + i\epsilon - E_b} - \lim_{\eta\to 0^+} P_b\frac{1}{E - i\eta - E_b} = 0. \end{aligned}$$

Let us now prove the third one. We have

$$\begin{aligned} (2\pi i)^2 P(E)P(E') &= \lim_{\epsilon\to 0^+}[G(E + i\epsilon) - G(E - i\epsilon)] \lim_{\eta\to 0^+}[G(E' + i\eta) - G(E' - i\eta)] \\ &= \lim_{\epsilon\to 0^+}\left[\frac{G(E + i\epsilon) - \lim_{\eta\to 0^+} G(E' + i\eta)}{E' - i\epsilon - E} + \frac{G(E - i\epsilon) - \lim_{\eta\to 0^+} G(E' - i\eta)}{E' + i\epsilon - E}\right. \\ &\quad \left. + \frac{\lim_{\eta\to 0^+} G(E' - i\eta) - G(E + i\epsilon)}{E' - i\epsilon - E} + \frac{\lim_{\eta\to 0^+} G(E' + i\eta) - G(E - i\epsilon)}{E' + i\epsilon - E}\right] \\ &= [\lim_{\eta\to 0^+} G(E' + i\eta) - \lim_{\eta\to 0^+} G(E' - i\eta)]\left[\lim_{\epsilon\to 0^+}\left(\frac{1}{E' + i\epsilon - E} - \frac{1}{E' - i\epsilon - E}\right)\right] = (2\pi i)^2 P(E')\ \delta(E - E'). \end{aligned}$$

Thus we have completed the proofs of formulas (11).

We next prove that

$$P_b(E_b - H) = (E_b - H)P_b = 0, \qquad P(E)(E - H) = (E - H)P(E) = 0. \tag{12}$$

In fact, we have

$$P_b(E_b - H) = \lim_{z \to E_b} (z - E_b)G(z)(E_b - z + z - H)$$

$$= -\lim_{z \to E_b} (z - E_b)^2 G(z) + \lim_{z \to E_b} (z - E_b)G(z)(z - H).$$

Since $z = E_b$ is a simple pole and $G(z)(z - H) = 1$, we get $P_b(E_b - H) = 0$. The same proof can be given also for the product of the factors in reversed order, so that the first of Eqs. (12) is established. The verification of the second equation is trivial using the identity

$$\lim_{z \to E \pm i0} G(z)(E - H) = \lim_{z \to E \pm i0} (E - H)G(z) = I.$$

Using Eq. (10) we easily get

$$H = \sum_b E_b P_b + \int_E^\infty EP(E)\, dE. \tag{13}$$

Therefore, Eq. (10) is a completeness relation for the eigenvectors of the operator H.

We now show how to relate the completeness relation (10) to the solutions of the eigenvalue equation of the given linear operator, belonging to the continuous spectrum and satisfying outgoing- or incoming-wave boundary conditions.

We suppose that H is the sum of two operators K and V, the former is self-adjoint and the latter is not necessarily Hermitian. The continuous spectrum of K is supposed to cover the interval $(\bar{E}, +\infty)$. For simplicity, we assume that K has no discrete eigenvalue. In the language common to physicists, we write

$$\int da\, |\chi_a\rangle\langle\chi_a| = I, \tag{14}$$

$|\chi_a\rangle$ being a non-normalizable eigenvector of K belonging to the eigenvalue E_a. The variable a is a shorthand for the eigenvalues of a complete set of commuting observables containing K. Let us define the non-normalizable eigenvectors of H and $H^\dagger$ in terms of $|\chi_a\rangle$.

$$|U_a^+\rangle = [I + \lim_{\epsilon \to 0^+} G(E_a + i\epsilon)V]\, |\chi_a\rangle, \qquad |V_a^+\rangle = [I + \lim_{\epsilon \to 0^+} G^+(E_a - i\epsilon)V^+]\, |\chi_a\rangle. \tag{15}$$

Let us consider the integral

$$-2\pi i \int_E^\infty dE \lim_{\epsilon \to 0^+} \lim_{\eta \to 0^+} [G(E + i\epsilon) - G(E - i\eta)]$$

$$= \int_E^\infty dE \lim_{\epsilon \to 0^+} \lim_{\eta \to 0^+} (-i)(\epsilon + \eta)G(E + i\epsilon)G(E - i\eta).$$

Using Eq. (14) and the relations

$$G(z) = G_0(z) + G_0(z)VG(z) = G_0(z) + G(z)VG_0(z),$$

where $G_0(z) = (zI - K)^{-1}$, we get

$$-2\pi i \int_E^\infty P(E)\, dE = \int_E^\infty dE \lim_{\substack{\epsilon \to 0^+ \\ \eta \to 0^+}} (-i)(\epsilon + \eta) \int da\, G(E + i\epsilon)\, |\chi_a\rangle\langle\chi_a|\, G(E - i\eta)$$

$$= \int_E^\infty dE \lim_{\substack{\epsilon \to 0^+ \\ \eta \to 0^+}} (-i)(\epsilon + \eta) \int da\, [I + G(E + i\epsilon)V]\, |\chi_a\rangle\langle\chi_a|\, [I + VG(E - i\eta)]$$

$$\times (E + i\epsilon - E_a)^{-1}(E - i\eta - E_a)^{-1}$$

$$= \int_E^\infty dE \lim_{\substack{\epsilon \to 0^+ \\ \eta \to 0^+}} \int da\, [I + G(E + i\epsilon)V]\, |\chi_a\rangle\langle[I + G^+(E - i\eta)V^+]\chi_a|$$

$$\times [(E + i\epsilon - E_a)^{-1} - (E - i\eta - E_a)^{-1}]$$

$$= -2\pi i \int_E^\infty dE \int da\, |U_a^+\rangle\langle V_a^+|\, \delta(E - E_a).$$

We finally obtain

$$\int_E^\infty P(E)\, dE = \int da\, |U_a^+\rangle\langle V_a^+|. \tag{16}$$

Let us consider the linear manifold L_b, range of the operator P_b. In L_b we introduce a complete set

$\{U_{b,\alpha}\}$. Then one can define the vectors $\{V_{b,\alpha}\}$ orthogonal to all U_a^+ and satisfying

$$(V_{b,\alpha}, U_{b',\alpha'}) = \delta_{bb'}\,\delta_{\alpha\alpha'}.$$

It is easily realized that the set $\{U, V\}$ is bi-orthogonal and that

$$P_b = \sum_\alpha |U_{b,\alpha}\rangle\langle V_{b,\alpha}|. \tag{17}$$

Then, from (16), (17), and (10), one gets the completeness of such a bi-orthogonal set

$$\sum_{b,\alpha} |U_{b,\alpha}\rangle\langle V_{b,\alpha}| + \int da\, |U_a^+\rangle\langle V_a^+| = I. \tag{18}$$

Eq. (18), of course, holds also, when substituting the outgoing with the incoming wavefunctions and, in particular, when H is Hermitian.

The method used to obtain Eq. (16) applies also in the multichannel case.[4] Consider, in fact, the splitting of the self-adjoint Hamiltonian H pertaining to the cth channel,

$$H = K_c + V_c.$$

Here, V_c describes the interaction between the particles of channel c, and K_c is the Hamiltonian for these particles when they are separated and do not interact. The operator K_c has a continuous spectrum starting from E_c, and χ_a^c is the complete set of its eigenvectors. Let us define the outgoing- and incoming-wave boundary-condition eigenvectors of H in terms of the vectors χ_c^a,

$$\psi_a^{c^{\pm}} = [I + \lim_{\epsilon\to 0^+} (E_a \pm i\epsilon - H)^{-1} V_c]\chi_a^c.$$

Repeating now the procedure leading to Eq. (16) with the care of integrating from E_c to ∞, it is easily shown that the sets $\{\psi_a^{c^+}\}$ and $\{\psi_a^{c^-}\}$ are complete in the subspace of the Hilbert space spanned by the eigenvectors of H whose energy is greater than or equal to E_c.

[4] For previous consideration of this problem, see, for example, M. L. Goldberger and K. M. Watson, *Collision Theory* (John Wiley & Sons, Inc., New York, 1964).

Relativistic Hydromagnetic Waves and Group Velocity*

ALBERT C. GIERE†
Department of Physics, The Pennsylvania State University, University Park, Pennsylvania
(Received 25 January 1966)

The velocities of relativistic hydromagnetic waves in a compressible, perfect fluid of infinite conductivity are calculated in the framework of general relativity. In the absence of viscuous and Joule heat losses, the flow is isentropic, and, therefore, the wave surfaces are propagated without change of shape. The velocities are first obtained in terms of the four-vector magnetic field and then in terms of the three-dimensional field. Several limiting cases are considered, and, in particular, it is shown under what conditions the expressions reduce to the nonrelativistic forms. Finally, the group velocities are calculated. The existence of a group velocity for such waves is based on the fact that the velocities exhibit a directional dependence. The group velocity in this case is of significance because it is the velocity of energy propagation, just as in the case when dispersion exists.

1. INTRODUCTION

THE hydromagnetic wave velocities and corresponding group velocities in a compressible, perfect fluid of infinite conductivity are determined in the framework of general relativity. Hydromagnetic waves are usually defined as small amplitude oscillations which involve variations of the physical parameters from equilibrium, and furthermore, the wave velocities are frequency-independent. We consider the propagation of such small disturbances in a homogeneous, perfectly conducting fluid in a uniform constant magnetic field. The flow is assumed to be adiabatic and without dissipation. In the absence of dissipative losses, there is no dispersion in the wave velocities. However, since the velocities exhibit a directional dependence, there still exists a group velocity for such waves. Lighthill[1] has shown by a very general argument that the velocity of energy propagation for a plane monochromatic wave in an anisotropic medium is equal to its group velocity.

Various aspects of hydromagnetic waves in special relativity have been treated by several authors, among them Reichel,[2] Akhiezer and Polovin,[3] and Harris.[4] Bruhat[5] has determined the phase velocities within the framework of general relativity in terms of total rest-mass energy and the proper-frame magnetic field. The former condition does not require the use of the entropy form of the continuity equation. In our work, we write the total energy as a sum of rest-mass energy and thermodynamic internal energy, and make explicit use of the entropy equation. Furthermore, we obtain formulas for the velocities in terms of the four-vector magnetic field, i.e., in terms of arbitrary coordinates. Then, by a direct and relatively simple method, it is shown how the expressions may be written in terms of the three-dimensional magnetic field.

The wave velocities are calculated here by a method founded on the theory of singular surfaces which arises in the solution of hyperbolic differential equations with given Cauchy data. In the singular case, i.e., when no unique solution exists, the surfaces on which the data are given are known as characteristics and may be identified with wave surfaces.[6]

We give first a brief summary of the relativistic macroscopic formulation of an infinitely conducting fluid in an electromagnetic field, and obtain the necessary equations from which the hydromagnetic wave velocities can be found. The formulation is based on the rationalized Gaussian system of units. A detailed treatment of the general equations for an arbitrary macroscopic fluid is given by Pham.[7] In the following notations, Greek indices take on the values 0, 1, 2, 3 and Latin indices the values 1, 2, 3. The symbol ∇_α denotes the covariant derivative with respect to the space–time coordinates x^α, and a metric of signature -2 is assumed, with the time coordinate $x^0 = ct$.

In a relativistic formulation, the four-vector elec-

* A part of a Ph. D. thesis submitted to the Pennsylvania State University.
† Presently at Clarkson College of Technology, Potsdam, New York.
[1] M. J. Lighthill, Phil. Trans. Roy. Soc. London **252**, 397 (1960).
[2] P. Reichel, "Basic Notions of Relativistic Hydromagnetics," Courant Institute of Mathematical Sciences, New York University, NYO-7697 (1958).
[3] I. A. Akhiezer and R. V. Polovin, Zh. Eksperim. i Teor. Fiz. **36**, 1845 (1959) [English transl.: Soviet Phys.—JETP **9**, 1316 (1959)].
[4] E. G. Harris, Phys. Rev. **108**, 1357 (1957).
[5] Y. Bruhat, Astronaut. Acta **6**, 354 (1960).
[6] G. Duff, *Partial Differential Equations* (University of Toronto Press, Toronto, Canada, 1956).
[7] M. Q. Pham, J. Ratl. Mech. Anal. 5, 473 (1956).

tric and magnetic fields are defined, respectively, by

$$e_\alpha = H_{\beta\alpha}u^\beta, \qquad h_\alpha = G^*_{\beta\alpha}u^\beta, \tag{1}$$

where u^α is the four-velocity of the fluid, whose magnitude is

$$u_\alpha u^\alpha = 1. \tag{2}$$

The quantities $H_{\alpha\beta}$ and $G_{\alpha\beta}$ are the electromagnetic field tensors defined by

$$H_{\alpha\beta} = \begin{bmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & B_3 & -B_2 \\ -E_2 & -B_3 & 0 & B_1 \\ -E_3 & B_2 & -B_1 & 0 \end{bmatrix},$$

$$G_{\alpha\beta} = \begin{bmatrix} 0 & D_1 & D_2 & D_3 \\ -D_1 & 0 & H_3 & -H_2 \\ -D_2 & -H_3 & 0 & H_1 \\ -D_3 & H_2 & -H_1 & 0 \end{bmatrix},$$

and $G^*_{\alpha\beta}$ is the dual tensor defined by

$$G^*_{\alpha\beta} = \tfrac{1}{2}\eta^{\alpha\beta\delta\gamma}G_{\delta\gamma}$$

with the usual definitions of $\eta^{\alpha\beta\delta\gamma}$ and its covariant form, namely

$$\eta^{\alpha\beta\delta\gamma} = -[1/(-g)^{\frac{1}{2}}]\epsilon^{\alpha\beta\delta\gamma},$$

$$\eta_{\alpha\beta\delta\gamma} = (-g)^{\frac{1}{2}}\epsilon_{\alpha\beta\delta\gamma},$$

The relative tensors $\epsilon_{\alpha\beta\delta\gamma} = \epsilon^{\alpha\beta\delta\gamma}$ are zero unless all indices are different, in which case they are equal to plus one if the indices are an even permutation of 0, 1, 2, 3, and minus one if this permutation is odd. The symbol g represents the determinant of the space–time metric tensor $g_{\alpha\beta}$. Furthermore, the electric and magnetic induction vectors are defined, respectively, by

$$d_\alpha = G_{\beta\alpha}u^\beta, \qquad b_\alpha = H^*_{\beta\alpha}u^\beta. \tag{3}$$

If the dielectric and magnetic properties of the medium were assumed constant, the definitions of the electric and magnetic four-vectors lead readily to a generalization of the constitutive relations

$$d_\alpha = \epsilon e_\alpha, \qquad b_\alpha = \mu h_\alpha. \tag{4}$$

Because of the antisymmetric character of the electromagnetic field tensors, one can easily verify that these four-vectors are orthogonal to u_α

$$e_\alpha u^\alpha = h_\alpha u^\alpha = d_\alpha u^\alpha = b_\alpha u^\alpha = 0. \tag{5}$$

Combining (1), (3), and (4), we obtain a single relation involving the electromagnetic field tensors

$$G_{\alpha\beta} = (1/\mu)H_{\alpha\beta} + [(1 - \epsilon\mu)/\mu] \times (H_{\gamma\alpha}u^\gamma u_\beta - H_{\gamma\beta}u^\gamma u_\alpha). \tag{6}$$

In general ϵ and μ are frequency dependent. In the limit of perfect conductivity, ϵ does not enter into the equations and the permeability μ can be assumed to be unity (or very close to it).

In a proper frame, i.e., in a frame moving with the fluid velocity, the four-velocity components are $u^0 = 1$, $u^i = 0$. It follows that in this frame $e_0 = h_0 = d_0 = b_0 = 0$, and the space components of the field vectors reduce to the ordinary three-dimensional quantities. Since the electric field within a perfectly conducting medium is zero, it follows that e_α must vanish in a proper frame. Inasmuch as e_α is a four-vector it must vanish in all frames. Hence

$$e_\alpha = H_{\beta\alpha}u^\beta = 0.$$

This expression can be taken as a definition of an infinitely conducting fluid. By assuming a Galilean metric such that

$$u^0 = (1 - \beta^2)^{-\frac{1}{2}}, \qquad u^i = (v^i/c)(1 - \beta^2)^{-\frac{1}{2}},$$

where $\beta = v^2/c^2$, v being the three-dimensional fluid velocity, the vanishing of e_α is equivalent to

$$\mathbf{E} + (1/c)\mathbf{v}\times\mathbf{B} = 0,$$

which is the expression one assumes in nonrelativistic magnetohydrodynamics in the limit of infinite conductivity. Furthermore, for the infinitely conducting case, the connecting relation (6) clearly reduces to

$$G_{\alpha\beta} = (1/\mu)H_{\alpha\beta} \tag{7}$$

so that the second relation of (1) becomes

$$h_\alpha = (1/\mu)H^*_{\beta\alpha}u^\beta. \tag{8}$$

From (8) and the definition of the dual tensor, we obtain

$$H_{\alpha\beta} = -\mu\eta_{\alpha\beta\delta\gamma}h^\delta u^\gamma, \tag{9}$$

$$H^*_{\alpha\beta} = \mu(h_\beta u_\alpha - h_\alpha u_\beta). \tag{10}$$

We also need the general Maxwell equations

$$\nabla_\alpha H^{\alpha\beta *} = 0,$$

which in view of (10) become

$$\nabla_\alpha(h^\beta u^\alpha - h^\alpha u^\beta) = 0. \tag{11}$$

The general nonsymmetric energy-momentum tensor of the electromagnetic field is given by

$$\tau^{\alpha\beta} = \tfrac{1}{4}g^{\alpha\beta}(H_{\lambda\gamma}G^{\lambda\gamma}) - H^{\gamma\beta}G^{\cdot\alpha}_{\gamma},$$

which in view of (7) takes the form

$$\tau^{\alpha\beta} = (1/\mu)[\tfrac{1}{4}g^{\alpha\beta}(H_{\lambda\gamma}H^{\lambda\gamma}) - H^{\gamma\beta}H_{\gamma}^{\ \alpha}]$$

and which is clearly symmetric. Employing the relation (9), $\tau^{\alpha\beta}$ may be expressed in terms of h^{α}. Thus

$$\tau^{\alpha\beta} = \mu(\tfrac{1}{2}g^{\alpha\beta} - u^{\alpha}u^{\beta})h_{\lambda}h^{\lambda} - \mu h^{\alpha}h^{\beta}. \tag{12}$$

Furthermore, with the aid of (12), (11), and the second relation of (5), it follows that the four-vector Lorentz force $\nabla_{\alpha}\tau^{\alpha\beta}$ is orthogonal to both u_{α} and h_{α}

$$u_{\beta}\nabla_{\alpha}\tau^{\alpha\beta} = 0, \tag{13}$$

$$h_{\beta}\nabla_{\alpha}\tau^{\alpha\beta} = 0. \tag{14}$$

The total energy-momentum tensor of the fluid is taken as

$$T^{\alpha\beta} = (w + p)u^{\alpha}u^{\beta} - pg^{\alpha\beta} + \tau^{\alpha\beta}, \tag{15}$$

where the first two terms constitute the mechanical stress tensor of a perfect fluid and $\tau^{\alpha\beta}$ is given by (12). The quantity p is the fluid pressure and w is the proper energy density. Denoting the proper mass density of the fluid by ρ, and the proper specific internal energy by $\mathcal{E}$, we write w as a sum of rest-mass energy and internal energy

$$w = \rho c^2 + \rho\mathcal{E}. \tag{16}$$

The equations of motion are given by

$$\nabla_{\alpha}T^{\alpha\beta} = 0. \tag{17}$$

Introducing (15) in (17) we obtain

$$wu^{\alpha}\nabla_{\alpha}u^{\beta} + u^{\beta}\nabla_{\alpha}(wu^{\alpha}) - (g^{\alpha\beta} - u^{\alpha}u^{\beta})\nabla_{\alpha}p + pu^{\alpha}\nabla_{\alpha}u^{\beta} + pu^{\beta}\nabla_{\alpha}u^{\alpha} + \nabla_{\alpha}\tau^{\alpha\beta} = 0. \tag{18}$$

By inner multiplication of (18) with u_{β} and simplifying this result with the use of (13), (2), and the identity

$$u_{\beta}\nabla_{\alpha}u^{\beta} = 0, \tag{19}$$

which follows from (2) by differentiation, the resultant expression is

$$\nabla_{\alpha}[(w + p)u^{\alpha}] - u^{\alpha}\nabla_{\alpha}p = 0. \tag{20}$$

Combining this result with the original expression (18), we have

$$(w + p)u^{\alpha}\nabla_{\alpha}u^{\beta} - (g^{\alpha\beta} - u^{\alpha}u^{\beta})\nabla_{\alpha}p + \nabla_{\alpha}\tau^{\alpha\beta} = 0. \tag{21}$$

The scalar equation (20) plays the role of a continuity equation, whereas the four equations of (21) determine the world lines of flow. Assuming a single-component fluid and the first law of thermodynamics, the energy density w may be expressed in terms of the proper temperature and the proper specific entropy S. This, in conjunction with the mass conservation relation

$$\nabla_{\alpha}(\rho u^{\alpha}) = 0, \tag{22}$$

allows us to write (20) as

$$u^{\alpha}\nabla_{\alpha}S = 0. \tag{23}$$

As would be expected the flow is isentropic. Introducing (12) in (21), we have

$$(w + p + \mu\,|h|^2)u^{\alpha}\nabla_{\alpha}u^{\beta} + (u^{\alpha}u^{\beta} - g^{\alpha\beta})\nabla_{\alpha}p - 2\mu(u^{\alpha}u^{\beta} - \tfrac{1}{2}g^{\alpha\beta})h_{\lambda}\nabla_{\alpha}h^{\lambda} + \mu\,|h|^2\,u^{\beta}\nabla_{\alpha}u^{\alpha} - \mu h^{\alpha}\nabla_{\alpha}h^{\beta} - \mu h^{\beta}\nabla_{\alpha}h^{\alpha} = 0, \tag{24}$$

where we wrote $|h|^2 = -h_{\alpha}h^{\alpha}$, since h_{α} is spacelike. Differentiating the second relation of (5) results in

$$h_{\beta}\nabla_{\alpha}u^{\beta} + u^{\beta}\nabla_{\alpha}h_{\beta} = 0. \tag{25}$$

Finally, we need an equation of state connecting the thermodynamic quantities, which will, in general, be taken as

$$\rho = \rho(p, S)$$

and which, by differentiation, yields

$$\nabla_{\alpha}\rho - a^2\nabla_{\alpha}p - b^2\nabla_{\alpha}S = 0, \tag{26}$$

where

$$a^2 = (\partial\rho/\partial p)_S, \qquad b^2 = (\partial\rho/\partial S)_p.$$

2. HYDROMAGNETIC WAVE VEOLCITIES

The system of equations we need is (11), (22), (23), (24), and (26). For convenience we collect these equations here in the same order,

$$h^{\beta}\nabla_{\alpha}u^{\alpha} + u^{\alpha}\nabla_{\alpha}h^{\beta} - h^{\alpha}\nabla_{\alpha}u^{\beta} - u^{\beta}\nabla_{\alpha}h^{\alpha} = 0, \tag{27}$$

$$u^{\alpha}\nabla_{\alpha}\rho + \rho\nabla_{\alpha}u^{\alpha} = 0, \tag{28}$$

$$u^{\alpha}\nabla_{\alpha}S = 0, \tag{29}$$

$$(w + p + \mu\,|h|^2)u^{\alpha}\nabla_{\alpha}u^{\beta} + (u^{\alpha}u^{\beta} - g^{\alpha\beta})\nabla_{\alpha}p - 2\mu(u^{\alpha}u^{\beta} - \tfrac{1}{2}g^{\alpha\beta})h_{\lambda}\nabla_{\alpha}h^{\lambda} + \mu\,|h|^2\,u^{\beta}\nabla_{\alpha}u^{\alpha} - \mu h^{\alpha}\nabla_{\alpha}h^{\beta} - \mu h^{\beta}\nabla_{\alpha}h^{\alpha} = 0, \tag{30}$$

$$\nabla_{\alpha}\rho - a^2\nabla_{\alpha}p - b^2\nabla_{\alpha}S = 0. \tag{31}$$

Multiplication of (27) by u_{β} and h_{β}, taking into account (19), (5), and (2), results in, respectively,

$$u^{\alpha}u_{\beta}\nabla_{\alpha}h^{\beta} - \nabla_{\alpha}h^{\alpha} = 0, \tag{32}$$

$$|h|^2\,\nabla_{\alpha}u^{\alpha} - u^{\alpha}h_{\beta}\nabla_{\alpha}h^{\beta} + h^{\alpha}h_{\beta}\nabla_{\alpha}u^{\beta} = 0. \tag{33}$$

Similarly, multiplication of (30) by u_β and h_β gives

$$\mu u^\alpha h_\beta \nabla_\alpha h^\beta - \mu |h|^2 \nabla_\alpha u^\alpha + \mu u_\beta h^\alpha \nabla_\alpha h^\beta = 0, \quad (34)$$

$$(w + p + \mu |h|^2) h_\beta u^\alpha \nabla_\alpha u^\beta - h^\alpha \nabla_\alpha p + \mu |h|^2 \nabla_\alpha h^\alpha = 0. \quad (35)$$

The equations (27) through (35) are 15 equations in the first derivatives of the physical variables. Since these equations must hold on either side of a hypersurface, on which the variables and their derivatives may be discontinuous, we now take the "jump" of these equations. The jump of a quantity F, denoted by $[F]$, is defined formally by

$$[F] = \lim_{\lambda \to 0} \{F(x^\alpha - \lambda \xi^\alpha) - F(x^\alpha + \lambda \xi^\alpha)\},$$

where x^α are the coordinates of a point on the singular hypersurface, λ is some positive quantity and ξ^α is arbitrary. We assume a weak discontinuity, i.e., the variables themselves are continuous but their first derivatives are discontinuous. That is

$$[F] = 0, \qquad [F_{,\alpha}] \neq 0.$$

This is the same assumption which is valid in the theory of ordinary sound waves. We assume further that the metric tensor and its first and second partial derivatives are continuous

$$[g_{\alpha\beta}] = [g_{\alpha\beta,\delta}] = [g_{\alpha\beta,\delta\gamma}] = 0.$$

Then, from the theory of singular surfaces, the above assumptions imply[8]

$$[F_{,\alpha}] = N_\alpha \, \delta F,$$

where N_α are the components of the four-vector unit normal to the hypersurface and δF is a quantity defined on the hypersurface and is related to the strength of the discontinuity. Moreover, since the metric tensor and its first derivatives are continuous, it follows that the Christoffel symbols are continuous, and, therefore, the jump of the covariant derivative is the same as the jump of the ordinary derivative.

Taking the jump of (27) through (35) we obtain, in the same order,

$$h^\beta N_\alpha \, \delta u^\alpha + L \, \delta h^\beta - P \, \delta u^\beta - u^\beta N_\alpha \, \delta h^\alpha = 0, \quad (36)$$

$$L \, \delta \rho + \rho N_\alpha \, \delta u^\alpha = 0, \quad (37)$$

$$L \, \delta S = 0, \quad (38)$$

$$(w + p + \mu |h|^2) L \, \delta u^\beta + (L u^\beta - N^\beta) \, \delta p - \mu (2 L u^\beta - N^\beta) h_\alpha \, \delta h^\alpha + \mu |h|^2 u^\beta N_\alpha \, \delta u^\alpha - \mu P \, \delta h^\beta - \mu h^\beta N_\alpha \, \delta h^\alpha = 0, \quad (39)$$

$$\delta \rho - a^2 \, \delta p - b^2 \, \delta S = 0, \quad (40)$$

$$L u_\alpha \, \delta h^\alpha - N_\alpha \, \delta h^\alpha = 0, \quad (41)$$

$$|h|^2 N_\alpha \, \delta u^\alpha - L h_\alpha \, \delta h^\alpha + P h_\alpha \, \delta u^\alpha = 0, \quad (42)$$

$$\mu L h_\alpha \, \delta h^\alpha - \mu |h|^2 N_\alpha \, \delta u^\alpha + \mu P u_\alpha \, \delta h^\alpha = 0, \quad (43)$$

$$(w + p + \mu |h|^2) L h_\alpha \, \delta u^\alpha - P \, \delta p + \mu |h|^2 N_\alpha \, \delta h^\alpha = 0, \quad (44)$$

where L and P are defined by

$$L = u^\alpha N_\alpha, \quad (45)$$

$$P = h^\alpha N_\alpha. \quad (46)$$

It is necessary to distinguish between the quantities that characterize the undisturbed fluid and the small variations of these quantities. That is, when we take the jump of an equation, it is implied that the continuous variables are replaced by their equilibrium values. Multiplication of the vector equation (36) by N_β vanishes identically. However, multiplication of (39) by this same quantity yields a further independent scalar equation, namely,

$$(w + p + 2\mu |h|^2) L N_\alpha \, \delta u^\alpha + (L^2 - Q^2) \, \delta p - \mu (2L^2 - Q^2) h_\alpha \, \delta h^\alpha - 2\mu P N_\alpha \, \delta h^\alpha = 0, \quad (47)$$

where

$$Q^2 = N_\alpha N^\alpha. \quad (48)$$

We may state at this point that a knowledge of the quantity L allows us to calculate the wave velocities through the use of the following expression,[9] valid in arbitrary coordinates:

$$U^2 = (u^\alpha f_{,\alpha})^2 [(u^\alpha u^\beta - g^{\alpha\beta}) f_{,\alpha} f_{,\beta}]^{-1}, \quad (49)$$

where $f(x^\alpha)$ is a function of the space–time coordinates such that $f(x^\alpha) = 0$ represents the equation of the singular hypersurface, and where $U = V/c$, V being the normal coordinate velocity of the wave surface relative to the fluid and c is the light velocity *in vacuo*. The wave surface is the two-dimensional surface associated with the hypersurface. This point is elaborated upon in a later paragraph. Since the components of the four-vector normal N_α are proportional to the derivatives $f_{,\alpha}$, and since this formula for U is homogeneous in these derivatives, it is readily clear that

$$U^2 = L^2/(L^2 - Q^2). \quad (50)$$

Also, since L is real it follows that the four-vector

[8] T. Y. Thomas, Proc. Natl. Acad. Sci. U. S. 51, 1 (1964).

[9] A. Lichnerowicz, *Theories Relativistes de la Gravitation et de L' Electromagnetisme* (Masson, Paris, 1955).

unit normal must be spacelike, i.e., $Q^2 = -1$, in order for U to be less than unity.

We now have 16 equations, (36) through (44) and (47), in the jumps of the derivatives of the physical variables. By a rather lengthy but straightforward process, we reduce the number of equations by successive elimination of the scalar products $N_\alpha \delta h^\alpha$, $h_\alpha \delta h^\alpha$, $u_\alpha \delta h^\alpha$, and $h_\alpha \delta u^\alpha$. The resulting equations are

$$\{P\phi(2Lu^\beta - N^\beta) + \mu\,|h|^2\, LQ^2(Lh^\beta - Pu^\beta) - \phi L^4 h^\beta\}\,\delta\rho + \rho\{PQ^2(Lu^\beta - N^\beta) + (L^2 - Q^2)[L^2 h^\beta - P(2Lu^\beta - N^\beta)]\}\,\delta p + LPQ^2\rho(\phi + \mu\,|h|^2)\,\delta u^\beta - \mu P^2 Q^2 \rho\,\delta h^\beta = 0, \quad (51)$$

$$L\{\mu PQ^2 h^\beta - (\mu\,|h|^2\,Q^2 + \phi L^2)Lu^\beta\}\,\delta\rho + L^2\rho u^\beta(L^2 - Q^2)\,\delta p + \mu P^2 Q^2 \rho\,\delta u^\beta - \mu LPQ^2\rho\,\delta h^\beta = 0, \quad (52)$$

$$L^2\phi(\mu\,|h|^2\,Q^2 + \phi L^2)\,\delta\rho - \rho[\phi L^2(L^2 - Q^2) + \mu P^2 Q^2]\,\delta p = 0, \quad (53)$$

$$\delta\rho - a^2\,\delta p - b^2\,\delta S = 0, \quad (54)$$

$$L\,\delta S = 0, \quad (55)$$

where we put

$$\phi = w + p = \rho c^2 + \rho\mathcal{E} + p. \quad (56)$$

The two vector equations, (51) and (52), each yields four equations. These, together with the three scalar equations (53), (54), and (55), comprise a system of 11 equations in the 11 quantities $\delta\rho$, δp, δS, δu^β, and δh^β, with $\beta = 0, 1, 2, 3$. For a nontrivial solution to exist the determinant of this system must vanish. One may expand the resulting eleventh-order determinant and thus obtain the characteristic equation for L. However, a simpler procedure is as follows. In (51) let A^β, B^β, C, and D denote the respective coefficients of $\delta\rho$, δp, δu^β, and δh^β. Similarly, in (52), let E^β, F^β, G, and H denote the respective coefficients of these same quantities, and let I and J represent the coefficients in (53). The equations now assume the form

$$A^\beta\,\delta\rho + B^\beta\,\delta p + C\,\delta u^\beta + D\,\delta h^\beta = 0, \quad (51')$$

$$E^\beta\,\delta\rho + F^\beta\,\delta p + G\,\delta u^\beta + H\,\delta h^\beta = 0, \quad (52')$$

$$I\,\delta\rho + J\,\delta p = 0, \quad (53')$$

$$\delta\rho - a^2\,\delta p - b^2\,\delta S = 0, \quad (54')$$

$$L\,\delta S = 0. \quad (55')$$

From (55'), the solution $L = 0$ corresponds to a "contact discontinuity" and implies a wave surface moving with the fluid velocity, for in this case (50) yields the velocity $U = 0$. For the case $L \neq 0$, $\delta S = 0$, we may eliminate δp and obtain

$$\left(A^\beta + \frac{1}{a^2}B^\beta\right)\delta\rho + C\,\delta u^\beta + D\,\delta h^\beta = 0, \quad (57)$$

$$\left(E^\beta + \frac{1}{a^2}F^\beta\right)\delta\rho + G\,\delta u^\beta + H\,\delta h^\beta = 0, \quad (58)$$

$$\left(I + \frac{1}{a^2}J\right)\delta S = 0, \quad (59)$$

which equations yield the determinant

$$\begin{vmatrix} A^\beta + \frac{1}{a^2}B^\beta & C & D \\ E^\beta + \frac{1}{a^2}F^\beta & G & H \\ I + \frac{1}{a^2}J & 0 & 0 \end{vmatrix} = 0, \quad (60)$$

from which follows

$$[I + (1/a^2)J](CH - DG) = 0. \quad (61)$$

Thus

$$I + (1/a^2)J = 0, \quad (62)$$

$$CH - DG = 0. \quad (63)$$

Introducing the values of the coefficients in these latter equations, we find, respectively, the characteristic equations for L

$$\phi(\phi a^2 - \rho)L^4 + \phi Q^2(a^2\mu\,|h|^2 + \rho)L^2 - \rho\mu P^2 Q^2 = 0, \quad (64)$$

$$(\phi + \mu\,|h|^2)L^2 - \mu P^2 = 0. \quad (65)$$

We note that (64) can be derived from (53), (54), and (55) alone, since these equations form a closed system in themselves. The two vector equations (51) and (52) are necessary for the derivation of (65), from which the transverse Alfvén wave velocity follows. With the aid of (50), the expressions (64) and (65) may be written in a form involving U

$$[a^2Q^2\phi(\phi + \mu\,|h|^2) - \rho\mu P^2]U^4 + [2\rho\mu P^2 - Q^2\phi(\rho + a^2\mu\,|h|^2)]U^2 - \rho\mu P^2 = 0, \quad (66)$$

$$[Q^2(\phi + \mu\,|h|^2) - \mu P^2]U^2 + \mu P^2 = 0. \quad (67)$$

The solution of (66) yields the velocities of the magnetosonic waves, whereas (67) yields the velocity

of the Alfvén waves. Thus, letting U_a represent this latter velocity, we have, respectively,

$$U^2 = \tfrac{1}{2}[a^2Q^2\phi(\phi + \mu\,|h|^2) - \rho\mu P^2]^{-1} \times \{Q^2\phi(\rho + a^2\mu\,|h|^2) - 2\rho\mu P^2 \pm [Q^4\phi^2(\rho + a^2\mu\,|h|^2)^2 + 4\rho\mu P^2Q^2\phi(\phi a^2 - \rho)]^{\frac{1}{2}}\}, \quad (68)$$

$$U_a^2 = [\mu P^2 - Q^2(\phi + \mu\,|h|^2)]^{-1}\mu P^2. \quad (69)$$

These expressions are valid in arbitrary coordinates since $|h|^2$ and P are relativistic invariants. In order for U to be less than c we must put $Q^2 = -1$. Thus, these become

$$U^2 = \tfrac{1}{2}[a^2\phi(\phi + \mu\,|h|^2) + \rho\mu P^2]^{-1}\{\phi(\rho + a^2\mu\,|h|^2) + 2\rho\mu P^2 \pm [\phi^2(\rho + a^2\mu\,|h|^2)^2 - 4\rho\mu P^2\phi(\phi a^2 - \rho)]^{\frac{1}{2}}\}, \quad (70)$$

$$U_a^2 = [\mu P^2 + \phi + \mu\,|h|^2]^{-1}\mu P^2. \quad (71)$$

In the absence of a magnetic field, $|h|^2 = P = 0$, and (70) reduces to the speed of sound U_s.

$$U_s^2 = \rho/\phi a^2 = \rho[a^2(\rho c^2 + \rho\mathcal{E} + p)]^{-1} \quad (72)$$
$$= [a^2c^2(1 + \rho\mathcal{E}/\rho c^2 + p/\rho c^2)]^{-1}$$
$$\simeq (a^2c^2)^{-1} \quad \text{if} \quad \left.\begin{matrix}\rho\mathcal{E}/\rho c^2 \\ p/\rho c^2\end{matrix}\right\} \ll 1.$$

This limiting form we recognize as the nonrelativistic formula for the speed of sound.

Moreover, putting $|h|^2 = P = 0$ in (53), and combining the resultant equation with (54) and (55) for the nontrivial case $L \neq 0$, we obtain the expression

$$\{[(a^2\phi/\rho) - 1]L^2 + Q^2\}\,\delta\rho = 0.$$

For $\delta\rho \neq 0$ and $Q^2 \leq 0$, and since L is real, it follows that

$$a^2 \geq \rho/\phi, \quad (73)$$

which is a restrictive relation on the equation of state in order for the wave speeds to be less than c. The equality sign corresponds formally to an incompressible fluid for, by putting $a^2 = \rho/\phi$ in (64), the resulting equation is (65), which, as we have seen, yields the Alfvén velocity.

We now express the invariant P in terms of the three-dimensional magnetic field, which is done by going over to a proper coordinate system. First, however, we need certain relations between the four-vector normal to a hypersurface and its three-dimensional velocity. These relations are readily obtained by an argument similar to one given by Synge.[10] Let the equation of a hypersurface in Galilean coordinates be given by

$$f(x^0, x^1, x^2, x^3) = 0. \quad (74)$$

For any time section $x^0 = \text{const}$, this equation yields a definite two-dimensional surface corresponding to an ordinary surface of physical space. The two-dimensional surface for $x^0 = X^0$ and the corresponding surface for $x^0 = X^0 + dx^0$ will be separated in the observer's time by an amount $dt = (1/c)dx^0$. If in this time interval the surface is displaced by an amount dx^i along its normal, the normal coordinate velocity will be

$$V^i = c(dx^i/dx^0). \quad (75)$$

Differentiating (74), dividing the result by dx^0, and taking (75) into account, we have

$$f_{,0} + (1/c)V^if_{,i} = 0. \quad (76)$$

The second term of this expression represents the scalar product of V^i with the ordinary gradient of f. Since the normal coordinate velocity is proportional to this gradient, we write

$$V^i = ka^{ij}f_{,j}, \quad (77)$$

where k is some constant and a^{ij} is the three-space metric tensor. Substituting (77) into (76), solving for k, and introducing the result back into (77), we obtain

$$V^i = -c(a^{ij}f_{,j}f_{,0}/a^{ij}f_{,i}f_{,j}).$$

The square of the normal coordinate speed is therefore,

$$V^2 = V_iV^i = -c^2[(f_{,0})^2/a^{ij}f_{,i}f_{,j}], \quad (78)$$

which may be rearranged to read

$$\frac{V^2}{c^2} = 1 - \frac{(f_{,0})^2 + a^{ij}f_{,i}f_{,j}}{a^{ij}f_{,i}f_{,j}},$$

or equivalently

$$U^2 = 1 + (g^{\alpha\beta}f_{,\alpha}f_{,\beta})(\delta^{ij}f_{,i}f_{,j})^{-1}, \quad (79)$$

where $U = V/c$, and $g^{\alpha\beta}$ is the space–time metric tensor in Galilean coordinates

$$g^{00} = 1, \qquad g^{ij} = a^{ij} = -\delta^{ij}, \qquad g^{i0} = g^{0i} = 0.$$

This is as far as we need to go for our purpose. It might be pointed out that Eq. (49), which is valid in arbitrary coordinates, follows from (79) by choosing an orthonormal tetrad of which the time-

[10] J. L. Synge, *Relativity: The Special Theory* (North-Holland Publishing Company, Amsterdam, 1958).

like vector is the unit four-velocity, and transforming the denominator in (79). It also follows from (79) that the four-vector normal must be spacelike for the wave speed to be less than c.

The four-vector normal is defined by

$$N_\alpha = f_{,\alpha}(-g^{\alpha\beta}f_{,\alpha}f_{,\beta})^{-\frac{1}{2}}, \tag{80}$$

which, combined with (79), results in

$$N_\alpha = f_{,\alpha}(1 - U^2)^{-\frac{1}{2}}(-a^{ij}f_{,i}f_{,j})^{-\frac{1}{2}}.$$

Hence, in view of (78)

$$N_0 = U(1 - U^2)^{-\frac{1}{2}}. \tag{81}$$

This result may also be obtained directly from (50) by setting $Q^2 = -1$ and evaluating L in a proper frame. Since in a proper frame $u^0 = 1$, $u^i = 0$, it follows that $L = N_0$. Furthermore,

$$N_i = -n_i(1 - U^2)^{-\frac{1}{2}},$$

in which the definition of n_i is obvious, and represents the three-dimensional unit normal. Thus, with the aid of these relations we find

$$\begin{aligned} P = h^\alpha N_\alpha &= (Uh^0 - h^i n_i)(1 - U^2)^{-\frac{1}{2}} \\ &= -h_n(1 - U^2)^{-\frac{1}{2}}, \end{aligned} \tag{82}$$

since h^0 vanishes in a proper frame. The quantity $h_n = h^i n_i$ is the component of the three-dimensional magnetic field vector normal to the wave surface. In the above expressions and in those that follow, it is assumed that $U \neq 1$.

The introduction of (82) in (66) yields a sixth-degree equation from which the quantity $(1 - U^2)$ can be factored out, with the result that

$$\begin{aligned} &a^2Q^2\phi(\phi + \mu\,|h|^2)U^4 \\ &+ [\rho\mu h_n^2 - Q^2\phi^2(\rho + a^2\mu\,|h|^2)]U^2 - \rho\mu h_n^2 = 0. \end{aligned} \tag{83}$$

Similarly, (67) becomes

$$Q^2(\phi + \mu\,|h|^2)U^2 + \mu h_n^2 = 0. \tag{84}$$

The solutions of (83) and (84) with $Q^2 = -1$ are, respectively,

$$\begin{aligned} U^2 = [2a^2\phi(\phi + \mu\,|h|^2)]^{-1}\{&\rho\mu h_n^2 + \phi(\rho + a^2\mu\,|h|^2) \\ &\pm [(\rho\mu h_n^2 + \phi\rho + a^2\phi\mu\,|h|^2)^2 \\ &- 4\rho\mu h_n^2 a^2\phi(\phi + \mu\,|h|^2)]^{\frac{1}{2}}\}, \end{aligned} \tag{85}$$

$$U_a^2 = \mu h_n^2/(\phi + \mu\,|h|^2), \tag{86}$$

where again we wrote U_a for the Alfvén velocity, and which, in view of the defining relation (56) for ϕ, becomes

$$U_a^2 = \mu h_n^2/(\rho c^2 + \rho\mathcal{E} + p + \mu\,|h|^2) \simeq \mu h_n^2/\rho c^2. \tag{87}$$

This last result represents the nonrelativistic limit, if the internal energy, fluid pressure, and magnetic energy are much less than the rest-mass energy.

We may write (86) in the form

$$U_a^2 = (\mathbf{U}_A \cdot \mathbf{n})^2,$$

where

$$\mathbf{U}_A = \mathbf{h}[\mu/(\phi + \mu\,|h|^2)]^{\frac{1}{2}}, \tag{88}$$

and is the velocity of an Alfvén wave traveling along the direction of the magnetic field. Since $h_n^2 = |h|^2 \cos\theta$, θ being the angle between $\mathbf{n}$ and $\mathbf{h}$, the magnetosonic velocities (85) may be written as

$$\begin{aligned} U^2 = {} &\tfrac{1}{2}[U_A^2 + \phi(\phi + \mu\,|h|^2)^{-1}U_s^2 + U_A^2U_s^2\cos^2\theta] \\ &\pm \tfrac{1}{2}\{[U_A^2 + \phi U_s^2(\phi + \mu\,|h|^2)^{-1} \\ &+ U_A^2U_s^2\cos^2\theta]^2 - 4U_A^2U_s^2\cos^2\theta\}^{\frac{1}{2}}. \end{aligned} \tag{89}$$

If $h_n = 0$, the nonzero solution of (85) is

$$\begin{aligned} U^2 &= (\rho + a^2\mu\,|h|^2)/a^2(\phi + \mu\,|h|^2) \\ &= U_A^2 + U_s^2\phi/(\phi + \mu\,|h|^2). \end{aligned} \tag{90}$$

Again, in the limit when the rest-mass energy predominates and $\mu\,|h|^2 \ll \phi$, then

$$\begin{aligned} \phi/(\phi + \mu\,|h|^2) &\simeq 1, \\ U_s^2 &\simeq (ac)^{-1}, \\ U_A^2 &\simeq \mu\,|h|^2/\rho c^2, \end{aligned} \tag{91}$$

and if, moreover, we neglect terms of order $1/c^2$ or smaller with respect to U_A and U_s, the expressions (89) and (90) assume their nonrelativistic values

$$\begin{aligned} U^2 \simeq \tfrac{1}{2}\{&U_A^2 + U_s^2 \pm [(U_A^2 + U_s^2)^2 \\ &- 4U_A^2U_s^2\cos^2\theta]^{\frac{1}{2}}\}. \end{aligned} \tag{92}$$

$$U^2 \simeq U_A^2 + U_s^2, \tag{93}$$

with the understanding that, in these approximate forms, U_A and U_s are the nonrelativistic values (91).

It might be noted that, in the usual problems of nonrelativistic magnetohydrodynamics, the displacement current is neglected, since it yields a small correction of order $1/c^2$. In a relativistic treatment, however, the displacement current is automatically taken into account when the Maxwell equations are cast in covariant form. Therefore, the nonrelativistic expressions given here, namely, the Alfvén velocity in (91), and the expressions (92) and (93) are those resulting from the neglect of the displacement current, since they were derived by neglecting terms of order $1/c^2$.

Finally, we classify the wave velocities according to their relative magnitudes. Manipulating Eq. (83) we observe that it can be written as

$$(U^2 - U_a^2)(U^2 - U_s^2)$$
$$= \mu(|h|^2 - h_n^2)(a^2\phi - \rho)U^2/a^2\phi(\phi + \mu\,|h|^2), \quad (94)$$

with U_a^2 and U_s^2 given by (86) and (72). Inasmuch as $a^2\phi \geq \rho$ and $|h|^2 \geq h_n^2$, it follows that the right member is a positive quantity. Denoting the larger and smaller of the roots of this equation by U_{fast} and U_{slow}, we deduce that

$$U_{slow} \leq U_s \leq U_{fast},$$
$$U_{slow} \leq U_a \leq U_{fast}.$$

The upper and lower values correspond to the positive and negative sign options, respectively, of (85). Thus, the relativistic sound and Alfvén velocities have magnitudes between the fast and slow magnetosonic waves, just as in the nonrelativistic case.

3. GROUP VELOCITIES

Denoting by $\mathbf{k}$ and ω as the propagation vector and circular frequency, respectively, the magnitude of the wave or phase velocity is ω/k, and the group velocity is given by $\partial\omega/\partial\mathbf{k}$. Since $\mathbf{k} = k\mathbf{n}$, $\mathbf{n}$ being the normal to the wave front, hence $h_n = \mathbf{h}\cdot\mathbf{n} = \mathbf{h}\cdot\mathbf{k}/k$. Furthermore,

$$U = V/c = (1/c)(\omega/k).$$

Making these substitutions in (86), the Alfvén wave velocity may be written as

$$\omega/c = [\mu/(\phi + \mu\,|h|^2)]^{\frac{1}{2}}\mathbf{h}\cdot\mathbf{k},$$

and the group velocity is found to be

$$(1/c)(\partial\omega/\partial\mathbf{k}) = [\mu/(\phi + \mu\,|h|^2)]^{\frac{1}{2}}\mathbf{h}. \quad (95)$$

The scalar product of this with $\mathbf{n}$ yields

$$[\mu/(\phi + \mu\,|h|^2)]^{\frac{1}{2}}\mathbf{h}\cdot\mathbf{n} = U_a. \quad (96)$$

Thus, the Alfvén group velocity is along $\mathbf{h}$ and its component along the propagation direction is equal to the phase velocity.

The group velocity of the magnetosonic modes (85) may be found in a similar fashion. However, it is more convenient to start with (83). With $Q^2 = -1$, (83) may be written in the form

$$U^4 - [a^2\phi(\phi + \mu\,|h|^2)]^{-1}[\rho\mu h_n^2 + (\rho + a^2\mu\,|h|^2)\phi]U^2$$
$$+ \rho\mu h_n^2[a^2\phi(\phi + \mu h_n^2)]^{-1} = 0,$$

or

$$U^4 - \alpha^2(h_n^2 + \beta^2)U^2 + \alpha^2 h_n^2 = 0, \quad (97)$$

where

$$\alpha^2 = \rho\mu/a^2\phi(\phi + \mu\,|h|^2),$$
$$\beta^2 = \phi(\rho + a^2\mu\,|h|^2)/\rho\mu. \quad (98)$$

Equation (97) is a quartic of the form $x^4 + qx^2 + r = 0$. Its solution may be written as

$$x = \pm\tfrac{1}{2}\{[-q + 2r^{\frac{1}{2}}]^{\frac{1}{2}} \pm [-q - 2r^{\frac{1}{2}}]^{\frac{1}{2}}\}.$$

The solution of (97) is therefore

$$U = \pm\tfrac{1}{2}\{[\alpha^2(h_n^2 + \beta^2) + 2\alpha h_n]^{\frac{1}{2}}$$
$$\pm [\alpha^2(h_n^2 + \beta^2) - 2\alpha h_n]^{\frac{1}{2}}\}, \quad (99)$$

which, in terms of ω and $\mathbf{k}$, becomes

$$\omega/c = \pm\tfrac{1}{2}\{[\alpha^2(\mathbf{h}\cdot\mathbf{k})^2 + \alpha^2\beta^2k^2 + 2\alpha k\mathbf{h}\cdot\mathbf{k}]^{\frac{1}{2}}$$
$$\pm [\alpha^2(\mathbf{h}\cdot\mathbf{k})^2 + \alpha^2\beta^2k^2 - 2\alpha k\mathbf{h}\cdot\mathbf{k}]^{\frac{1}{2}}\}. \quad (100)$$

Taking into account the fact that $k^2 = \mathbf{k}\cdot\mathbf{k}$, the group velocities in terms of $\mathbf{n}$ are

$$\frac{1}{c}\frac{\partial\omega}{\partial\mathbf{k}} = \pm\tfrac{1}{2}\{[(\alpha^2h_n^2 + \alpha)\mathbf{h} + (\alpha h_n + \alpha^2\beta^2)\mathbf{n}]$$
$$\times [\alpha^2(h_n^2 + \beta^2) + 2\alpha h_n]^{-\frac{1}{2}}$$
$$\pm [(\alpha^2h_n^2 - \alpha)\mathbf{h} - (\alpha h_n - \alpha^2\beta^2)\mathbf{n}]$$
$$\times [\alpha^2(h_n^2 + \beta^2) - 2\alpha h_n]^{-\frac{1}{2}}\}. \quad (101)$$

One can show that the scalar product of (101) with $\mathbf{n}$ results in (99), so that, as in the case of the Alfvén wave, the component of the magnetosonic group velocity in the direction of propagation is equal to the phase velocity.

We consider two special cases of (101). If $\mathbf{h} = 0$, (101) reduces to

$$(\rho/a^2\phi)^{\frac{1}{2}}\mathbf{n} = \mathbf{U}_s, \quad (102)$$

which is the group velocity of sound and is identical to the phase velocity, as would be expected, since the speed of sound is the same in all directions for the fluid being considered.

If $h_n = 0$, then (101) reduces to

$$[(\rho + a^2\mu\,|h|^2)/a^2(\phi + \mu\,|h|^2)]^{\frac{1}{2}}\mathbf{n}, \quad (103)$$

which is equal to the phase velocity (90). Thus, the phase and group velocities of this particular

mode are identical, as might be expected, since this velocity mode is independent of h_n. Equation (103) may also be derived directly from (90).

In the case when the rest-mass energy is much greater than the internal energy, the magnetic energy, and the pressure, we have

$$\alpha^2 \simeq \mu/\rho a^2 c^4, \qquad \beta^2 \simeq \rho c^2/\mu + |h|^2 a^2 c^2,$$

$$\alpha^2\beta^2 \simeq 1/a^2c^2 + \mu\,|h|^2/\rho c^2.$$

Thus, if in (101) we neglect the terms involving α^2 (not the terms involving the product $\alpha^2\beta^2$) we obtain the nonrelativistic limit

$$\frac{1}{c}\frac{\partial\omega}{\partial\mathbf{k}} \simeq \pm\tfrac{1}{2}\{[\alpha\mathbf{h} + (\alpha h_n + \alpha_1^2\beta^2)\mathbf{n}][\alpha^2\beta^2 + 2\alpha h_n]^{-\frac{1}{2}}$$

$$\pm [-\alpha\mathbf{h} - (\alpha h_n - \alpha^2\beta^2)\mathbf{n}][\alpha^2\beta^2 - 2\alpha h_n]^{-\frac{1}{2}}\}$$

with the understanding that α and β are now the approximate forms above.

ACKNOWLEDGMENTS

The author thanks Professor R. P. Kanwal for suggesting the problem and guiding this research to completion. Thanks are also due to Professor J. J. Gibbons and Professor E. Kazes for some valuable comments on certain aspects of this work.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 7, NUMBER 9 SEPTEMBER 1966

Local States*

A. L. LICHT†

U. S. Naval Ordnance Laboratory, Silver Spring, Maryland, and Department of Physics and Astronomy, University of Maryland, College Park, Maryland

(Received 8 April 1965; final manuscript received 1 February 1966)

A heuristic discussion is given of the preparation of states in finite space–time regions. Some axioms concerning such states are derived from this discussion. It is shown that states requiring for their preparation a selection of events according to the outcome of a measurement will not be "strictly local." Nonselective states will, however, be "strictly local." The mathematical structure of such states is investigated.

1. INTRODUCTION

A. Motivation

THE following is an analysis of the preparation of states in finite space–time regions. We attempt to derive the local properties of such states. It is felt[1,2] that this will eventually yield information concerning the rings[3,4] generated by the local observables.[5–7]

This in a sense is an investigation into the quantum theory of measurement. It differs from the usual treatments of this subject[8,9] mainly in the emphasis placed on the local character of the processes of measurement and state preparation.

A local measurement is an observation made in some finite space–time region α. It is a measurement of a local observable. To each such observable there corresponds an operator in the Hilbert space H. The set of such operators generates a ring, $R(\alpha)$, which we assume to be a von Neumann algebra.[3,4] It seems reasonable that if we wish to discuss local states, which generally involve local measurements, then we must discuss them in terms of the local rings $R(\alpha)$. The relevant assumptions that we make, following Araki,[6,7] concerning these rings are listed in Sec. 2.

The subject of local rings of observables grew out of quantum field theory.[5] The observables were

* An earlier version of part of this article was circulated as an N. O. L. preprint entitled, "Local States," in 1964.

† Presently at The Institute for Advanced Study, Princeton, New Jersey.

[1] A. L. Licht, J. Math. Phys. 4, 1443 (1963).

[2] A. L. Licht, Proceedings, Second Eastern Theoretical Physics Conference, Chapel Hill (1963).

[3] M. A. Naimark, *Normed Rings*, translated from the first Russian edition by L. F. Boron (P. Noordhoff, Ltd., Groningen, The Netherlands, 1959).

[4] J. Dixmier, *Les algebres d'operateurs dans l'espace Hilbertien* (Guathier-Villars, Paris, 1957).

[5] R. Haag, *Colloque Int. sur les problemes math. de la theorie quantique des champs. Lille (1957)*, CNRS, Ed. (Paris, 1959).

[6] H. Araki, Lecture Notes, Part (I), Zürich (1961).

[7] H. Araki, Progr. Theoret. Phys. (Kyoto) 32, 844 (1964).

[8] J. von Neumann, Göttingen Nachr. 245 (1927) [also in *Collected Works* (Pergamon Press, Ltd., London, 1961), Vol. I, p. 208]. J. Schwinger, Proc. Natl. Acad. Sci. U. S. 45, 1542 (1959).

[9] G. Ludwig, *Die Grundlagen der Quantenmechanik* (Springer-Verlag, Berlin, 1954).

mode are identical, as might be expected, since this velocity mode is independent of h_n. Equation (103) may also be derived directly from (90).

In the case when the rest-mass energy is much greater than the internal energy, the magnetic energy, and the pressure, we have

$$\alpha^2 \simeq \mu/\rho a^2 c^4, \qquad \beta^2 \simeq \rho c^2/\mu + |h|^2 a^2 c^2,$$

$$\alpha^2\beta^2 \simeq 1/a^2c^2 + \mu\,|h|^2/\rho c^2.$$

Thus, if in (101) we neglect the terms involving α^2 (not the terms involving the product $\alpha^2\beta^2$) we obtain the nonrelativistic limit

$$\frac{1}{c}\frac{\partial\omega}{\partial\mathbf{k}} \simeq \pm\tfrac{1}{2}\{[\alpha\mathbf{h} + (\alpha h_n + \alpha_1^2\beta^2)\mathbf{n}][\alpha^2\beta^2 + 2\alpha h_n]^{-\frac{1}{2}}$$

$$\pm [-\alpha\mathbf{h} - (\alpha h_n - \alpha^2\beta^2)\mathbf{n}][\alpha^2\beta^2 - 2\alpha h_n]^{-\frac{1}{2}}\}$$

with the understanding that α and β are now the approximate forms above.

ACKNOWLEDGMENTS

The author thanks Professor R. P. Kanwal for suggesting the problem and guiding this research to completion. Thanks are also due to Professor J. J. Gibbons and Professor E. Kazes for some valuable comments on certain aspects of this work.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 7, NUMBER 9 SEPTEMBER 1966

Local States*

A. L. LICHT†

U. S. Naval Ordnance Laboratory, Silver Spring, Maryland, and Department of Physics and Astronomy, University of Maryland, College Park, Maryland

(Received 8 April 1965; final manuscript received 1 February 1966)

A heuristic discussion is given of the preparation of states in finite space–time regions. Some axioms concerning such states are derived from this discussion. It is shown that states requiring for their preparation a selection of events according to the outcome of a measurement will not be "strictly local." Nonselective states will, however, be "strictly local." The mathematical structure of such states is investigated.

1. INTRODUCTION

A. Motivation

THE following is an analysis of the preparation of states in finite space–time regions. We attempt to derive the local properties of such states. It is felt[1,2] that this will eventually yield information concerning the rings[3,4] generated by the local observables.[5–7]

This in a sense is an investigation into the quantum theory of measurement. It differs from the usual treatments of this subject[8,9] mainly in the emphasis placed on the local character of the processes of measurement and state preparation.

A local measurement is an observation made in some finite space–time region α. It is a measurement of a local observable. To each such observable there corresponds an operator in the Hilbert space H. The set of such operators generates a ring, $R(\alpha)$, which we assume to be a von Neumann algebra.[3,4] It seems reasonable that if we wish to discuss local states, which generally involve local measurements, then we must discuss them in terms of the local rings $R(\alpha)$. The relevant assumptions that we make, following Araki,[6,7] concerning these rings are listed in Sec. 2.

The subject of local rings of observables grew out of quantum field theory.[5] The observables were

* An earlier version of part of this article was circulated as an N. O. L. preprint entitled, "Local States," in 1964.

† Presently at The Institute for Advanced Study, Princeton, New Jersey.

[1] A. L. Licht, J. Math. Phys. 4, 1443 (1963).

[2] A. L. Licht, Proceedings, Second Eastern Theoretical Physics Conference, Chapel Hill (1963).

[3] M. A. Naimark, *Normed Rings*, translated from the first Russian edition by L. F. Boron (P. Noordhoff, Ltd., Groningen, The Netherlands, 1959).

[4] J. Dixmier, *Les algebres d'operateurs dans l'espace Hilbertien* (Guathier-Villars, Paris, 1957).

[5] R. Haag, *Colloque Int. sur les problemes math. de la theorie quantique des champs. Lille (1957)*, CNRS, Ed. (Paris, 1959).

[6] H. Araki, Lecture Notes, Part (I), Zürich (1961).

[7] H. Araki, Progr. Theoret. Phys. (Kyoto) 32, 844 (1964).

[8] J. von Neumann, Göttingen Nachr. 245 (1927) [also in *Collected Works* (Pergamon Press, Ltd., London, 1961), Vol. I, p. 208]. J. Schwinger, Proc. Natl. Acad. Sci. U. S. 45, 1542 (1959).

[9] G. Ludwig, *Die Grundlagen der Quantenmechanik* (Springer-Verlag, Berlin, 1954).

originally assumed to be generated by smoothed polynomials in the field operators. The properties of these rings can, however, be formulated independently of the existence of any field operator.

Here we assume only that the ring $R(\alpha)$ is generated by the observables in α, without specifying exactly what these observables are. We do expect that these observables will eventually be determined by the analysis of local experiments. Among these observables there may be some that lead to the notion of a field operator, but we do not feel that it is necessary here to require this.

B. Historical

1. Newton–Wigner States

In the work of Newton and Wigner,[10] the vector states $\psi_{\mathbf{a}}$, for all three-vectors $\mathbf{a}$, are said to be localized at the points $\mathbf{a}$ if, among other things,

$$(\psi_{\mathbf{a}}, \psi_{\mathbf{b}}) = 0, \quad \text{for} \quad \mathbf{a} \neq \mathbf{b}.$$

This kind of localized state represents a physical situation in which some condition is known to hold at the point $\mathbf{a}$, and also known *not* to hold at all other points. Such a state could in practice be set up only by an infinite physicist, who would be able to examine all of space at a fixed time.

In the following we wish to examine those states that may be set up by a finite observer. For such an observer, there is some bounded region α which he may examine, and an infinite region surrounding α which he may not. It is conceivable that he may be able to set up some states that look like the Newton–Wigner states within α. Here, however, we are concerned with the properties of the set of all states that he may prepare, $\mathfrak{S}(\alpha)$, as seen from both inside and outside α.

2. The Reeh–Schlieder Theorem

A naive point of view would be to say that a state in $\mathfrak{S}(\alpha)$ could be obtained by letting an operator from $R(\alpha)$ act on the vacuum vector Ω, and that

$$\mathfrak{S}(\alpha) = R(\alpha)\Omega. \tag{1.1}$$

However, the following lemma leads to an immediate paradox.

Lemma 1[11,12]*:* (The Reeh–Schlieder theorem) If α contains an open region, then

$$\text{strong closure } (R(\alpha)\Omega) = H,$$

where H denotes the Hilbert space.

This lemma, although originally proved on the basis of field theory,[11] does not depend on the existence of a field, and has also been proved on the basis of the general ring axioms of Sec. 2.[12]

If Eq. (1.1) were to be true, then this lemma tells us that an observer in the region α, using an apparatus located entirely in α, could construct a state that was arbitrarily close to any state whatsoever. In particular, he could construct a state in which there was essentially nothing present in α, but with an arbitrarily large disturbance present 10 light years away from α in a spacelike direction. We conclude that Eq. (1.1) cannot be true.

It is shown in Theorem (4) below that, in principle, the expectation value of any given local operator in any vector state can be computed from measurements made in the vacuum. In a sense then, any state is physically producible from the vacuum. The physical process needed is, however, not simply a multiplication of the vacuum by an operator. It is seen below that the above objection to Eq. (1.1) does not apply to this process.

3. Strictly Localized States

It has been suggested that $\mathfrak{S}(\alpha)$ should consist of only "strictly localized" vector states.[1,13] A unit vector ψ is said to be strictly localized in α if for all operators $A \in R(\alpha')$, where α' denotes the spacelike complement of α,

$$(\psi, A\psi) = (\Omega, A\Omega). \tag{1.2}$$

The intuitive justification for this is the following:

(a). If a state has been prepared by an observer in α, then this should have no effect on measurements in α'. An observer in α' should then obtain the same values for a measurement of an operator as if no state had been prepared anywhere.

(b). If no state has been prepared anywhere, then it is felt that the expectation value for the measurement of an operator B should be $(\Omega, B\Omega)$.

Assumption (a) can be expected to be true only for what we call in Sec. 3A "nonselective" states. It will not be true for "selective" states.

Assumption (b) essentially says that the "background" state in which the experiment takes place is the vacuum vector state. There are certainly many experiments in which the vacuum is a good approxi-

[10] T. D. Newton and E. P. Wigner, Rev. Mod. Phys. **21**, 400 (1949); A. S. Wightman, *ibid.* **34**, 845 (1962). The strict orthogonality condition for such states is dropped in: T. O. Philips, Phys. Rev. **136**, B893 (1964); S. Schlieder, Commun. Math. Phys. **1**, 265 (1965).

[11] H. Reeh and S. Schlieder, Nuovo Cimento **22**, 1051 (1961).

[12] Reference 6, p. 57, Satz (10.2).

[13] J. M. Knight, J. Math. Phys. **2**, 459 (1961).

mation to the actual background. There are many in which it is not. We discuss in the following both the general case of an arbitrary background and the particular case of a vacuum background.

C. Summary

Section 2 lists the relevant assumptions made concerning local rings. Local states are discussed heuristically in Sec. 3. Their mathematical properties are investigated in Sec. 4.

The various stages of a local experiment are discussed in Sec. 3A. The effects of the locations of the regions of state preparation and measurement are discussed in Sec. 3B. In Sec. 3C we make certain assumptions concerning the average values. The various methods of composing states are covered in Sec. 3D.

In Sec. 4A, Gleason's theorem is used to show that these states must correspond to density matrices. Theorem 1 in Sec. 4B then gives the particular density matrix structure appropriate to a pure selective state.

Certain assumptions are made about a particularly ideal background, the uniformly empty background, in Sec. 4C. In Theorem 2 it is shown that these assumptions imply that this background is the vacuum vector state.

The structure of nonselective states prepared in the presence of the vacuum is then given by Theorem 3, in Sec. 4D.

Theorem 4 in Sec. 4E gives a physical procedure for approximating to any state in any region. This leads, in Sec. 4F, to Theorem 5, which gives the structure of nonselective states prepared in the presence of an arbitrary background.

2. RINGS

Unless otherwise specified, each space–time region α considered below is assumed to be such that both α and α' contain open regions. The algebra $R(X)$, where X denotes all space–time, we assume to equal $L(H)$, the ring of all bounded linear operators on H. Following Araki,[14] we assume

$$R(\alpha) \supset R(\beta), \quad \text{if} \quad \alpha \supset \beta, \tag{2.1}$$

where α could be X here.

There exists a unitary representation $U(a, \Lambda)$ of the inhomogeneous orthochromous Lorentz group $I^\uparrow$ such that

$$U(a, \Lambda)R(\alpha)U^\dagger(a, \Lambda) = R(\Lambda\alpha + a). \tag{2.2}$$

The usual spectrum conditions are also assumed.[14,15] In particular, we assume the existence of a unique vacuum vector $\Omega \in H$ such that

$$U(a, \Lambda)\Omega = \Omega. \tag{2.3}$$

Spacelike operators are assumed to commute. That is,

$$R(\beta) \subset R'(\alpha) \quad \text{if} \quad \beta \subset \alpha', \tag{2.4}$$

where R' denotes the commutant of the ring R.

Each $R(\beta)$, where β may be unbounded, is assumed to be generated by the $R(\alpha_i)$ for α_i bounded and contained in β. That is, let $\{\alpha_i, i = 1, 2 \cdots\}$ be a partitioning of β by bounded sets α_i, i.e., $\beta = \bigcup_{i=1}^{\infty}\alpha_i$.

Then

$$R(\beta) = \left\{\bigcup_{i=1}^{\infty} R(\alpha_i)\right\}''. \tag{2.5}$$

There exists a representation $\sigma(a, \Lambda)$ of $I^\uparrow$ such that

$$\sigma(a, \Lambda)\mathfrak{S}(\alpha) = \mathfrak{S}(\Lambda\alpha + a). \tag{2.6}$$

We also make the following postulate.

The Diamond Property[16]: Let $V(y, R, T)$ be any "pill box," i.e.,

$$V(y, R, T) = \{x \in X \mid |x^0 - y^0| < T, |\mathbf{x} - \mathbf{y}| < R\}.$$

Let $\tilde{V}(y, R + T)$ denote the "diamond" completion of V, i.e.,

$$V(y, R + T) = \{x \in X \mid |x^0 - y^0| < R + T - |\mathbf{x} - \mathbf{y}|\}.$$

Then

$$R(V(y, R, T)) = R(\tilde{V}(y, R + T)). \tag{2.7}$$

The set of points that cannot be causally affected by any point in a region α is denoted by α_-.

$$\alpha_- = \{x \in X| \quad \text{for all } y \in \alpha, \text{ either } (x - y)^2 < 0, \text{ or } (x - y)^2 > 0 \text{ and } (x^0 - y^0) < 0\}. \tag{2.8}$$

A set of regions, states, and operators $(\alpha_1, \cdots, \alpha_k; S_1, \cdots, S_l; B_1, \cdots, B_m)$ is said to be Lorentz congruent to the set $(\beta_1, \cdots, \beta_k; T_1, \cdots, T_l; C_1, \cdots, C_m)$ if there exists a common transformation (a, Λ) such that

[14] Reference 6, Secs. 7 and 8. Ref. 7, assumption (a).

[15] A. S. Wightman, Phys. Rev. 101, 860 (1956).

[16] R. Haag and B. Schroer, J. Math. Phys. 3, 248 (1962), Postulate (8b).

$$\beta_i = \Lambda\alpha_i + a, \quad i = 1, 2, \cdots, k,$$
$$T_i = \sigma(a, \Lambda)S_i, \quad i = 1, 2, \cdots, l,$$
$$C_i = U(a, \Lambda)B_iU^\dagger(a, \Lambda), \quad i = 1, 2, \cdots, m. \tag{2.9}$$

An operator A known to be in $R(\alpha)$ is occasionally written as A_α. Similarly, a state S in $\mathfrak{S}(\alpha)$ is written S_α.

In the following, we frequently use the fact[17] that if K is a ring, K'' is the ultra-weak[18] closure of K as well as the strong and weak closure.

3. LOCAL STATES

A. Experiments

There are five aspects to every experiment. These are the *background*, the *preparation* of the state, the *measurement*, the *repeated trials*, and the *averaging* of the results.

The *background* is that state of the universe which would have been present if the experiment had not taken place. It may not be completely known or controllable by the observer; therefore we do not consider it to be a local state. The background is usually assumed to be nearly constant and empty over the duration of the repeated trials which take place in the experiment. We consider below the ideal case of a uniformly empty background. In Theorem 2 this is shown to correspond to the vacuum vector state, as might be expected.

A local state is *prepared* by an observer in a finite space–time region α by performing a set of physical operations in α. These operations include the turning on of apparatus and the making of measurements. This produces a certain disturbance which propagates into the forward light cone subtended by α. Let S_α denote the set of such physical operations. We say that S_α is the local state prepared in α. The results of measurements will, however, in general be determined not by S_α alone, but by both S_α and the background state T. We say then that the actual state is not S_α but some kind of product TS_α.

A measurement of an observable A_β in the region β is a physical process that yields a number, the observed value for A_β, which we denote by $[A_\beta]$. As in ordinary quantum mechanics, we identify each such measured number with one of the eigenvalues of the operator A_β, and vice versa; each eigenvalue of A_β is a possible measured value.

The object of the experiment is to find whatever correlation might exist between the average value of an operator A_β and the preparation of a state S_α in the presence of the background T. An average value cannot be obtained from a single measurement. A large number of independent *repeated trials* is necessary. A single trial consists of the preparation of the state S_{α_i} in the region α_i, and the measurement of the operator A_{β_i} in the region β_i, where S_{α_i}, A_{β_i}, α_i, and β_i are all Lorentz-congruent to S_α, A_β, α, β, respectively.

The final *averaging* can be done in two different ways. All the experimental determinations of the observable A_β could be averaged together. We then say that the state S_α is *nonselective*, and we sometimes express this by a careted symbol $\hat{S}_\alpha$; or, only a subset of the experimental determinations could be averaged together. Namely, those determinations in which some preassigned condition had been found to be true in the region α. This we call a *selective* state, and denote it occasionally by a dotted symbol $\dot{S}_\alpha$.

In the following, we refer to the experiment consisting of the preparation of the state S_α and the measurement of A_β in the presence of the background T as the experiment $\{TS_\alpha, A_\beta\}$, using curly brackets. An individual trial in the same experiment is written occasionally as $(TS_{\alpha_i}, A_{\beta_i})$, using parenthesis.

B. The Effects of Location

The results of an experiment $\{TS_\alpha, A_\beta\}$ could be affected by the intervals between the repeated trials $(TS_{\alpha_i}, A_{\beta_i})$. In an ideal experiment these trials must be independent. This is usually arranged by making the space–time distances between the regions $\alpha_i \cup \beta_i$ sufficiently great.

The background T may be such that it is impossible to do some experiments in certain regions α and β. All regions would be allowed in a uniformly empty background, and perhaps there are also other nonrestrictive backgrounds.

Even assuming that one has a nonrestrictive background, not all relative locations of the regions α, β may be permissible. It is often impossible to put two different apparatuses in the same place. Therefore, if α and β overlap, it may not be possible to both prepare S_α and measure A_β. Also, if β can causally affect α, then it is conceivable that the measurement of A_β may so affect conditions in α as to make it impossible to prepare S_α. However, if α and β are disjoint, and if β cannot causally affect α, then the experiment $\{TS_\alpha, A_\beta\}$ should be possible, the background permitting.

If α and β overlap, and the experiment $\{TS_\alpha, A_\beta\}$

[17] Reference 4, p. 44, Corollary (1).
[18] Reference 4, p. 35.

is still possible, then we say that A_β is *compatible* with S_α. This is the case when the same apparatus can both prepare S_α and measure A_β.

When α and β are disjoint, it should in some cases be possible to do $\{TS_\alpha, A_\beta\}$, the background permitting, even if β is in the *past* relative to α. In the four-dimensional viewpoint taken here, the difference between the preparation of the state S_α and the measurement of the operator A_β is not in the time order of α and β, but in the treatment of the results of the experiment. The measurement of A_β is a certain physical process that yields a set of numbers that are later averaged. The state S_α is that set of all other physical processes that are performed in conjunction with each independent measurement of A_β, and which may be used to determine the manner of averaging. The processes involved in S_α may be as well carried out after, as before the measurement of A_β. The dependence of A_β on S_α will, of course, be very different in the two cases.

C. The Average Value

In this section we give three axioms concerning the average value to be found for the measurement of an observable. Axiom I postulates the existence of the average value, subject to certain restrictions discussed above. Axiom II gives the locality properties of this average value. Finally, in Axiom III we postulate the existence of a certain linear functional which is identical with the average value for certain experiments.

Axiom I:

(a) If the experiment $\{TS_\alpha, A_\beta\}$ is possible, then the average value for A_β in the state TS_α exists. We denote it by $M(TS_\alpha, A_\beta)$.

(b) Except for certain backgrounds T, the experiment is possible if α and β are disjoint, and β cannot causally affect α.

(c) The experiment in some cases may be possible even if α and β overlap, or if β can causally affect α.

Suppose that the region β is either spacelike or past relative to α; that is, $\beta \subset \alpha_-$. Then β cannot be causally affected by whatever is done in α. In the series of trials $(TS_{\alpha_i}, A_{\beta_i})$, the sequence of numbers $[A_{\beta_i}]$ will be the same as if the states S_{α_i} had not been set up. The average over all the numbers $[A_{\beta_i}]$ will then be the average value to be expected for A_β in the presence of the background state T. If S_α is a nonselective state, this same over-all average is the one appropriate for the state TS_α. We assume therefore the following axiom.

Axiom II: If $\beta \subset \alpha_-$, and if S_α is nonselective, then

$$M(TS_\alpha, A_\beta) = M(T, A_\beta).$$

If, however, S_α is a selective state, the average appropriate for TS_α will depend on the results of measurements made in the regions α_i. Let $\{i = n_j, j = 1, 2, \cdots\}$ index the subset of trials in which these measurements show that $S_{\alpha_{n_j}}$ is successfully set up. Then we must average only the numbers $[A_{\beta_{n_j}}]$. This average could depend on S_α, as well as on the background, even if β is spacelike relative to α. This is because events occurring in the background T, in the common past of α and β, could set up spacelike correlations between α and β. We thus restrict Axiom II to nonselective states.

Consider now the case when S_α is selective, and β can causally affect α, and $\{TS_\alpha, A_\beta\}$ is still possible. Measurements made in α may then be influenced by the measurement of A_β. In particular, the measurements that monitor S_α can be so influenced. The average over the numbers $[A_{\beta_{n_j}}]$ could then depend in a roundabout, complicated way on A_β. Indeed, we show below in Appendix A that in at least one such case the average is a nonlinear function of A_β.

The function $M(TS_\alpha, A_\beta)$ is what is determined in actual experiments. However, it is not a very convenient thing to handle mathematically. It is not defined for all T, S_α and A_β, and it may not be a linear function of A_β if β can causally affect α. We therefore require the following axiom.

Axiom III: For every possible state TS_α, there exists a normalized, positive linear functional, ultra-weakly continuous[18] on $L(H)$, $E(TS_\alpha, A)$ such that, if β cannot causally affect α, and if the experiment $\{TS_\alpha, A_\beta\}$ is possible, then

$$E(TS_\alpha, A_\beta) = M(TS_\alpha, A_\beta).$$

We require that E be normalized and positive because M must be so. The set of measured values of an operator has been identified with its set of eigenvalues. An average of observations of the unit observable can therefore only be one. An average of observations of a positive operator can likewise only be positive. Thus M must be both normalized and positive. We require ultra-weak continuity to simplify the mathematics.

If β can causally affect α, then $E(TS_\alpha, A_\beta)$ may not equal $M(TS_\alpha, A_\beta)$. An example of this is given in Appendix A. It is often assumed that $E(TS_\alpha, A_\beta)$ is even in this case a physically determinable average value. Namely, it is the expectation value of

A_β in some state, prepared in the past of both β and α, that could have led in time to S_α. Such a state may very well exist, but we do not consider it here.

D. Composition of States

Independent Superposition

Suppose that in a sequence of independent trials in the presence of the background T, a state S_α^1 is prepared a fraction r times, and a state S_β^2 is prepared a fraction $(1 - r)$ times. This procedure provides us with a certain kind of composite state. We denote it by

$$TS_\gamma = T(rS_\alpha^1 \oplus (1 - r)S_\beta^2),$$

and we regard it as a local state, preparable in the region $\gamma = \alpha \cup \beta$. This state is usually called the independent superposition of the states TS_α^1 and TS_β^2. Clearly, if the experiments $\{TS_\alpha^1, A_\delta\}$ and $\{TS_\beta^2, A_\delta\}$ are possible, then $\{TS_\gamma, A_\delta\}$ is possible and

$$M(TS_\gamma, A_\delta) = rM(TS_\alpha^1, A_\delta) + (1 - r)M(TS_\beta^2, A_\delta).$$

Products of States

Suppose that during each trial in the presence of the background T, an observer attempts to prepare S_α^1, and another observer attempts to prepare S_β^2. A composite state is formed which we call the product state $TS_\alpha^1 S_\beta^2$. We regard it as a local state, if it exists.

The state $TS_\alpha^1 S_\beta^2$ may not, however, actually exist. This composition could be forbidden by the background T. Regardless of T, the preparation of one state may make it impossible to prepare the other.

Even if $TS_\alpha^1 S_\beta^2$ does exist, the expectation value $M(TS_\alpha^1 S_\beta^2, A_\gamma)$ will have no simple relationship to $M(TS_\alpha^1, A_\gamma)$ and $M(TS_\beta^2, A_\gamma)$. The following theorems do, however, make clear the relationships for some cases.

Coherent Superposition

In ordinary quantum mechanics there is still a third kind of composition of states. There one has vector states, $\psi_1, \psi_2 \in H$, which can be added to produce a new vector state

$$\psi_3 = a\psi_1 + b\psi_2,$$

which is called a "coherent superposition" of ψ_1 and ψ_2.

It is possible to show that what is actually done in finite laboratories when coherent superposition is practiced, is to combine in a certain way local states S_α^1, S_β^2 which are in general mixtures, not vector states.[19] This will be discussed in greater detail in a later publication.

4. STRUCTURE

A. Gleason's Theorem

Axiom III strongly restricts the structure of the local states. For, consider the following lemma.[20]

Lemma 2: (Gleason's theorem) If F is a normalized, ultra-weakly continuous,[18] positive linear functional on $L(H)$, then there exists a sequence of vectors $\psi_i \in H$, $i = 1, 2, \cdots$, such that

$$F(C) = \sum_{i=1}^{\infty} (\psi_i, C\psi_i),$$

for all $C \in L(H)$. In particular, there exists an orthogonal sequence,

$$(\psi_i, \psi_j) = 0, \qquad i \neq j.$$

The function $E(TS_\alpha, C)$, according to Axiom III, satisfies the conditions of this lemma. Thus we have

Lemma 3: For every state TS_α, there exists a sequence of orthogonal vectors $\{\psi_i\}$, such that

$$E(TS_\alpha, C) = \sum_{i=1}^{\infty} (\psi_i, C\psi_i).$$

In other words, TS_α corresponds to the density matrix $\sum_i |\psi_i)(\psi_i|$.

B. Pure Selective States

Let P_α be a projection measurable in α. We let $\dot{P}_\alpha$ denote the selective state formed by measuring P_α in α and requiring it to have the value one. Let T be some background for which $M(T, P_\alpha) \neq 0$. Let Q_β be any projection $\leq P_\alpha$. The experiment $\{T\dot{P}_\alpha, Q_\beta\}$ is certainly possible. We need only measure the operator $O = 2(P_\alpha - Q_\beta) + 3Q_\beta$, select the events for which $[O] \neq 0$, and count the number of times $[O] = 3$. We require for this experiment the following axiom[21]

Axiom IV:

$$M(T, P_\alpha)M(T\dot{P}_\alpha, Q_\beta) = M(T, Q_\beta),$$

from which follows, setting $Q_\beta = P_\alpha$,

$$M(T\dot{P}_\alpha, P_\alpha) = 1. \tag{4.1}$$

This axiom expresses the compatibility of the observables P_α and Q_β.

Any state that involves only selecting trials ac-

[19] A. L. Licht, Bull. Am. Phys. Soc. 10, 47 (1965).
[20] A. M. Gleason, J. Math. Mech. 6, 885 (1953).
[21] This is essentially the same as in Ref. 9, p. 57, Axiom VI.

cording to the outcome of measurements in α can be expressed as a $\dot{P}_\alpha$ by taking the intersections of appropriate projections. The following theorem gives the structure of such states.

Theorem 1: If T is such that $E(T, P_\alpha) \neq 0$, then

$$E(T\dot{P}_\alpha, A) = E(T, P_\alpha A P_\alpha)(E(T, P_\alpha))^{-1},$$

for all $A \in L(H)$.

Proof: From Axioms III and IV it follows that[22]

$$E(T\dot{P}_\alpha, Q) = E(T, Q)(E(T, P_\alpha))^{-1}, \qquad (4.2)$$

for any projection $Q \in R(\beta)$, $Q \leq P_\alpha$, and β in the future of α. By assumptions (2.1) and (2.7), there exists such a β, β_1, with $R(\beta_1) \supset R(\alpha)$. Then $P_\alpha \in R(\beta_1)$, which implies[23]

$$E(T\dot{P}_\alpha, P_\alpha) = 1. \qquad (4.3)$$

Let $\{\psi_i\}$ be the sequence of vectors corresponding to $T\dot{P}_\alpha$, as in Lemma 3. By Eq. (4.3),

$$\sum_{i=1}^{\infty} (\psi_i, P_\alpha \psi_i) = 1 = E(T\dot{P}_\alpha, 1) = \sum_{i=1}^{\infty} (\psi_i, \psi_i).$$

The norm of P_α is 1; therefore,

$$(\psi_i, P_\alpha \psi_i) \leq (\psi_i, \psi_i).$$

If the inequality held for any i, we would have

$$\sum_{i=1}^{\infty} (\psi_i, P_\alpha \psi_i) < \sum_{i=1}^{\infty} (\psi_i, \psi_i) = 1,$$

a contradiction. Thus $(\psi_i, P_\alpha \psi_i) = (\psi_i, \psi_i)$, which implies that

$$\psi_i = P_\alpha \psi_i, \quad \text{for all } i.$$

Thus, for all $A \in L(H)$,

$$\begin{aligned} E(T\dot{P}_\alpha, A) &= \sum_{i=1}^{\infty} (\psi_i, A\psi_i) \\ &= \sum_{i} (\psi_i, P_\alpha A P_\alpha \psi_i) \\ &= E(T\dot{P}_\alpha, P_\alpha A P_\alpha). \end{aligned}$$

For any $A \in R(\beta)$, β future relative to α, $P_\alpha A P_\alpha$ is the strong limit of sums of projections in $R(\beta)$ and $\leq P_\alpha$. Thus by Eq. (4.2), linearity and continuity,

$$E(T\dot{P}_\alpha, A_\beta) = E(T, P_\alpha A_\beta P_\alpha)(E(T, P_\alpha))^{-1},$$

for all A in such $R(\beta)$. By assumptions (2.5) and (2.7), $L(H)$ is the ultra-weak limit of such $R(\beta)$. The theorem then follows by ultra-weak continuity.

[22] The usual proof of this theorem as given in Ref. 9, p. 57, starts with Eq. (4.2) holding for all $Q \leq P_\alpha$. Here the restrictions placed on the Q_β's require the use of the special properties of the rings $R(\beta)$.

[23] Equation (4.3) does not follow from Axiom III and Eq. (4.1). The region α can causally affect itself, so we cannot assume that $E(TS_\alpha, A_\alpha) = M(TS_\alpha, A_\alpha)$.

C. Uniform Background

Consider an ordinary laboratory experiment in which a state S_α is prepared and an operator A_β measured. It is customary in such an experiment to ensure that there is a region γ including $\alpha \cup \beta$ in which there is nothing extraneous present which might significantly influence the experiment.

If such a region cannot be found, then the presence of the extraneous influences is usually noted and regarded as part of the specification of the state. The state is then not S_α, but some T_γ. Often a correction is applied to the final results to deduce the result that would have been obtained if γ had been empty.

We conclude that the ideal experiment takes place in uniform, empty region. In the following we make the further idealization that this empty region is all of space–time. All departures from emptiness we regard as included in the specification of the state to be prepared.

Let V denote this empty background. It seems reasonable to assume that it is a nonrestrictive background; therefore, we require

Axiom V:

(a) For all operators A_β, β finite, the experiment $\{V, A_\beta\}$ is possible.

(b) For any state S_α, the compound state VS_α is preparable.

Consider the measurement of an operator A_β in the presence of V. If the background is truly uniformly empty, then the expectation value $M(V, A_\beta)$ should not depend on the position of β, on its orientation, or on its state of motion. That is

(c) For any $(a, \Lambda) \in I^\uparrow$

$$M(V, U(a, \Lambda)A_\beta U^\dagger(a, \Lambda)) = M(V, A_\beta).$$

There seems to be no reason to assume that $M(V, A_\beta)$ is not linear. Therefore we assume that it is linear, i.e.,

(d) For all finite β,

$$E(V, A_\beta) = M(V, A_\beta).$$

Theorem 2: For all $A \in L(H)$,

$$E(V, A) = (\Omega, A\Omega).$$

The proof requires the following lemma,[24] which we state without proof.

[24] This is proved on the basis of ring theory in Ref. 6, p. 78, Eq. (12.63), also in Ref. 7, Proposition 4. A field-theoretic proof is given by H. J. Borchers, R. Haag, and B. Schroer, Nuovo Cimento 29, 148 (1963), Theorem IV.

Lemma 4: Let a be any spacelike vector, $\lambda > 0$, ψ a vector in H, and let B be in $R(\alpha)$, where α is finite. Then

$$\lim_{\lambda\to\infty} (\psi, U(\lambda a, 1)BU^{\dagger}(\lambda a, 1)\psi) = (\psi, \psi)(\Omega, B\Omega).$$

Proof of Theorem 2: By Lemma 4, there exist $\psi_i \in H$ such that, for all $A \in L(H)$,

$$E(V, A) = \sum_{i=1}^{\infty} (\psi_i, A\psi_i).$$

In particular, suppose $A \in R(\beta)$, β finite, $E(V, A)$ is assumed to be Lorentz-invariant for such A. Therefore,

$$E(V, A_\beta) = E(V, U(\lambda a, 1)AU^{\dagger}(\lambda a, 1)),$$
$$= \sum_{i=1}^{\infty} (\psi_i, U(\lambda a, 1)AU^{\dagger}(\lambda a, 1)\psi_i),$$

for any spacelike vector a. Let $\epsilon > 0$. Since

$$\sum_{i=1}^{\infty} (\psi_i, \psi_i) = 1,$$

there must be some N such that

$$\sum_{i=N+1}^{\infty} (\psi_i, \psi_i) \leq \frac{\epsilon}{3|A|}.$$

Clearly, for all a,

$$\left|\sum_{i=N+1}^{\infty} (\psi_i, U(a, 1)AU^{\dagger}(a, 1)\psi_i)\right| \leq \tfrac{1}{3}\epsilon.$$

Now

$$|E(V, A) - (\Omega, A\Omega)|$$
$$\leq \sum_{i=N+1}^{\infty} |(\psi_i, U(a, 1)AU^{\dagger}(a, 1)\psi_i)|$$
$$+ |A| \sum_{i=N+1}^{\infty} (\psi_i, \psi_i) + \sum_{i=1}^{N} |(\psi_i, U(a, 1)$$
$$\times AU^{\dagger}(a, 1)\psi_i) - (\psi_i, \psi_i)(\Omega, A\Omega)|.$$

By Lemma 4, we may choose a so large that for $i = 1, 2, \cdots, N$,

$$|(\psi_i, U(a, 1)AU^{\dagger}(a, 1)\psi_i) - (\psi_i, \psi_i)(\Omega, A\Omega)| \leq \epsilon(3N)^{-1}.$$

Thus

$$|E(V, A) - (\Omega, A\Omega)| \leq \epsilon.$$

Since ϵ is arbitrary, we conclude that

$$E(V, A) = (\Omega, A, \Omega) \quad \text{for all} \quad A \in R(\beta),$$

where β is arbitrary, but bounded. From assumption (2.5), we see that each element of $R(X)$ is the ultra-weak limit of such operators. Since $E(V, A)$ is continuous in the ultra-weak topology, we conclude that the theorem holds for all $A \in L(H)$.

D. Nonselective States in the Vacuum

Now, let $\hat{S}_\alpha$ be a nonselective state, and consider the experiment $\{V\hat{S}_\alpha, A_\beta\}$. If β is spacelike relative to α, i.e., $\beta \subset \alpha'$, then β cannot causally affect α, and by Axiom III,

$$E(V\hat{S}_\alpha, A_\beta) = M(V\hat{S}_\alpha, A_\beta).$$

But β also cannot be causally affected by α, thus Axiom II applies, and by Axiom V (d) and Theorem 2, we have

$$E(V\hat{S}_\alpha, A_\beta) = (\Omega, A_\beta\Omega). \tag{4.4}$$

By ultra-weak continuity and assumption (2.5), this equation must also hold for all $A \in R(\alpha')$.

Equation (4.4) states that a nonselective state $\hat{S}_\alpha$ must be strictly localized outside α'.[1,13] A unit vector state ψ that is strictly localized outside α' has been shown[1] to have the form

$$\psi = V\Omega,$$

where $V \in R'(\alpha')$, $V^{\dagger}V = 1$. Here we do not assume that $\hat{S}_\alpha$ corresponds to a vector state. The following theorem gives the structure of these more general states.

Theorem 3: Let $\hat{S}_\alpha$ be as above. Then there exists a sequence $\{A_i \in R'(\alpha'), i = 1, 2, \cdots\}$ such that, for all $B \in L(H)$,

(1) $E(V\hat{S}_\alpha, B) = \sum_{i=1}^{\infty} (\Omega, A_i^{\dagger}BA_i\Omega)$,

(2) $(\Omega, A_i^{\dagger}A_j\Omega) = 0, \quad i \neq j$,

(3) $\sum_{i=1}^{n} A_i^{\dagger}A_i$ (a) ≤ 1 for all n; (b) converges weakly to 1, as $n \to \infty$; (c) converges strongly to 1 as $n \to \infty$.

For the proof, we need the following lemma.[25]

Lemma 5: Let K be a von Neumann algebra. Let $\varphi, \psi \in H$ be such that $(\psi, A\psi) \leq (\varphi, A\varphi)$, for all positive $A \in K$. Then there exists $B \in K'$ such that $\psi = B\varphi$.

Proof: Define strong closure $(K\varphi) = H_1$, and let P_1 denote the projection onto H_1. It can be shown that $P_1 \in K'$.

Define the map $C: K\varphi \to K\psi$ by $C(A\varphi) = A\psi$, for

[25] Suggested in part by H. Araki, private communication (1963).

any $A \in K$. C is clearly linear and commutes with K. It is also bounded on $K\varphi$ as

$$|C(A\varphi)|^2 = |A\psi|^2 = (\psi, A^\dagger A\psi) \leq (\varphi, A^\dagger A\varphi) = |A\varphi|^2.$$

Moreover,

$$C\varphi = \psi.$$

The operator $B' = CP_1$ therefore admits of a closure B, which can be seen to possess the required properties.

Proof of Theorem 3:

Parts (1) and (2): By Lemma 3 there exists a sequence of orthogonal vectors ψ_i such that for all $A \in L(H)$,

$$E(V\hat{S}_\alpha, A) = \sum_{i=1}^{\infty} (\psi_i, A\psi_i).$$

By Eq. (4.4), for all $A \in R(\alpha')$,

$$\sum_{i=1}^{\infty} (\psi_i, A\psi_i) = (\Omega, A\Omega).$$

This implies that for positive $B \in R(\alpha')$,

$$(\psi_i, B\psi_i) \leq (\Omega, B\Omega), \quad \text{for all} \quad i.$$

Lemma 5 then implies that there exists $A_i \in R'(\alpha')$ such that

$$\psi_i = A_i\Omega.$$

Parts (1) and (2) of the theorem follow immediately.

Part (3): Define

$$S_n = \sum_{i=1}^{n} A_i^\dagger A_i.$$

We show first that $|S_n| \leq 1$ for all n, then that S_n converges weakly to one, and finally that S_n converges strongly to one.

Part (3a): $|S_n| \leq 1$. For, since $S_n \geq 0$, there exists some $C_n \geq 0$, $C_n \in R'(\alpha')$, such that

$$S_n = C_n^2 \quad \text{and} \quad |S_n| = |C_n|^2.$$

Let $B \in R(\alpha')$.

$$\begin{aligned} |C_nB\Omega|^2 &= \sum_{i=1}^{n} (\Omega, B^\dagger A_i^\dagger A_i B\Omega) \\ &= \sum_{i=1}^{n} (\Omega, A_i^\dagger B^\dagger B A_i\Omega) \\ &\leq \sum_{i=1}^{\infty} (\Omega, A_i^\dagger B^\dagger B A_i\Omega) \\ &= E(V\hat{S}_\alpha, B^\dagger B) \\ &= (\Omega, B^\dagger B\Omega) = |B\Omega|^2. \end{aligned}$$

Thus, $|C_n|$ is bounded by one on $R(\alpha')\Omega$. But C_n is a bounded operator, and $R(\alpha')\Omega$ is dense in H by Lemma 1. This implies that $|C_n| \leq 1$, and therefore $|S_n| \leq 1$, for all n.

Part (3b): S_n converges weakly to one. For, let B, D be any two operators in $R(\alpha')$,

$$\begin{aligned} (\Omega, B^\dagger S_n D\Omega) &= \sum_{i=1}^{n} (\Omega, B^\dagger A_i^\dagger A_i D\Omega) \\ &= \sum_{i=1}^{n} (\Omega, A_i^\dagger B^\dagger D A_i\Omega) \\ &\to \sum_{i=1}^{\infty} (\Omega, A_i^\dagger B^\dagger D A_i\Omega) \\ &= E(V\hat{S}_\alpha, B^\dagger D) \\ &= (\Omega, B^\dagger D\Omega). \end{aligned}$$

Thus S_n converges weakly to one on $R(\alpha')\Omega$. The sequence $\{S_n\}$ is uniformly bounded, and $R(\alpha')\Omega$ is dense in H. This implies that $S_n \to 1$ weakly.

Part (3c): $S_n \to 1$ strongly. Let $\psi \in H$. Consider

$$\begin{aligned} |(S_n - 1)\psi|^2 &= (\psi, S_n^2\psi) - 2(\psi, S_n\psi) + (\psi, \psi) \\ &\leq |S_n|\,|C_n\psi|^2 - 2(\psi, S_n\psi) + (\psi, \psi), \end{aligned}$$

which, by part (3a), is

$$\begin{aligned} &\leq (\psi, \psi) - (\psi, S_n\psi) \\ &= (\psi, (1 - S_n)\psi), \end{aligned}$$

which goes to zero by part (3b). Thus $S_n \to 1$ strongly.

Example (1): The strictly localized states are an example of the above functionals. There $A_i = \delta_{i1}V$, $V^\dagger V = 1$.

Example (2): Let $\hat{P}_\alpha$ denote the state prepared by measuring, in the presence of the uniform background, the nonzero projection P_α without selecting events according to the outcome. In a series of independent trials in the congruent regions $\{\alpha_i, i = 1, 2, \cdots\}$ the result $[P_{\alpha_i}] = 1$ will occur with probability $(\Omega, P_\alpha\Omega)$, and the result $[P_{\alpha_i}] = 0$ will occur with probability $(\Omega, (1 - P_\alpha)\Omega)$. This state is the independent sum of the states $V\dot{P}_\alpha$ and $V(1 - P_\alpha)^\cdot$ with the respective weights $(\Omega, P_\alpha\Omega)$ and

$$(\Omega, (1 - P_\alpha)\Omega).$$

For all B, we have then

$$\begin{aligned} E(V\dot{P}_\alpha, B) &= (\Omega, P_\alpha\Omega)E(V\dot{P}_\alpha, B) \\ &\quad + (\Omega, (1 - P_\alpha)\Omega)E(V(1 - P_\alpha)^\cdot, B) \\ &= (\Omega, P_\alpha B P_\alpha\Omega) \\ &\quad + (\Omega, (1 - P_\alpha)B(1 - P_\alpha)\Omega), \end{aligned}$$

by Theorem 1. This functional is an example of the above with

$$A_i = \delta_{i1}P_\alpha + \delta_{i2}(1 - P_\alpha).$$

Example (3): Let P_α^i be a set of orthogonal projections in $R(\alpha)$ such that $\sum_i P_\alpha^i = 1$. Let $B_\alpha = \sum_i \lambda_i P_\alpha^i$, for some λ_i, with $\lambda_i \neq \lambda_j$ for $i \neq j$. If the operator B_α is measured, and if the independent trials are not selected according to the outcome of the measurements, then we obtain a certain nonselective state $\hat{B}_\alpha$. Essentially the same argument as in example (2) leads to a representing sequence for this state with

$$A_i = P_\alpha^i, \qquad i = 1, 2, \cdots .$$

E. Arbitrary Backgrounds

From the uniformly empty background V we can obtain other states VS_α. The following theorem shows that any state may be in a certain sense approximated by such a state.

Theorem 4: Let T be any state. Let the region α be given. Then T may be approximated uniformly in $R(\alpha')$ by a local selective state prepared in α in the presence of the vacuum. That is, for any $\epsilon > 0$, there is a selective state preparable in α, T_α^ϵ, such that, for all $A \in R(\alpha')$,

$$|E(VT_\alpha^\epsilon, A) - E(T, A)| \leq \epsilon\,|A|.$$

Proof:

Part (1): Suppose T is a pure vector state ψ. Then

$$E(T, A) = (\psi, A\psi).$$

By Lemma 1, there exists $C \in R(\alpha)$ such that

$$|C\Omega - \psi| < \min\left[\tfrac{1}{2}, \epsilon\left(\tfrac{1}{24}\right)\right].$$

The operator $C^\dagger C$ is Hermitian and positive, and therefore has a spectral resolution,

$$C^\dagger C = \int_0^a \lambda\, dP(\lambda),$$

where $a = |C|^2$ and where the $P(\lambda)$ are in $R(\alpha)$. With ϵ given, there exists a partitioning $\{\Delta_i, i = 1, 2, \cdots, n\}$ of the interval $[0, a]$, such that

$$\left|C^\dagger C - \sum_{i=1}^n \lambda_i P(\Delta_i)\right| < \epsilon\left(\tfrac{1}{24}\right),$$

for some $\lambda_i \in \Delta_i$, $\lambda_i \geq 0$. Let $P^i = P(\Delta_i)$, and define

$$Q = \sum_{i=1}^n \lambda_i P^i.$$

Consider the state

$$\psi_\alpha^\epsilon = \sum_{i=0}^n \oplus\, t_i \dot{P}^i,$$

where

$$t_i = \lambda_i(\Omega, P^i\Omega)((\Omega, Q\Omega))^{-1}.$$

Clearly $t_i \geq 0$, and $\sum_{i=1}^n t_i = 1$. The state ψ_α^ϵ is then an independent superposition of local states, and is thus itself a local state.

Consider now the expectation value, for any $A \in R(\alpha')$,

$$E(V\psi_\alpha^\epsilon, A) = \sum_{i=1}^n t_i E(V\dot{P}_\alpha^i, A)$$

$$= \sum_{i=1}^n \lambda_i(\Omega, P_\alpha^i A P_\alpha^i \Omega)((\Omega, Q\Omega))^{-1}$$

and by Theorem 1,

$$= (\Omega, QA\Omega)((\Omega, Q\Omega))^{-1},$$

since A commutes with all the P_α^i.

Now

$$|E(V\psi_\alpha^\epsilon - A)| - (\psi, A\psi)| \leq \{|(\Omega, (Q - C^\dagger C)A\Omega)| + |(\Omega, C_\natural^\dagger AC\Omega) - (\psi, A\psi)(\Omega, Q\Omega)|\}(\Omega, Q\Omega)^{-1}.$$

After some manipulation this becomes

$$|E(V\psi_\alpha^\epsilon, A) - (\psi, A\psi)| \leq \epsilon\,|A|.$$

Part (2): If T is an arbitrary state, then it has a density matrix $\sum_{i=1}^\infty \mu_i\, |\psi^i)(\psi^i|$, where

$$(\psi^i, \psi^j) = \delta_{ij}, \quad \sum_{i=1}^\infty \mu_i = 1, \quad \text{and} \quad \mu_i \geq 0.$$

Consider the state

$$T_\alpha^\epsilon = \sum_{i=1}^\infty \oplus\, \mu_i \psi_\alpha^{i\epsilon}.$$

This, being the independent sum of local states, must itself be a local state. We have now, for any $A \in R(\alpha')$,

$$|E(VT_\alpha^\epsilon, A) - E(T, A)|$$

$$\leq \sum_{i=1}^\infty \mu_i\, |E(V\psi_\alpha^{i\epsilon}, A) - (\psi^i, A\psi^i)|$$

$$\leq \sum_{i=1}^\infty \mu_i \epsilon\, |A| = \epsilon\,|A|.$$

Remarks: The above construction is a physical one, in the sense that there is a set of physical processes that could in principle be carried out, that would produce VT_α^ϵ. It is not, however, a very practical construction. Loosely speaking, we monitor

the vacuum with an instrument located in α, and wait for a suitable vacuum fluctuation to come along.

The construction depends on the fact that vacuum fluctuations are spacelike correlated. We interpret these correlations as being produced by vacuum fluctuations occurring in the past of α. The construction does not therefore imply the existence of a direct causal link between α and α'.

The states VT_α^ϵ approach T only in α'. They look very different from T as seen in α, or from any region causally dependent on α. Indeed, as $\epsilon \to 0$, they may not even converge in these regions to any state.

We conclude that the objections to Eq. (1.1) do not apply to this construction.

The state VT_α^ϵ could itself serve as the background for the preparation of some other state, S_β, say, with $\beta \subset \alpha'$. By varying ϵ, we can make VT_α^ϵ approach T as closely as we like uniformly in $R(\beta)$. We would like to conclude that $VT_\alpha^\epsilon S_\beta$ must also approach TS_β uniformly in $R(\beta)$. We cannot, however, do this without making a special assumption about the continuity of state composition. Therefore we assume the following.

Axiom VI: Suppose the sequence of states $\{T^i, i = 1, 2, \cdots\}$ approach a state T uniformly in $R(\beta)$. Let S_β be any local state such that $\{TS_\beta, T^iS_\beta, i = 1, 2, \cdots\}$ are possible states. Then we require that the sequence T^iS_β also approaches TS_β uniformly in $R(\beta)$.

F. Nonselective States in an Arbitrary Background

Suppose that a nonselective state $\hat{S}_\alpha$ is prepared in the presence of the background T. A sequence of operators $\{A_i \in R'(\alpha'), i = 1, 2, \cdots\}$ such that, for all $B \in L(H)$,

$$E(T\hat{S}_\alpha, B) = \sum_{i=1}^{\infty} E(T, A_i^\dagger B A_i)$$

we call a representing sequence for $\hat{S}_\alpha$ relative to T. According to Theorem 3, every $\hat{S}_\alpha$ has a representing sequence relative to the vacuum. We show below that this same sequence will also represent $\hat{S}_\alpha$ relative to any background.

We need the following lemma.

Lemma 6: Let $\hat{S}_\alpha$ be represented by $\{A_i \in R'(\alpha'), i = 1, 2, \cdots\}$ relative to V.

(1) There is a linear map of $L(H)$ into $L(H)$, which we denote by $\hat{S}_\alpha(B)$, defined for all $B \in L(H)$ by

$$\hat{S}_\alpha(B) = \text{strong } \lim_{n\to\infty} \sum_{i=1}^{n} A_i^\dagger B A_i.$$

It has the further properties:

(2) $$|S_\alpha(B)| \leq |B|,$$

(3) $$E(V\hat{S}_\alpha, B) = (\Omega, \hat{S}_\alpha(B)\Omega).$$

(4) The map $\hat{S}_\alpha(B)$ is independent of the particular representing sequence chosen for $\hat{S}_\alpha$.

(5) For any background T, and any $B \in L(H)$,

$$\sum_{i=1}^{\infty} E(T, A_i^\dagger B A_i) = E(T, \hat{S}_\alpha(B)).$$

The proof is given in Appendix B. We call this the representing map for $\hat{S}_\alpha$ relative to V. Note that it is in general not an automorphism.

We also need the following axiom, expressing the independence of states prepared in spacelike separated regions in the presence of the vacuum.

Axiom VII: Let α and β be two spacelike separated regions. Then for any two states, S_α^1, S_β^2, the compound state $VS_\alpha^1 S_\beta^2$ is always possible. Moreover, it should make no difference if we regard VS_α^1 as background to S_β^2, or VS_β^2 as background to S_α^1.

Theorem 5: Let $\hat{S}_\alpha$ be a nonselective state. For any background T, and any $B \in L(H)$,

$$E(T\hat{S}_\alpha, B) = E(T, \hat{S}_\alpha(B)),$$

where $\hat{S}_\alpha(B)$ is the map given in Lemma 5.

Proof: Suppose first $B \in R(\beta)$, β finite. Take some $\gamma \subset (\alpha \cup \beta)'$. By Theorem 4 and Axioms VI and VII, there exists a selective state T_γ^ϵ such that

$$|E(T\hat{S}_\alpha, B_\beta) - E(VT_\gamma^\epsilon \hat{S}_\alpha, B_\beta)| \to 0$$

as $\epsilon \to 0$. Moreover,

$$T_\gamma^\epsilon = \sum_i \oplus \lambda_i \dot{P}_\gamma^i,$$

for certain projections $P_\gamma^i \in R(\gamma)$.

Consider the state $V\dot{P}_\gamma^i\hat{S}_\alpha$. According to Axiom VII, this is a possible state. Also,

$$V\dot{P}_\gamma^i\hat{S}_\alpha = (V\dot{P}_\gamma^i)\hat{S}_\alpha = (V\hat{S}_\alpha)\dot{P}_\gamma^i.$$

By Theorem 1 we then have

$$\begin{aligned} &E(V\dot{P}_\gamma^i\hat{S}_\alpha, B_\beta) \\ &\quad = E((V\hat{S}_\alpha)\dot{P}_\gamma^i, B_\beta) \\ &\quad = E(V\hat{S}_\alpha, P_\gamma^i B_\beta P_\gamma^i)(E(V\hat{S}_\alpha, P_\gamma^i))^{-1}, \end{aligned}$$

and by Theorem 3,

$$= \sum_j E(V, A_j^\dagger P_\gamma^i B_\beta P_\gamma^i A_j)(E(V\hat{S}_\alpha, P_\gamma^i))^{-1}.$$

Since $\gamma \subset \alpha'$, we then have, using Lemma 6,

$$= E(V, P_\gamma^i \hat{S}_\alpha(B_\beta) P_\gamma^i)(E(V, P_\gamma^i))^{-1},$$

and by Theorem 1,

$$= E(VP_\gamma^i, \hat{S}_\alpha(B_\beta)).$$

Thus,

$$E(VT_\gamma^\epsilon \hat{S}_\alpha, B_\beta) = E(VT_\gamma^\epsilon, \hat{S}_\alpha(B_\beta)),$$

and

$$|E(T, \hat{S}_\alpha(B_\beta)) - E(VT_\gamma^\epsilon, \hat{S}_\alpha(B_\beta))|$$
$$\leq \epsilon\, |\hat{S}_\alpha(B_\beta)|, \quad \text{by Theorem 4,}$$
$$\leq \epsilon\, |B_\beta|, \quad \text{by Lemma 6.}$$

Now,

$$\begin{aligned}
&|E(T\hat{S}_\alpha, B_\beta) - E(T, \hat{S}_\alpha(B_\beta)| \\
&\quad \leq |E(T\hat{S}_\alpha, B_\beta) - E(VT_\gamma^\epsilon \hat{S}_\alpha, B_\beta)| \\
&\qquad + |E(VT_\gamma^\epsilon \hat{S}_\alpha, B_\beta) - E(VT_\gamma^\epsilon, \hat{S}_\alpha(B_\beta))| \\
&\qquad + |E(VT_\gamma^\epsilon, \hat{S}_\alpha(B_\beta)) - E(T, \hat{S}_\alpha(B_\beta))| \\
&\quad \leq \tfrac{1}{2}|E(T\hat{S}_\alpha, B_\beta) - E(VT_\gamma^\epsilon \hat{S}_\alpha, B_\beta)| + 0 + \epsilon\, |B_\beta|.
\end{aligned}$$

This goes to zero as $\epsilon \to 0$, ϵ is arbitrary, therefore

$$E(T\hat{S}_\alpha, B_\beta) = E(T, \hat{S}_\alpha(B_\beta)). \tag{4.5}$$

The functional $E(T\hat{S}_\alpha, B_\beta)$ is continuous in the ultra-weak topology, by Lemma 2. Let $\sum_i |\psi_i)(\psi_i|$ be the density matrix corresponding to the state T. Then

$$E(T, \hat{S}_\alpha(B_\beta)) = \sum_{i,j} (\psi_i, A_j^\dagger B_\beta A_j \psi_i).$$

Clearly $\sum_{i,j} (A_j\psi_i, A_j\psi_i) = 1$, therefore $E(T, \hat{S}_\alpha(B_\beta)$ is also ultra-weak-continuous on B_β. By assumption (2.5) then, both sides of Eq. (4.5) extend to all $L(H)$, and the theorem follows.

5. DISCUSSION

We have seen that every physical state consists of a nonlocal part, the background, which is not completely under the control of the experimenter, and a local part, that is, the local state.

There are basically two types of local states, pure selective and pure nonselective. There are also other types, obtainable from the basic two by the various processes of state composition.

The local properties of these states are very different. The pure selective states are sensitive to spacelike correlations set up in the background state by past events. The pure nonselective states are not. The composite states will therefore have a composite type of locality.

The mathematical structure of the pure selective states is familiar from ordinary quantum mechanics. They are simply the states "projected out" from the background density matrix by a local projection.

The structure of those pure nonselective states which are prepared by a nonselective measurement is also familiar. These states are discussed above in Example (3) following Theorem 3. They have density matrices of the form

$$M = \sum_i P_\alpha^i T P_\alpha^i,$$

where the P_α^i are as in Example (3), and T is the density matrix corresponding to the background state.

However, here we consider that a nonselective state can be prepared by any physical process that does not involve a selection. There certainly are many such processes which also do not involve a measurement. For such states, the structure given by Theorems 3 and 5 and Lemma 6 seems to be new.

To obtain these results we have had to use certain axioms.

We postulate, in Axiom I, the existence of that average value of an operator, the M function, which is determinable by experiment. In Axiom II, we require that the M function have certain locality properties. A distinction is made in Axiom III between this experimental average value and the mathematical expectation value, the E function. This distinction is not made in ordinary quantum mechanics. However, it must be made here as it depends on the relative locations of the regions involved in the experiments.

Axiom IV concerns the expectation values of compatible observables and is also used in ordinary quantum mechanics.

It may not be possible to prepare certain local states in the presence of some backgrounds. We therefore find it convenient in Axiom V to postulate the existence of an ideally empty and nonrestrictive background, the vacuum state. In Axiom VI we assume that states have a certain continuity property. The notion of the nonrestrictivity of the vacuum is strengthened in Axiom VII.

These axioms have been found to be very useful in the above analysis. It is possible that these axioms are not independent, and a smaller set might do just as well. It is also possible that some of these axioms are over idealized. (This may be particularly true of those concerning the vacuum state.) Never-

theless, it is felt that these axioms are sufficiently close to actuality as to still be very useful in the future.

ACKNOWLEDGMENTS

The author wishes to express his thanks to Professor H. Araki for several very beneficial suggestions, and to Dr. D. Currie, Dr. E. J. Woods, and Professor O. W. Greenberg for many stimulating criticisms. He also wishes to thank Professor J. R. Oppenheimer for the hospitality extended to him at the Institute for Advanced Study during the revision of this manuscript.

APPENDIX

A. The M Function

In Sec. (3c) it was suggested that $M(S_\alpha, A_\beta)$ may not be a linear function of A_β if S_α is selective and if also β can causally affect α. We prove here that this is so for the case when α, β are disjoint, A_β is a projection Q_β, and $S_\alpha = V\dot{P}_\alpha$.

The function $M(V\dot{P}_\alpha, Q_\beta)$ is obtained as an average over the results $[P_{\alpha i}]$, $[Q_{\beta i}]$ of measurements made in the independent trials $\{(V\dot{P}_{\alpha i}, Q_{\beta i}),\ i = 1, 2, \cdots\}$. In fact

$$M(V\dot{P}_\alpha, Q_\beta) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} [P_{\alpha i}][Q_{\beta i}]\left\{\frac{1}{N}\sum_{i=1}^{N}[P_{\alpha i}]\right\}^{-1}. \tag{A1}$$

The limit of the denominator is, however, the expectation value of P_α in a state in which Q_β has been measured but not selected. That is,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} [P_{\alpha i}] = M(V\hat{Q}_\beta, P_\alpha).$$

Since β cannot be causally affected by α, Axiom III applies and

$$\begin{aligned} M(V\hat{Q}_\beta, P_\alpha) &= E(VQ_\beta, P_\alpha) \\ &= (\Omega, Q_\beta P_\alpha Q_\beta \Omega) \\ &\quad + (\Omega, (1 - Q_\beta)P_\alpha(1 - Q_\beta)\Omega), \end{aligned} \tag{A2}$$

by Eq. (4.5).

From the results of the same sequence of trials, one can also calculate

$$M(V\dot{Q}_\beta, P_\alpha) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} [Q_{\beta i}][P_{\alpha i}]\left(\frac{1}{N}\sum_{i=1}^{N}[Q_{\beta i}]\right)^{-1}, \tag{A3}$$

$$= E(V\dot{Q}_\beta, P_\alpha) \tag{A4}$$

$$= (\Omega, Q_\beta P_\alpha Q_\beta \Omega)\ |Q_\beta\Omega|^{-2},$$

by Axiom III and Theorem 1. However, the limit of the denominator $\sum[Q_{\beta i}]/N$ in Eq. (A3) is $M(V\dot{P}_\alpha, Q_\beta)$, and by Axiom II and Theorem 2, it must be $(\Omega, Q_\beta\Omega)$. Thus from Eq. (A3) and Eq. (A4) we conclude that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} [Q_{\beta i}][P_{\alpha i}] = (\Omega, Q_\beta P_\alpha Q_\beta \Omega). \tag{A5}$$

Equations (A1), (A2), and (A5), however, imply that

$$M(V\dot{P}_\alpha, Q_\beta) = (\Omega, Q_\beta P_\alpha Q_\beta \Omega)[(\Omega, Q_\beta P_\alpha Q_\beta \Omega) + (\Omega, (1 - Q_\beta)P_\alpha(1 - Q_\beta)\Omega)]^{-1},$$

which is clearly nonlinear in Q_β.[26]

Note that by Theorem 1

$$E(V\dot{P}_\alpha, Q_\beta) = (\Omega, P_\alpha Q_\beta P_\alpha \Omega)\ |P_\alpha\Omega|^{-2},$$

even if β can causally affect α. Thus for such β, and if P_α and Q_β do not commute,

$$E(V\dot{P}_\alpha, Q_\beta) \neq M(V\dot{P}_\alpha, Q_\beta).$$

It can be seen that we do have equality if P_α and Q_β commute.

B. Proof of Lemma 5

Part (1): For any $B \in L(H)$, denote

$$W_n(B) = \sum_{i=1}^{n} A_i^\dagger B A_i.$$

Let D be any positive operator in $L(H)$. Clearly

$$W_{n+m}(D) \geq W_n(D) \geq 0, \quad \text{for all} \quad n, m.$$

This implies that the operator

$$C_n^m(D) = [W_{n+m}(D) - W_n(D)]^{\frac{1}{2}}$$

exists.

For any $\psi \in H$,

$$\begin{aligned} (\psi, W_n(D)\psi) &\leq \sum_{i=1}^{\infty} (\psi, A_i^\dagger D A_i \psi), \\ &\leq |D| \sum_{i=1}^{\infty} (\psi, A_i^\dagger A_i \psi), \\ &= |D|\ |\psi|_n^2. \end{aligned}$$

by Part (3) of Theorem 3. Thus

$$|W_n(D)| \leq |D|.$$

The sequence of positive numbers $\{(\psi, W_n(D)\psi),\ n = 1, 2, \cdots\}$ is monotonically increasing, and bounded by $|D|\ |\psi|^2$. It therefore converges to some

[26] A similar result has been obtained by S. Watanabe, Rev. Mod. Phys. 27, 179 (1955).

limit. This implies that given $\epsilon > 0$, there exists $N(\epsilon)$ such that for all $n > N(\epsilon)$, and for all m,

$$(\psi, (W_{n+m}(D) - W_n(D))\psi) \leq \epsilon(2\,|D|)^{-1}.$$

This is the same as

$$|C_n^m(D)\psi|^2 \leq \epsilon(2\,|D|)^{-1}.$$

Now consider, for arbitrary $\psi \in H$,

$$\begin{aligned}|(W_{n+m}(D) - W_n(D))\psi|^2 &= (\psi, C_n^m(D)(W_{n+m}(D) - W_n(D))C_n^m(D)\psi)\\ &\leq (|W_{n+m}(D)| + |W_n(D)|)\,|C_n^m(D)\psi|^2\\ &\leq 2\,|D|\,|C_n^m(D)\psi|^2.\end{aligned}$$

By the above, n can be taken so large that this is smaller than any preassigned ϵ. We conclude that $\{W_n(D), n = 1, 2, \cdots\}$ is a strong Cauchy sequence, hence possesses a strong limit

$$\hat{S}_\alpha(D) = \text{strong}\lim_{n\to\infty} W_n(D).$$

Any $B \in L(H)$ can be written as the sum of four positive elements, therefore Part (1) holds.

Part (2): Let $\varphi, \psi \in H$, $|\varphi| = |\psi| = 1$. Let $B \in L(H)$. By part (1), $\hat{S}_\alpha(B)$ exists and

$$\begin{aligned}|(\varphi, \hat{S}_\alpha(B)\psi)| &= \left|\sum_{i=1}^{\infty} (\varphi, A_i^\dagger B A_i \psi)\right|\\ &\leq |B| \sum_{i=1}^{\infty} |A_i\varphi|\,|A_i\psi|\\ &\leq |B| \left[\sum_{i=1}^{\infty} |A_i\varphi|^2\right]^{\frac{1}{2}} \left[\sum_{j=1}^{\infty} |A_j\psi|^2\right]^{\frac{1}{2}}\\ &= |B|, \quad \text{by Theorem 3, Part (3).}\end{aligned}$$

In particular, let $\varphi = \hat{S}_\alpha(B)\,|\hat{S}_\alpha(B)\psi|^{-1}$. This yields

$$|\hat{S}_\alpha(B)| \leq |B|.$$

Part (3): Proof immediate.

Part (4): Suppose that S_α is represented relative to V by both

$$\{A_i', i = 1, 2, \cdots\} \quad \text{and} \quad \{A_i'', i = 1, 2, \cdots\}.$$

By Part (1),

$$\hat{S}_\alpha'(B) = \text{strong}\lim_{n\to\infty} \sum_{i=1}^{n} A_i'^\dagger B A_i',$$

$$\hat{S}_\alpha''(B) = \text{strong}\lim_{n\to\infty} \sum_{i=1}^{n} A_i''^\dagger B A_i'';$$

both exist for any $B \in L(H)$, and

$$E(V\hat{S}_\alpha, B) = (\Omega, \hat{S}_\alpha'(B)\Omega) = (\Omega, \hat{S}_\alpha''(B)\Omega).$$

In particular, let $B = D_1^\dagger C D_2$, where D_1, D_2 are arbitrary operators in $R(\alpha')$, and C is an arbitrary element of $L(H)$. Then clearly

$$(\Omega, \hat{S}_\alpha'(D_1^\dagger C D_2)\Omega) = (D_1\Omega, \hat{S}_\alpha'(C)D_2\Omega),$$

$$(\Omega, \hat{S}_\alpha''(D_1^\dagger C D_2)\Omega) = (D_1\Omega, \hat{S}_\alpha''(C)D_2\Omega),$$

which implies that

$$(D_1\Omega, \hat{S}_\alpha'(C)D_2\Omega) = (D_1\Omega, \hat{S}_\alpha''(C)D_2\Omega).$$

By Lemma 1 we conclude that

$$\hat{S}_\alpha'(C) = \hat{S}_\alpha''(C), \quad \text{for all} \quad C \in L(H).$$

Part (5): Let $\sum_{i=1}^{\infty} |\psi_i)(\psi_i| \lambda_i$, $\lambda_i \geq 0$, $\sum_{i=1}^{\infty} \lambda_i = 1$, be the density matrix corresponding to the background state T. We must prove only that

$$\lim_{n\to\infty} E(T, W_n(B)) = E(T, \hat{S}_\alpha(B)).$$

Now E is continuous on the ultra-weak topology. The operators $W_n(B) \to \hat{S}_\alpha(B)$ in the strong topology and are uniformly bounded. Any strongly convergent, and uniformly bounded sequence is also an ultra-weakly convergent sequence, from which the theorem follows.

Absorptive Parts and the Bethe–Salpeter Equation for Forward Scattering*

S. NUSSINOV† AND J. ROSNER
University of Washington, Seattle, Washington
(Received 15 February 1966)

We have developed a Laplace transform approach to the Bethe–Salpeter equation for absorptive parts of forward scattering amplitudes. The method appears direct and unsophisticated and is useful for computation. It is essentially identical to decomposition into four-dimensional partial waves, but the inversion formula is more straightforward. We have obtained the high-energy behavior of sums of several types of ϕ^4 graphs in the weak- and strong-coupling limits. These examples illustrate some general results we obtain for Mth order ϕ^4 kernels. Specifically, the absorptive part behaves as $s^{n_0}(\log s)^\beta$ for high s; for weak coupling λ, $n_0 \sim \lambda^{1M}$, while for strong coupling in the ladder graph approximation, $n_0/\lambda^{\frac{1}{2}} \to 1/2\pi$. We have also proven an interesting inequality related to absorptive parts. One of its corollaries is that uncrossed ladder graphs "majorize" crossed ones.

I. INTRODUCTION

THE Bethe-Salpeter equation for forward scattering is often decomposed using four-dimensional partial waves[1-4] and a Euclidean formalism.[5] An alternative approach deals with the Bethe–Salpeter equation for absorptive parts.[6-9] The high-energy behavior of the solution is usually the desired end result in both cases. It is obtained in the partial-wave approach from the singularities in the n plane, and in the absorptive part approach by direct but often somewhat cumbersome means. The absorptive part approach is, however, appealing for the following reasons: (a) It deals with real quantities. (b) It often circumvents divergences by placing internal lines on the mass shell. (c) It does not entail passing to and from a Euclidean metric, thus avoiding the temporary inconvenience of a bounded center-of-mass energy. (d) It avoids singularities in propagators.

In this paper we discuss and utilize an approach to the Bethe–Salpeter equation for absorptive parts which is based on the Laplace transform. We obtain an equation similar to the four-dimensional partial-wave Bethe–Salpeter equation, but in a continuous variable n. The operation of inverse Laplace transformation replaces integration along the deformed Sommerfeld–Watson contour. We obtain the high-energy behavior of the absorptive part from the leading singularity in the n plane.

We have summed several types of graphs in ϕ^4 theory, including one set for which the kernel is of fifth order in λ, using a generalization of the Bethe–Salpeter equation.[4] We have also obtained some interesting results in weak- and strong-coupling limits for a wide class of ϕ^4 kernels. The specific examples illustrate these general results.

We also prove an inequality for absorptive parts based on the Schwarz inequality. One of its consequences, roughly speaking, is that uncrossed ladder graphs "majorize" crossed ones.

Section II is devoted to a discussion of the Laplace transform method. In Sec. III we present specific examples of summed graphs, and in Sec. IV we obtain some general results which these examples illustrate. The above-mentioned inequality is proved in Sec. V.

Our metric is $(-1, 1, 1, 1)$.

II. ABSORPTIVE PART EQUATIONS AND THE LAPLACE TRANSFORM METHOD

We write the Bethe–Salpeter equation symbolically as

$$A = B + KA. \tag{1}$$

For forward scattering of particles of momenta p_1 and p_2, A and B have cuts for $s = -(p_1 + p_2)^2 \geq 0$ and/or $u = -(p_1 - p_2)^2 \geq 0$. (We assume that p_1^2, p_2^2, and the internal masses are such that no anomalous thresholds occur.) We write the imaginary (or absorptive) part of (say) A evaluated above the s cut as A_s, and above the u cut as A_u. One calculates A_s by "cutting" the graph in all possible ways[10]

* Supported in part by the U. S. Atomic Energy Commission.
† Present address: Palmer Physical Laboratory, Princeton University, Princeton, New Jersey.
[1] J. D. Bjorken, J. Math. Phys. 5, 192 (1964).
[2] M. Baker and I. J. Muzinich, Phys. Rev. 132, 2291 (1963).
[3] M. K. Banerjee, M. Kugler, C. A. Levinson, and I. J. Muzinich, Phys. Rev. 137, B1280 (1965).
[4] S. Nussinov, Phys. Rev. 140, B1367 (1965).
[5] G. Tiktopoulos, Phys. Rev. 136, B275 (1964).
[6] D. Amati, A. Stanghellini, and S. Fubini, Nuovo Cimento 26, 896 (1962).
[7] S. Treiman and G. Tiktopoulos, Phys. Rev. 135, B711 (1964).
[8] S. Treiman and G. Tiktopoulos, Phys. Rev. 136, B1217 (1964).
[9] S. Treiman and G. Tiktopoulos, Phys. Rev. 137, B1597 (1965).
[10] R. E. Cutkosky, J. Math. Phys. 1, 429 (1960).

so as to separate it into two pieces, one containing the initial two lines and the other containing the final two lines. A_u is calculated similarly except that the two pieces each contain one initial and one final line.

The above prescription allows us to write coupled equations for A_s and A_u based on Eq. (1):

$$\begin{aligned} A_s &= B_s + (KA)_s \\ &= B_s + K_sA_s + K_uA_u, \\ A_u &= B_u + (KA)_u \\ &= B_u + K_sA_u + K_uA_s. \end{aligned} \tag{2}$$

These may be uncoupled using the definitions

$$A^+ = A_s + A_u, \quad A^- = A_s - A_u, \quad \text{etc.} \tag{3}$$

to give

$$A^\pm = B^\pm + K^\pm A^\pm,$$

or, more explicitly,

$$A^\pm(p_1, p_2) = B^\pm(p_1, p_2) + \frac{1}{8\pi^4}\int \frac{d^4k}{(k^2+m^2)^2} K^\pm(p_1, k)A^\pm(k, p_2). \tag{4}$$

Let p_1 and p_2 be spacelike. (We discuss later the case that one or both is timelike.) Then both $K^\pm$ and $A^\pm$ vanish for $s < 0$, $u < 0$. This fact then implies that in (4) k ranges over spacelike values only. (See, e.g., Ref. 7.) Moreover, the above threshold conditions allow us to make the following definitions for real ϕ, ϕ_1, and ϕ_2:

$$\begin{aligned} \cosh\phi &\equiv |p_1\cdot p_2|/[(p_1^2)^{\frac{1}{2}}(p_2^2)^{\frac{1}{2}}], \\ \cosh\phi_1 &\equiv |p_1\cdot k|/[(p_1^2)^{\frac{1}{2}}(k^2)^{\frac{1}{2}}], \\ \cosh\phi_2 &\equiv |p_2\cdot k|/[(p_2^2)^{\frac{1}{2}}(k^2)^{\frac{1}{2}}]. \end{aligned} \tag{5}$$

One can then express $A^\pm$, $B^\pm$, and $K^\pm$ as functions of two invariant masses and the invariant angles ϕ. The integration over k in (4) can be transformed into one over k^2, ϕ_1, and ϕ_2 by standard means.[11] One obtains

$$d^4k\, K^\pm(p_1, k)A^\pm(k, p_2) = \frac{\pi\theta(-J)\sinh\phi_1\sinh\phi_2 k^2}{\sinh\phi}\, d\phi_1\, d\phi_2\, dk^2 \times K^\pm(\phi_1, p_1^2, k^2)A^\pm(\phi_2, k^2, p_2^2), \tag{6}$$

where

$$J = \begin{vmatrix} k^2 & k\cdot p_1 & k\cdot p_2 \\ k\cdot p_1 & p_1^2 & p_1\cdot p_2 \\ k\cdot p_2 & p_1\cdot p_2 & p_2^2 \end{vmatrix}$$
$$= k^2p_1^2p_2^2(1 - \cosh^2\phi - \cosh^2\phi_1 - \cosh^2\phi_2 + 2\cosh\phi\cosh\phi_1\cosh\phi_2). \tag{7}$$

The condition $J < 0$, together with conservation of energy to tell us the appropriate choice of root, then implies

$$\phi \geq \phi_1 + \phi_2. \tag{8}$$

This condition is just one of reality of the scattering angle in the center-of-mass frame. The absorptive part equation may then be written as

$$A^\pm(\phi, p_1^2, p_2^2) = B^\pm(\phi, p_1^2, p_2^2) + \frac{1}{8\pi^3\sinh\phi} \times \int \frac{k^2\, d(k^2)}{(k^2+m^2)^2}\,\theta(\phi - \phi_{10} - \phi_{20}) \times \int_{\phi_{1*}}^{\phi-\phi_{2*}} \sinh\phi_1\, d\phi_1 \int_{\phi_{2*}}^{\phi-\phi_1} \sinh\phi_2\, d\phi_2 \times K^\pm(\phi_1, p_1^2, k^2)A^\pm(\phi_2, k^2, p_2^2), \tag{9}$$

where the limits on ϕ_1 and ϕ_2 are defined by (8) and the thresholds of $K^\pm$ and $A^\pm$. Note that these thresholds are functions of momenta. They are related to the usual thresholds s_0, u_0 by

$$\phi_0 = \cosh^{-1}\left\{\frac{\min(s_0, u_0) + p_1^2 + p_2^2}{2(p_1^2)^{\frac{1}{2}}(p_2^2)^{\frac{1}{2}}}\right\}. \tag{10}$$

The following transform simplifies (9) considerably. We define

$$F_n(p_1^2, p_2^2) \equiv \frac{2}{\pi}\int_0^\infty d\phi \sinh\phi e^{-(n+1)\phi} \times F(\phi, p_1^2, p_2^2), \tag{11}$$

where F denotes $A^\pm$, $B^\pm$, or $K^\pm$. Applying this transform to (9), rearranging the order of integration in the KA term, and performing the trivial ϕ integration, one obtains

$$A_n^\pm(p_1^2, p_2^2) = B_n^\pm(p_1^2, p_2^2) + \frac{1}{16\pi^2(n+1)}\int_0^\infty \frac{k^2\, d(k^2)}{(k^2+m^2)^2} \times K_n^\pm(p_1^2, k^2)A_n^\pm(k^2, p_2^2). \tag{12}$$

Expression (11) is just the one given in Ref. 3 for the partial-wave amplitude and was used there to continue to complex n. We stress that in our approach n is complex from the start. Furthermore, we

[11] See, e.g., T. W. B. Kibble, Phys. Rev. 117, 1159 (1960).

see that the Euclidean approach needed for the partial-wave decomposition is not required here; by working with absorptive parts we can arrive at (11) more directly.

Equation (12) is the same as the partial-wave Bethe–Salpeter equation, the nature of whose solutions is described extensively in Refs. 1 and 3. For a ϕ^3 theory, the kernel is Fredholm and the right-most singularities in the n plane are (in general) simple poles. These poles may be obtained by solving the corresponding homogeneous equation.[6,12] For a ϕ^4 theory the singularity of the kernel for large k leads to cuts in the n plane.[2]

Equation (11) is a Laplace transform. It requires for its existence in some region Re $n \geq N$ only the fact that F obeys a threshold condition and the condition

$$F(\phi, p_1^2, p_2^2) \leq e^{N\phi} G(p_1^2, p_2^2), \tag{13}$$

which is equivalent to F being polynomially bounded in s and u. We shall see that such bounds can be assumed to exist. Then (11) may be inverted to give

$$(2/\pi) F(\phi, p_1^2, p_2^2) e^{-\phi} \sinh \phi = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} dn\, e^{n\phi} F_n(p_1^2, p_2^2), \tag{14}$$

where $c \geq N$. The contour in (14) is the same as the Sommerfeld–Watson contour after it has been opened up parallel to the imaginary axis. Note, however, that the integrand is of somewhat simpler form than in the Sommerfeld–Watson representation.

Suppose the Born term in (9) behaves when $\phi \to \infty$ as $e^{N\phi}\phi^\beta G(p_1^2, p_2^2)$ for some N and β. Then the right-most singularity of $B_n^\pm$ will be at $n = N$. For very small coupling, the right-most singularity of $A_n^\pm$ will also be very near $n = N$.[1,13] Then we may represent $A^\pm(\phi, p_1^2, p_2^2)$ in the form (14), with $c = N + \epsilon$. As we let the coupling constant increase, the right-most singularity of $A_n^\pm$ will move to the right and we must deform the contour in (14) to keep to the right of it. The contribution to $A^\pm(\phi, p_1^2, p_2^2)$ of this right-most singularity is easily evaluated as in Refs. 1, 2, and 3. An isolated pole at n_0 gives rise to a behavior

$$A^\pm(\phi, p_1^2, p_2^2) \underset{(\phi\to\infty)}{\longrightarrow} e^{n_0\phi}, \tag{15}$$

which is the same as a behavior s^{n_0} for high s, while a cut ending at n_0 gives rise to an additional factor of ϕ^β in (15), corresponding to an additional factor $(\log s)^\beta$. If the cut is of the form $A_n \approx \gamma/(n - n_0)^{\frac{1}{2}}$, for example, we find that $\beta = -\frac{1}{2}$.[2,3]

To extend the method described above to cases for which one or both of the external vectors are timelike, we simply note that, if certain mass inequalities are met, the representation (5) is still valid provided one takes the square root of positive quantities. For example, if q is timelike and k is spacelike,

$$2q\cdot k = -(q-k)^2 + q^2 + k^2 \geq 2(-q^2)^{\frac{1}{2}}(k^2)^{\frac{1}{2}},$$

$$\text{if} \qquad -(q-k)^2 + q^2 \geq -q^2,$$

$$\text{or} \qquad -(q-k)^2 \geq -2q^2. \tag{16}$$

The general discussion of continuation from spacelike to timelike momenta has been given by Tiktopoulos[5] and is illustrated by Tiktopoulos and Treiman.[7]

III. SOME SPECIFIC SUMS OF GRAPHS

The method of the previous section has actual computational value. In Ref. 3 a general discussion of the Bethe–Salpeter equation was given for ϕ^4 kernels. It is interesting to note for what kernels can in practice obtain the leading singularity n_0, and hence the high-energy behavior.

In what follows we restrict ourselves to spacelike p_1 and p_2. All internal masses are taken to be zero. We consider ϕ^4 kernels which can be renormalized by a single subtraction. This corresponds to assuming that their absorptive parts are bounded by a constant as $\phi \to \infty$, so that the expression (11) for $K_n(p_1^2, p_2^2)$ converges for $n > 0$. We treat only crossing-symmetric kernels (or ones which we symmetrize artificially) so that we may neglect the equation for A^-, and thus drop the superscript (+) in the remainder of this section.

An equation similar to the Bethe–Salpeter equation may be obtained by introducing two-particle intermediate states artificially.[4] (see Fig. 1.) The dotted lines are given momentum labels, but they

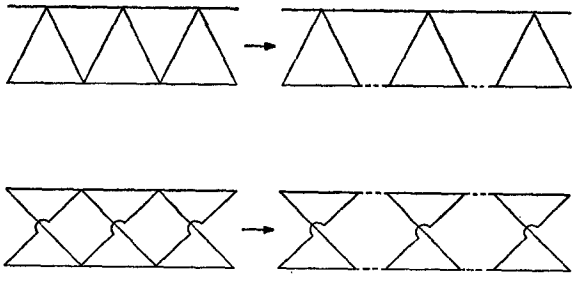


FIG. 1. Artificial introduction of two-particle intermediate states.

[12] J. Rosner, J. Math. Phys. 7, 875 (1966).
[13] M. Kugler and S. Nussinov (unpublished).

do not have propagators associated with them. One then obtains a generalization of the Bethe–Salpeter equation by replacing the $(k^2 + m^2)^{-2}$ factor in (12) by $(k^2 + m^2)^{-1}$ for the case illustrated in Fig. 1(a), or by 1 for Fig. 1(b).

If one lets all internal particles of K be massless, K can be only a function of p_1^2/p_2^2 times an over-all factor $(p_1^2 p_2^2)^{-\alpha}$. For the true Bethe–Salpeter equation $\alpha = 0$, for Fig. 1(a) $\alpha = \frac{1}{2}$, and for Fig. 1(b) $\alpha = 1$. We define the dimensionless functions $\tilde{A}_n$, $\tilde{B}_n$, and $\tilde{K}_n$ by the general formula

$$\tilde{F}_n \equiv (p_1^2 p_2^2)^{\alpha} F_n \tag{17}$$

so that $\tilde{A}_n$ satisfies

$$\tilde{A}_n(p_1, p_2) = \tilde{B}_n(p_1, p_2) + \frac{1}{8\pi^2(n+1)} \int_0^\infty \frac{dk}{k} \tilde{K}_n\left(\frac{p_1}{k}\right) \tilde{A}_n(k, p_2). \tag{18}$$

Here and later we mean $p = (p^2)^{\frac{1}{2}}$.

The dilatational invariance of the kernel of (18) allows the equation to be solved by taking the Mellin transform with respect to the variable

$$x \equiv p_1/p_2. \tag{19}$$

One then finds that the condition for the singularity is[3]

$$\tilde{K}_n = 8\pi^2(n+1), \tag{20}$$

where

$$\tilde{K}_n \equiv \int_0^\infty \frac{dx}{x} \tilde{K}_n(x). \tag{21}$$

We note here two very useful rules for composition of kernels. Iteration of the kernel $K^{(1+2)} \equiv K^{(1)} + K^{(2)}$ corresponds to random alternation of $K^{(1)}$ and $K^{(2)}$ in the summed graphs. Since (11), (17), and (21) are linear operations this is just reflected in the relation

$$\tilde{K}_n^{(1+2)} = \tilde{K}_n^{(1)} + \tilde{K}_n^{(2)}. \tag{22}$$

Regular alternation of $K^{(1)}$ and $K^{(2)}$ corresponds to iteration of the kernel

$$\tilde{K}_n^{(1\otimes 2)}\left(\frac{p_1}{p_2}\right) \equiv \frac{1}{8\pi^2(n+1)} \int_0^\infty \frac{dk}{k} \tilde{K}_n^{(1)}\left(\frac{p_1}{k}\right) \tilde{K}_n^{(2)}\left(\frac{k}{p_2}\right). \tag{23}$$

This is just a convolution whose Mellin-transformed counterpart is

$$\tilde{K}_n^{(1\otimes 2)} = \tilde{K}_n^{(1)} \tilde{K}_n^{(2)} / 8\pi^2(n+1). \tag{24}$$

By the artificial introduction of two-particle intermediate states at the points marked by arrows we have decomposed the following graphs into their respective "units," for which it has been possible to calculate K_n. [See Fig. 2. In the case (c), K_n has been expressed in terms of an infinite sum which can be evaluated in the weak-and strong-coupling limits.]

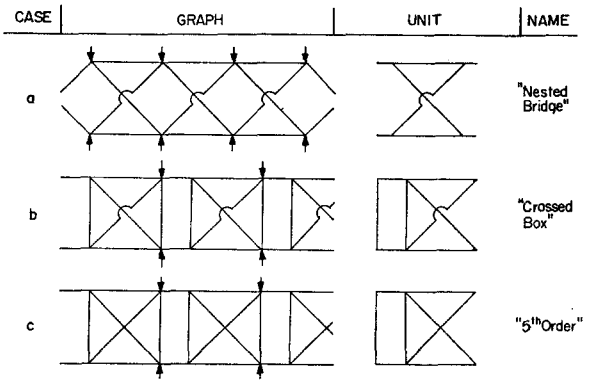


FIG. 2. Specific graphs summed.

The calculation of K_n for cases (b) and (c) of Fig. 2 makes use of the convolution formula (24). The decomposition of these kernels is illustrated in Fig. 3. We understand all subunits to be crossing-symmetrized as discussed above.

We thus need K_n for subunits A, B, and C of Fig. 3. The absorptive parts for these subunits are

$$p_1^2 p_2^2 K_A(\phi, p_1^2, p_2^2) = \frac{\lambda^2}{8\pi} \frac{\phi}{\sinh\phi \cosh\phi}, \tag{25}$$

$$p_1^2 p_2^2 K_B(\phi, p_1^2, p_2^2) = \frac{\lambda^2}{8\pi}, \tag{26}$$

$$p_1^2 p_2^2 K_C(\phi, p_1^2, p_2^2) = -\frac{\lambda^3}{(8\pi)^2} \frac{\phi}{\sinh\phi} \frac{P}{\pi} \int_{\phi_0}^\infty \frac{\phi'\, d\phi'}{\cosh\phi' - \cosh\phi}. \tag{27}$$

The common threshold for these absorptive parts is

$$\phi_0 = \cosh^{-1}\{(p_1^2 + p_2^2)/2p_1p_2\} = |\log x|. \tag{28}$$

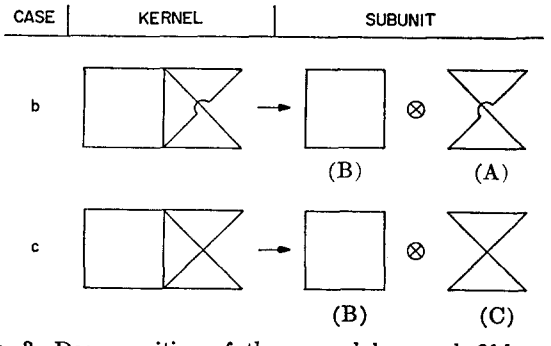


FIG. 3. Decomposition of the crossed box and fifth-order kernel.

Taking the transform (11) and recalling Eq. (17), one obtains

$$\bar{K}_{nA}\left(\frac{p_1}{p_2}\right) = \frac{\lambda^2}{4\pi^2}\int_{\phi_0}^{\infty}\frac{\phi\, d\phi\, e^{-(n+1)\phi}}{\cosh\phi}, \tag{29}$$

$$\bar{K}_{nB}\left(\frac{p_1}{p_2}\right) = \frac{\lambda^2}{4\pi^2}\int_{\phi_0}^{\infty} e^{-(n+1)\phi}\sinh\phi\, d\phi, \tag{30}$$

$$\bar{K}_{nC}\left(\frac{p_1}{p_2}\right) = \frac{\lambda^3}{(8\pi^2)^2}\int_{\phi_0}^{\infty}\phi\, d\phi\int_{\phi_0}^{\infty}\phi'\, d\phi' \times\frac{e^{-(n+1)\phi'} - e^{-(n+1)\phi}}{\cosh\phi' - \cosh\phi}. \tag{31}$$

(The last equation was obtained by symmetrizing the integrand in ϕ and ϕ'.) One then takes the Mellin transform (21) using the substitution (28) and reversing the order of integration so as to perform the ϕ_0 integration first. One obtains[14]

$$\bar{K}_{nA} = \frac{\lambda^2}{2\pi^2}\int_0^{\infty}\frac{\phi^2\, d\phi\, e^{-(n+1)\phi}}{\cosh\phi} = \frac{\lambda^2}{64\pi^2}\left[\psi^{(2)}\left(\frac{n+4}{4}\right) - \psi^{(2)}\left(\frac{n+2}{4}\right)\right], \tag{32}$$

where

$$\psi^{(n)}(x) \equiv \frac{d^{n+1}}{dx^{n+1}}\log\Gamma(x) = (-1)^{n+1}n!\sum_{k=0}^{\infty}\frac{1}{(x+k)^{n+1}}, \tag{33}$$

$$\bar{K}_{nB} = \frac{\lambda^2}{2\pi^2}\int_0^{\infty}\phi\, d\phi\sinh\phi e^{-(n+1)\phi} = \frac{\lambda^2(n+1)}{\pi^2 n^2(n+2)^2}, \tag{34}$$

$$\bar{K}_{nC} = \frac{\lambda^3}{(2\pi)^4}\int_0^{\infty}\phi\, d\phi\int_0^{\phi}\phi'^2\, d\phi' \times\frac{e^{-(n+1)\phi'} - e^{-(n+1)\phi}}{\cosh\phi - \cosh\phi'}. \tag{35}$$

The last expression can be evaluated exactly for $n = 0$, and its asymptotic behavior as $n \to \infty$ can be obtained. (See Appendix A.) We find

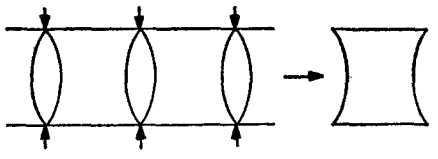


FIG. 4. An artifical decomposition of the bubble exchange diagram.

$$\bar{K}_{0C} = [\lambda^3/(8\pi^2)^2]5\zeta(5), \tag{36}$$

where $\zeta(5)$ is the Riemann zeta-function, and

$$\bar{K}_{nC} \underset{(n\to\infty)}{\to} \frac{\lambda^3\log n}{8\pi^4 n^3} + O\left(\frac{1}{n^3}\right). \tag{37}$$

We now put the "pieces" together to give the kernels of Figs. 2(a), 2(b), and 2(c), using the convolution formula (24). For general λ, Eq. (20) determines n_0 only implicitly. We can solve for n_0 in the weak- and strong-coupling limits.

For the graphs of Fig. 2(a),

n_0 (nested bridge)

$$= -1 + (\lambda^2/2^9\pi^4)[\psi^{(2)}(\tfrac{3}{4}) - \psi^{(2)}(\tfrac{1}{4})]$$

$$= -1 + (\lambda^2/2^7\pi) \quad \text{(weak coupling)} \tag{38}$$

$$= \lambda^{\frac{1}{2}}/8^{\frac{1}{2}}\pi \quad \text{(strong coupling)}. \tag{39}$$

Here we have used standard properties of $\psi^{(n)}(x)$.[15] For the graphs of Fig. 2(b),

n_0 (crossed box)

$$= \left[\frac{\lambda^4}{(8\pi^2)^3 32\pi^2}\{\psi^{(2)}(1) - \psi^{(2)}(\tfrac{1}{2})\}\right]^{\frac{1}{3}}$$

$$= (\lambda^2/2^6\pi^2)\{3\zeta(3)\}^{\frac{1}{3}} \quad \text{(weak coupling)} \tag{40}$$

$$= \lambda^{\frac{1}{2}}/8^{\frac{1}{2}}\pi \quad \text{(strong coupling)}. \tag{41}$$

For the graphs of Fig. 2(c),

n_0 (5th order)

$$= \left[\frac{2\cdot 5\lambda^5\zeta(5)}{(8\pi^2)^5}\right]^{\frac{1}{5}} \quad \text{(weak coupling)} \tag{42}$$

$$= \frac{\lambda^{5/8}(\frac{5}{8}\log\lambda)^{1/8}}{2^{9/8}(\pi^2)^{5/8}} \quad \text{(strong coupling)}. \tag{43}$$

We see that in (43) n_0 does not even have a power dependence on λ for strong coupling.

Notice that n_0 for the sum of bubble exchange diagrams[2] in Fig. 4 may be obtained very simply from the condition

$$\bar{K}_{nB} = 8\pi^2(n+1)$$

yielding

$$n_0 \text{ (bubble)} = -1 + (1 + \lambda/8^{\frac{1}{2}}\pi^2)^{\frac{1}{2}}. \tag{44}$$

We find an identical strong-coupling behavior in (39), (41), and (44). This is a general feature of certain ϕ^4 ladder graphs for forward scattering which we discuss further in the next section and in Appendix B. It resembles a result for $g\phi^3$ ladder graphs

[14] See, e.g., *Tables of Integral Transforms*, A. Erdelyi, Ed. (McGraw-Hill Book Company, Inc., New York, 1953), Vol. I, p. 163, Eq. (7).

[15] See, e.g., A. Erdelyi *et al.*, *Higher Transcendental Functions* (McGraw-Hill Book Company, Inc., New York, 1953), Vol. I, Chap. I.

for forward scattering with scalar photon exchange that

$$\lim_{g\to\infty} n_0/(g/4\pi m) = 1, \tag{45}$$

where m is the mass of the side lines of the ladder.[16] No such result seems to hold for graphs with "internal" vertices, however, as (43) shows.

For weak coupling we see from (38), (40), (42), and (44) that, of the four sets of graphs in Figs. 2 and 4, bubble exchange gives the dominant contribution to high-energy scattering. We show in the next section that for a general ϕ^4 kernel of Mth order, $n_0 \sim \lambda^{M/2}$ for weak coupling.

Note that we have summed all but one of the ϕ^4 kernels with no inserts up to and including fifth order. (See Fig. 5.) Iteration of the kernel (d) seems impractical from a purely computational standpoint.

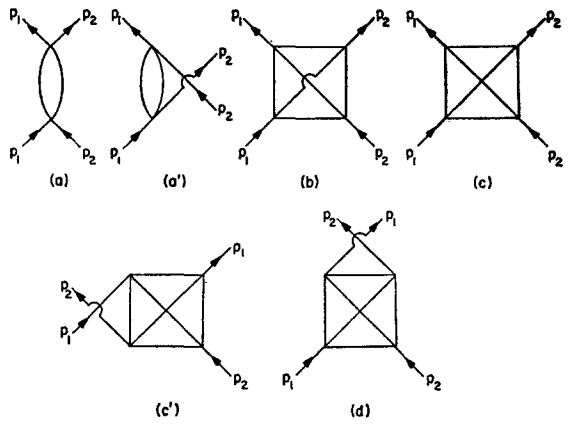


FIG. 5. The fifth-or-less-order ϕ^4 kernels without inserts.

IV. GENERAL WEAK- AND STRONG-COUPLING LIMITS

A convenient expression for $K^{\pm}(\phi, p_1^2, p_2^2)$ exists in terms of Feynman parameters. Consider a ϕ^4 kernel of order M with I internal lines requiring a single subtraction. The scattering amplitude associated with this kernel develops a singularity when the discriminant $D(\alpha, p_1, p_2)$[17] of its Feynman-parametrized denominator becomes negative. The absorptive part corresponding to this singularity may be written[3]

$$K^{\pm}(\phi, p_1^2, p_2^2) = \text{const} \times \lambda^M \times \int_0^1 \frac{d\alpha_1 \cdots d\alpha_I}{(\det a_{ij})^2} \times \delta\Big(1 - \sum_{i=1}^{I} \alpha_i\Big)\{\theta(-H) \pm \theta(H)\} \times \theta\{-D(\alpha, p_1, p_2)\}, \tag{46}$$

where

$$D(\alpha, p_1, p_2) = F(\alpha)p_1^2 + G(\alpha)p_2^2 - 2p_1p_2H(\alpha)\cosh\phi. \tag{47}$$

Det (a_{ij}) is a homogeneous function of degree L of the Feynman parameters α_l, where L is the number of loops in the graph. F, G, and H are homogeneous of degree $L + 1$. Convenient topological rules exist for the calculation of these functions.[18] The functions satisfy the inequalities

$$F \geq |H|, \qquad G \geq |H|. \tag{48}$$

Taking the transform (11) of (46), one obtains

$$K_n^{\pm}\Big(\frac{p_1}{p_2}\Big) = \text{const}\,\lambda^M \int_0^1 \frac{d\alpha_1 \cdots d\alpha_I}{(\det a_{ij})^2}\,\delta\Big(1 - \sum_{i=1}^{I}\alpha_i\Big) \times \{\theta(-H) \pm \theta(H)\}\Big[\frac{1}{n}\Big\{\frac{F}{2\,|H|}\frac{p_1}{p_2} + \frac{G}{2\,|H|}\frac{p_2}{p_1} + \Big[\Big(\frac{F}{2\,|H|}\frac{p_1}{p_2} + \frac{G}{2\,|H|}\frac{p_2}{p_1}\Big)^2 - 1\Big]^{\frac{1}{2}}\Big\}^{-n} - [n \to (n+2)]\Big]. \tag{49}$$

The Mellin transform (21) may be written using the substitution

$$e^y = (F/G)^{\frac{1}{2}}(p_1/p_2) \tag{50}$$

as

$$K_n^{\pm} = \text{const}\,\lambda^M \int_0^{\infty} dy \int_0^1 \frac{d\alpha_1 \cdots d\alpha_I}{(\det a_{ij})^2}\,\delta\Big(1 - \sum_{i=1}^{I}\alpha_i\Big) \times \Big[\Big\{\frac{H^2}{FG}\Big\}^{\frac{1}{2}n}\{\theta(-H) \pm \theta(H)\} \times \frac{1}{n}\Big\{\cosh y + \Big(\cosh^2 y - \frac{H^2}{FG}\Big)^{\frac{1}{2}}\Big\}^{-n} - [n \to (n+2)]\Big]. \tag{51}$$

In the weak-coupling limit one sees from (51) and (20) that $n \to 0$. In this limit the major contribution to (51) comes from the region of large y, where

$$\cosh y + [\cosh^2 y - (H^2/FG)]^{\frac{1}{2}} \to e^y.$$

[16] When massive particles are exchanged in $g\phi^3$ ladder graphs, n_0 increases as $(g/\mu)^{\frac{1}{2}}$ forl arge g, in the special cases of the one-particle-exchange kernel (Ref. 13) and the kernel for crossed exchange of two particles (J. Rosner, unpublished). Such a result, however, has not yet been proven in the general case.

[17] R. J. Eden, *1961 Brandeis Summer Institute Lecture Notes* (W. A. Benjamin Company, New York, 1962).

[18] J. B. Boyling, J. Math. Phys. 6, 1469 (1965).

Hence, as $n \to 0$,

$$\bar{K}_n^{\pm} \to \frac{(\text{const})\lambda^M}{n^2} \int \frac{d\alpha_1 \cdots d\alpha_I}{(\det a_{ij})^2} \delta(1 - \sum \alpha_l) \times \{\theta(-H) \pm \theta(H)\}, \quad (52)$$

showing that as $\lambda \to 0$,

$$n_0 \sim \lambda^{\frac{1}{2}M} \times \text{const} \quad (53)$$

for a graph of order M. As mentioned earlier, this implies that the bubble kernel dominates all others in the weak-coupling limit.

In the strong-coupling limit, $n \to \infty$ and only the region where $\cosh y + [\cosh^2 y - (H^2/FG)]$ is very close to 1 contributes to (50). This occurs for small y and $FG \approx H^2$. We define

$$\Delta \equiv (FG - H^2)/FG. \quad (54)$$

Note that $0 \leq \Delta \leq 1$. The limit

$$L(\Delta) \equiv \lim_{n\to\infty} \int_0^\infty dy \{\cosh y + (\sinh^2 y + \Delta)^{\frac{1}{2}}\}^{-n}$$

can then be calculated by scaling. Let $\Delta = \delta/n^2$ and $y = \xi/n$; one obtains

$$L(\Delta) = \Delta^{\frac{1}{2}} K_1(n\Delta^{\frac{1}{2}}), \quad (55)$$

where $K_1(x)$ is a Bessel function, so that

$$\bar{K}_n^{\pm} \underset{(n\to\infty)}{\to} \frac{\lambda^M}{n} \int \frac{d\alpha_1 \cdots d\alpha_I}{(\det a_{ij})^2} \delta(1 - \sum \alpha_l) \times (1 - \Delta)^{\frac{1}{2}n} \Delta^{\frac{1}{2}} K_1(n\Delta^{\frac{1}{2}}) \{\theta(-H) \pm \theta(H)\} - (n \to n + 2). \quad (56)$$

The major contribution to (56) is obtained when

$$\Delta \lesssim 1/n^2 \quad \text{and} \quad (1 - \Delta)^{\frac{1}{2}n} \sim 1.$$

Now Δ vanishes only when $F = G = |H|$, as can be seen from (48). Equality in both relations holds (as can be seen from the rules in Ref. 18) only if:

(i) a subset of the α_l vanishes so that the graph reduces to one of the forms of Fig. 6, or

(ii) a subset of the α_l vanishes so as to short-circuit a closed loop or so as to reduce the graph to the form shown in Fig. 7.

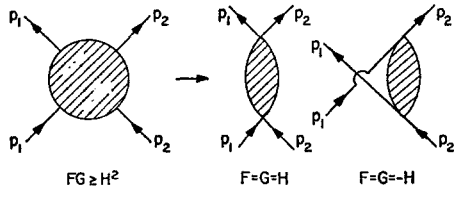


FIG. 6. Reduced graphs contributing to the strong-coupling limit.

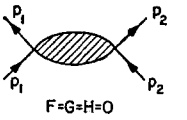


FIG. 7. Reduced graph which does not contribute to the strong-coupling limit.

Since in case (ii) F, G, and H all vanish linearly in some scaling parameter, Δ as defined in (54) will not vanish identically in the remaining α_l. Hence only case (i) need be considered.

Let s denote the minimum number of parameters, say $\alpha_1 \cdots \alpha_s$, whose vanishing reduces the graph as shown in Fig. 6. (It turns out that only these minimal subsets of the α_l need be considered.) The major contribution to (56) as $n \to \infty$ comes when these parameters are near zero.

A method of scaling[19] can be used to discuss these contributions. A direct scaling of the parameters in question, i.e.,

$$\alpha_1 = \mu\alpha_1', \cdots, \alpha_s = \mu\alpha_s',$$

$$\prod_{l=1}^{s} d\alpha_l = \prod_{l=1}^{s} d\alpha_l' \, \delta\Big(1 - \sum_{l=1}^{s} \alpha_l'\Big) \mu^{s-1} \, d\mu, \quad (57)$$

may not be adequate since the integrations over unscaled parameters may diverge when μ is set equal to zero. In general, a "complete set" of scalings of minimal length s, including "preliminary scalings" of various closed loops, must be performed. This is not attempted here; we restrict most of the discussion to the case where the direct scaling (57) is applicable.

For μ small we write

$$\Delta = \sum_{l=1}^{s} \alpha_l \Delta_l + O(\alpha_K^2) \qquad (K = 1, \cdots, s)$$
$$= f\mu + O(\mu^2),$$

where

$$f = \sum_{l=1}^{s} \alpha_l' \Delta_l = \sum_{l=1}^{s} \alpha_l' \Big(\frac{\partial \Delta}{\partial \alpha_l}\Big)_{\substack{\alpha_K = 0 \\ K=1,\cdots,s}}. \quad (58)$$

After the change of variables (57) in (56), the integration over μ then yields

$$\bar{K}_n^{\pm} \underset{(n\to\infty)}{\to} \text{const} \times \frac{\lambda^M}{n^{2s+2}} \int \prod_{l=1}^{s} d\alpha_l' \, \delta\Big(1 - \sum_{l=1}^{s} \alpha_l'\Big) \times \prod_{K=s+1}^{I} d\alpha_K \, \delta\Big(1 - \sum_{K=s+1}^{I} \alpha_K\Big) (\det \hat{a}_{ij})^{-2} f^{-s} - (n \to n + 2). \quad (59)$$

A carat ($\hat{}$) indicates that μ is set equal to zero in

[19] G. Tiktopoulos, Phys. Rev. 131, 480 (1963).

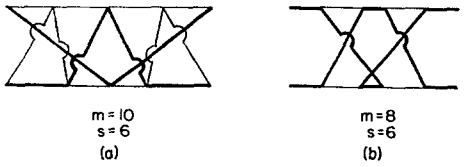


FIG. 8. Ladder graphs with more than one contribution to K_s. These graphs can be reduced in more than one way, e.g., by shrinking heavy lines.

evaluating the quantity in question. When (59) converges we then have

$$K_n \pm \underset{(n\to\infty)}{\to} \text{const}\ (\lambda^M/n^{2s+3}), \tag{60}$$

and (20) finally gives

$$n_0 \sim \lambda^{M/(2s+4)}\ \text{const}. \tag{61}$$

Let us apply the result (61) to ladder graphs. One can approximate the absorptive part of a ladder graph by cutting its rungs alone.[9] (That this is not the only contribution to K_s for general ladder graphs is illustrated in Fig. 8.) Since short-circuiting any rung will reduce a ladder diagram to one which cannot be cut in this manner, the only scaling relevant to the high-n limit in this approximation is that of the side lines, reducing the graph to the form of Fig. 9. In this case the function f [defined in (58)] is

$$f = \left(\sum_{i=1}^{s} \alpha_i'\right) \frac{\det \hat{a}_{ij}}{\hat{H}} = \left(\sum_{i=1}^{s} \alpha_i'\right)\left(\sum_{K=s+1}^{I} \frac{1}{\alpha_K}\right);$$

furthermore,

$$\det \hat{a}_{ij} = \left(\sum_{l=s+1}^{I} \frac{1}{\alpha_l}\right) \prod_{K=s+1}^{I} \alpha_K.$$

The α_i' integrations in (56) are trivial and the convergence of the remaining integral, i.e.,

$$\int \frac{\prod_{K=s+1}^{I} d\alpha_K\ \delta\left(1 - \sum_{K=s+1}^{I} \alpha_K\right)}{\left(\sum_{K=s+1}^{I} \alpha_K^{-1}\right)^{s+2}\left(\prod_{K=s+1}^{I} \alpha_K\right)^2}$$

is verified by power counting. Thus (61) is applicable, and since in this case $s = M - 2$ we obtain for all ladder graphs

$$\lim_{\lambda\to\infty} n_0\lambda^{-\frac{1}{2}} = c_0. \tag{62}$$

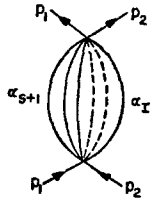


FIG. 9. The reduced ladder graph obtained by shrinking side lines.

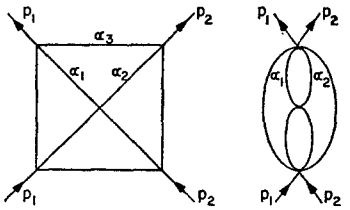


FIG. 10. The reduced ladder graph for the fifth-order diagram.

Indeed, the *same* constant, $c_0 = 1/2\pi$, appears when the contribution of K_n is neglected in all the examples worked in Sec. III except for the one of Fig. 2(c). [When we neglect K_n we must merely divide the right-hand sides of (32) and (34) by 2. Then (39), (41), and (44) will be modified accordingly.] In fact, as is shown in Appendix B, $c_0 = 1/2\pi$ for ϕ^4 ladder graphs in general in the above approximation.

A simple example for which (60) does not apply is the fifth-order graph of Fig. 5(c). In this case f is again $\det \hat{a}_{ij}/\hat{H}$ as calculated from the reduced graph of Fig. 10, and (59) diverges at $\alpha_1 = 0$, $\alpha_2 = 0$. This seems to be related to our result (37) that K_n does not have a simple power dependence on n in the strong-coupling limit. The fact that the power of n in (37) is predicted correctly by (60) suggests, however (bearing in mind analogous high-energy-behavior calculations[19]), that even in the general case (60) may be true up to a power of $\log n$.

Certain ladder graphs of order M can be reduced by shrinking less than $M - 2$ lines [cf. Fig. 8(a)]. These may have by (61) higher n_0 values as $\lambda \to \infty$; hence the approximation leading to (62) may not be valid in the strong-coupling limit. This approximation is, however, natural for weak coupling since for a given number of particles in the intermediate state the contribution to the absorptive part from the lowest-order production diagram is taken. (See Fig. 11.)

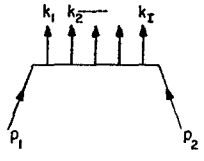


FIG. 11. The simple ϕ^3 production diagram.

V. AN INEQUALITY FOR CONTRIBUTIONS TO ABSORPTIVE PARTS

We now prove an inequality satisfied by contributions of the Cutcosky diagrams shown symbolically in Fig. 12 to the absorptive part. It is assumed that the intermediate lines joining A_1 with A_1 (or A_2 with A_2) in the right half of the diagram join "congruent" vertices with no crossing or per-

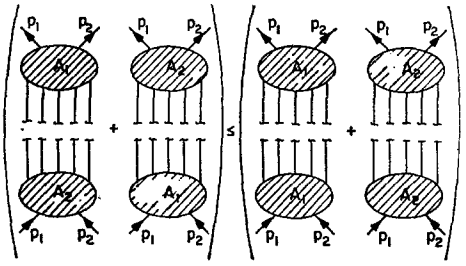


FIG. 12. A symbolic description of the inequality (63).

mutations of lines. Thus, for the particular case in which A_1 is the simple ϕ^3 production diagram (Fig. 11) this joining will form a simple ladder diagram. For the corresponding ϕ^4 case we would obtain the bubble series of Fig. 4.

The inequality, which explicitly reads

$$\int \cdots \int \prod_i \frac{d^3k_i}{2k_i^0} [A_1(p_1p_2k_i)A_2^*(p_1p_2k_i) + A_1^*(p_1p_2k_i)A_2(p_1p_2k_i)] \leq \int \cdots \int \prod_i \frac{d^3k_i}{2k_i^0} [|A_1(p_1p_2k_i)|^2 + |A_2(p_1p_2k_i)|^2] \tag{63}$$

follows directly from the Schwarz inequality.

Letting A_1 and A_2 be diagrams of Fig. 11, or the corresponding ϕ^4 diagrams, we find that the contribution to K_s of uncrossed ladders always majorizes that of the general ladder when the graph is cut in such a way so as to leave its sides intact. Our specific examples (25) and (26) for which only such a cut contributes to K_s illustrate the inequality. Moreover, the result[4,20] that n_0 (truss bridge diagrams) $\leq n_0$ (bubbles) is readily understood using this theorem.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Professor I. J. Muzinich for several helpful discussions and for his continuing interest.

APPENDIX A. BEHAVIOR OF K_{nc} AT $n = 0$ AND $n \to \infty$

In this Appendix we prove (36) and (37) from (35). First, we note that

$$\cosh\phi - \cosh\phi' = \tfrac{1}{2}e^{\phi'}(e^{-\phi'} - e^{-\phi})(e^{\phi} - e^{-\phi'}). \tag{A1}$$

For $n = 0$ Eq. (35) becomes

$$\bar{K}_{0c} = \frac{2\lambda^3}{(2\pi)^4}\int_0^\infty \phi\, d\phi \int_0^\phi \phi'^2\, d\phi'\, e^{-\phi'} \times (e^{\phi} - e^{-\phi'})^{-1}. \tag{A2}$$

We change variables: Define x and y by

$$\phi = \tfrac{1}{2}(x + y), \qquad \phi' = \tfrac{1}{2}(x - y). \tag{A3}$$

Then

$$\bar{K}_{0c} = \frac{\lambda^3}{8(2\pi)^4}\int_0^\infty \frac{e^{-x}\,dx}{1 - e^{-x}}\int_0^x dy\,(x - y)^2(x + y) = 5\lambda^3\zeta(5)/(8\pi^2)^2. \tag{36}$$

For general n, we use the property of the Laplace transform that if it is specified at a denumerable infinity of evenly spaced points along the positive real axis it has a unique analytic continuation. Let us compute expression (31) for the integers $n > 0$. We use (A1) and write

$$\frac{e^{-(n+1)\phi'} - e^{-(n+1)\phi}}{e^{-\phi'} - e^{-\phi}} = \sum_{\tau=0}^{n} e^{-\tau\phi'}e^{-(n-\tau)\phi}. \tag{A4}$$

We also make use of the expansion

$$\frac{1}{1 - e^{-(\phi+\phi')}} = \sum_{\sigma=0}^{\infty} e^{-\sigma(\phi+\phi')} \tag{A5}$$

and find that

$$\bar{K}_{nc}\left(\frac{p_1}{p_2}\right) = \frac{2\lambda^3}{(8\pi^2)^2}\sum_{\sigma=0}^{\infty} e^{-(2\sigma+n+2)\phi_0} \times \sum_{\tau=0}^{n} \{\sigma + n + 1 - \tau\}^{-1}\{\sigma + \tau + 1\}^{-1} \times \{\phi_0 + (\sigma + n + 1 - \tau)^{-1}\}\{\phi_0 + (\sigma + \tau + 1)^{-1}\},$$

or, defining $\sigma + \tau = k$ and $\sigma + n - \tau = m$,

$$\bar{K}_{nc}\left(\frac{p_1}{p_2}\right) = \frac{2\lambda^3}{(8\pi^2)^2}\sum_{m=0}^{\infty}\frac{e^{-(m+1)\phi_0}}{m+1}\left\{\frac{1}{m+1} + \phi_0\right\} \times \sum_{\substack{k=|m-n| \\ (m+n+k \text{ even})}}^{k=m+n} \frac{e^{-(k+1)\phi_0}}{k+1}\left\{\frac{1}{k+1} + \phi_0\right\}. \tag{A6}$$

(It is amusing to note that this result can also be obtained as a partial-wave projection of the kernel K_C^+ in a Euclidean formalism.)

Taking the Mellin transform (21), we obtain

$$\bar{K}_{nc} = \frac{\lambda^3}{(2\pi)^4}\sum_{m=0}^{\infty}\sum_{\substack{k=|m-n| \\ (m+n+k \text{ even})}}^{k=m+n} [(m+1)^{-2}(k+1)^{-2} \times (2 + k + m)^{-1} + (m+1)^{-1} \times (k+1)^{-2}(2 + k + m)^{-2} + (k+1)^{-1}(m+1)^{-2}(2 + k + m)^{-2} + 2(m+1)^{-1}(k+1)^{-1}(2 + k + m)^{-3}]. \tag{A7}$$

[20] J. D. Bjorken and T. T. Wu, Phys. Rev. 130, 2566 (1963).

For high n, we may approximate the condition $[(n + m + k) = \text{even}]$ by reducing the sum by a factor of 2. The asymptotically leading term in (A7) is then found by using partial fractions to be

$$\bar{K}_{nc} \underset{(n\to\infty)}{\to} \frac{\lambda^3 \log n}{8\pi^4 n^3} + O\left(\frac{1}{n^3}\right). \tag{37}$$

APPENDIX B. STRONG-COUPLING BEHAVIOR OF LADDER GRAPHS

In this Appendix we prove a result for ϕ^4 ladder graphs analogous to the result (45) for ϕ^3 ladder graphs. We neglect the contribution of K_n to A_n for formal convenience, as in Ref. 9. Then, using the kinematically obtained bounds of Ref. 9,

$$K_<(\phi, p_1^2, p_2^2) \leq K(\phi, p_1^2, p_2^2) \leq K_>(\phi, p_1^2, p_2^2), \tag{B1}$$

where

$$K_\gtrless(\phi, p_1^2, p_2^2) = \frac{16\pi^3\lambda^M}{(4\pi)^{2M}\Gamma(M)\Gamma(M-1)} \times [\cosh\phi - \cosh\phi_0]^{M-2}(2e^{\pm\phi})^{M-2} \tag{B2}$$

for an Mth-order kernel. [Recall the definition (28) of ϕ_0.] Now take the transform (11) of (B2), and follow with the Mellin transform (21), using the substitution (28). As usual, reverse the order of integration to perform the ϕ integration first. For large n, only small ϕ contributes to the ϕ integral, and we may hence make the approximation

$$\cosh\phi - \cosh\phi_0 \sim \tfrac{1}{2}(\phi^2 - \phi_0^2). \tag{B3}$$

It is then a straightforward matter to perform the ϕ_0 and ϕ integrations, and one finds that

$$\frac{\bar{K}_n}{8\pi^2 n} \underset{(\lambda\to\infty)}{\to} \frac{\lambda^M}{(2\pi n)^{2M}}\left[1 + O\left(\frac{1}{n}\right)\right]. \tag{B4}$$

Application of (20) then shows that

$$\frac{n_0}{\lambda^{\frac{1}{2}}} \underset{(\lambda\to\infty)}{\to} \frac{1}{2\pi}. \tag{B5}$$

One can arrive at (B5) more directly by noting that the bounds (B2) are independent of the detailed structure of the ladder graph. The corresponding bounds on $\bar{K}_n$ coalesce for $n \to \infty$ since the $e^{\pm(M-2)\phi}$ factors become negligible in comparison with the $e^{-(n+1)\phi}$ factor in the transform (11). Then every ladder graph kernel of Mth order is equivalent as $n \to \infty$ to an $\frac{1}{2}M$-fold iterated bubble kernel, for which [by (24)] (B5) can be seen immediately to hold.

A Formalism for Generalized Quantum Mechanics

EDUARD PRUGOVEČKI
*Palmer Physical Laboratory, Princeton University, Princeton, New Jersey**
(Received 1 July 1965)

A generalization of a non-Hilbert-space formalism, developed in two previous papers, is presented. In this generalization observables with a many-dimensional spectrum are considered. Furthermore, the fundamental mathematical tools are generalized by introducing generalized functions instead of measures. It is then proved that a physical equivalence of the present formalism with the Hilbert-space formalism can still be established.

1. INTRODUCTION

OUR aim in the present note is to generalize the formalism of Ref. 1—to which we refer as the $\mathfrak{L}$-formalism. We do that in Sec. 2. In the rest of the sections we investigate the relation of this formalism to that of conventional Hilbert space. After establishing, in Sec. 3, the concept of the Hilbert-space representation of the $\mathfrak{L}$-formalism, we prove in Sec. 4 the existence of such a representation.

We base the present $\mathfrak{L}$-formalism, as we did in I, on two fundamental concepts—the set $\mathfrak{O}$ of observables and the set $\mathfrak{S}$ of all physical states. These two sets constitute the set $\mathfrak{L}$, representing the present theoretical "language". We construct the $\mathfrak{L}$-formalism by endowing the set $\mathfrak{L}$ with a certain structure—imposed on it in the form of four axioms.

By comparison with the formalism brought forth in I, the first generalization consists in considering observables α with an n-dimensional spectrum $\mathbf{S}^\alpha (n = 1, 2 \cdots)$, i.e., with a spectrum which is a subset of an n-dimensional Euclidean space $\mathbf{R}^\alpha$.

The second essential generalization concerns the fundamental mathematical structure of the concept of physical state introduced in I. There, a physical state P was defined in terms of complex probability measures.[2]

$$P(B), \qquad B \in \mathfrak{B}^\alpha, \qquad \alpha \in \hat{\mathfrak{O}}, \tag{1.1}$$

i.e., normalized complex measures defined on the family $\mathfrak{B}^\alpha$ of all Borel sets in the Euclidean space $\mathbf{R}^\alpha$ adjoined to α. It is easy to see that to any such object can be assigned uniquely a continuous linear functional

$$F_P^\alpha(f) = \int f(\lambda)\, dP^\alpha(\lambda), \qquad f \in \mathfrak{D}_0^\alpha, \tag{1.2}$$

on the space $\mathfrak{D}_0^\alpha$. The space $\mathfrak{D}_0^\alpha$ consists of all continuous functions $f(\lambda)$, $\lambda \in \mathbf{R}^\alpha$, of compact support on the Euclidean space $\mathbf{R}^\alpha$. This space is the countable union[3,4] of the normed spaces $\mathfrak{D}^\alpha(r)$ $(r = 1, 2, 3, \cdots)$, consisting of all continuous functions $f(\lambda)$ in $\mathbf{R}^\alpha$ with the support within the set $\{\lambda : |\lambda| \leq r\}$.[5] The norm of an element $f(\lambda) \in \mathfrak{D}^\alpha(r)$ is defined as

$$||f|| = \operatorname*{suprem}_{\lambda \in \mathbf{R}^\alpha} |f(\lambda)|. \tag{1.3}$$

Due to a well-known theorem by Riesz,[6] to every continuous linear functional on $\mathfrak{D}_0^\alpha$ can be assigned[2,3] a finite complex measure by means of which it can be written in the form (2).

It is interesting now to establish whether we can build the $\mathfrak{L}$-formalism not only in terms of linear functionals of the above type, but also in terms of more general linear functionals, which are continuous on other types of fundamental linear spaces of test functions. Namely, if such a formalism would be adopted, then this would prevent the introduction of observables called "questions"—which form the basis of the Birkhoff–von Neumann[7] lattice approach

---

* Present address: Theoretical Physics Institute, University of Alberta, Edmonton, Alberta, Canada.

[1] E. Prugovečki, J. Math. Phys. 7, 1054 (1966); and 7, 1070 (1966); hereafter referred to as I and II, respectively.

[2] Contrary to the notation used in I and II, we drop the superscript on an element of the set $\hat{S}$ constituted of n-tuples of elements of S—except when ambiguities might arise. Thus, we write α instead of $\hat{a}$ if $\alpha \in \hat{\mathfrak{O}}$, and in general x instead of $\hat{x}$ for an element of $\hat{S}$.

[3] I. M. Gel'fand and G. E. Schilow, *Verallgemeinerte Funktionen* (Deutscher Verlag der Wissenschaften, Berlin, 1962), Vol. 2. An English translation by G. Tesch of this volume is in preparation. We very often follow the terminology introduced in this book.

[4] A condensed exposition in English of most of the contents of Ref. 3 can be found in the first four chapters of the book: A. Friedman, *Generalized Functions and Partial Differential Equations* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[5] If $\lambda \in \mathbf{R}^\alpha$, where $\mathbf{R}^\alpha$ is an n-dimensional Euclidean space and $\lambda = (\lambda_1, \cdots, \lambda_n)$, then $|\lambda|$ denotes $(\lambda_1^2 + \cdots + \lambda_n^2)^{\frac{1}{2}}$.

[6] F. Riesz and B. Sz-Nagy, *Functional Analysis*, translated by L. F. Boron (Frederick Ungar Publishing Company, New York, 1955), p. 110.

[7] G. Birkhoff and J. Von Neumann, Ann. Math. 37, 835 (1936).

to the formalism of quantum mechanics. The issue of not being able to introduce "questions" in a quantum mechanical way is directly related to the problem of whether it is meaningful to attach a probability to the outcome of a measurement in an absolutely sharply defined region of the spectrum of *any* observable with a continuous spectrum. Considerations like those of Bohr and Rosenfeld[8] and Heisenberg,[9] which have motivated the introduction of "smeared" fields in quantum field theory,[10] make this point worth considering.

As there is no theory of measurement which might indicate which fundamental spaces of test functions are worth considering, we will keep our considerations most general in this respect.

We define the concept of a probability functional by the following construction.

In the considered theory in which the set $\mathcal{O}$ of all observables is given, we assign to each single observable α a space $\mathfrak{F}^\alpha$ of functions $f(\lambda)$ defined on $\mathbf{R}^\alpha$, which fulfills the condition of a fundamental space in the Gel'fand sense,[3] or is a union of topological products of such spaces (see Sec. 2, Axiom IA) in case that α denotes an ordered finite set of observables. We require that $\mathfrak{F}^\alpha$, $\alpha \in \mathcal{O}$, is a complete normed or countably normed space, or a countable union of such spaces[11] of p times differentiable $[p=(0,\cdots,0),(1,0,\cdots,0),\cdots,(+\infty,\cdots,+\infty)]$[12] complex functions defined and finite at each point of $\mathbf{R}^\alpha$.

If it is required that the functions $f(\lambda)$ in $\mathfrak{F}^\alpha$, $\alpha \in \mathcal{O}$, are differentiable at most p times, where p is an s-tuple of finite integers, then $\mathfrak{F}^\alpha$ should be a normed space or a countable union of such spaces possessing the following structure: There should exist a real function $M(\lambda) \geq 1$, $\lambda \in \mathbf{R}^\alpha$, having at each point a finite or infinite value and continuous on the set $\{\lambda : M(\lambda) < +\infty\}$, i.e., on the subset of $\mathbf{R}^\alpha$ on which it has finite values. Each function $f(\lambda) \in \mathfrak{F}^\alpha$ should be identically zero on the set $R_M^\alpha = \{\lambda: M(\lambda) = +\infty\}$, i.e., if $M(\lambda) = +\infty$, then $f(\lambda) = 0$, $f \in \mathfrak{F}^\alpha$. Then we define

$$||f|| = \sup_{\substack{\lambda \in \mathbf{R}^\alpha \\ q \leq p}} M(\lambda)\,|D^q f(\lambda)|, \tag{1.4}$$

where

$$D^q f(\lambda) = \frac{\partial^{|q|} f(\lambda)}{\partial\lambda_1^{q_1} \cdots \partial\lambda_s^{q_s}}, \qquad |q| = q_1 + \cdots + q_s, \tag{1.5}$$

if $\mathbf{R}^\alpha$ is an s-dimensional space; here $q \leq p$ means that $q_1 \leq p_1, \cdots, q_s \leq p_s$ and $M(\lambda)\,|D^k f(\lambda)|$ is taken equal to zero when $M(\lambda) = +\infty$.

The more general case of a countable union[13] of such spaces which we consider occurs when we have an infinite sequence

$$1 \leq M_1(\lambda) \leq M_2(\lambda) \leq \cdots, \lambda \in \mathbf{R}^\alpha, \tag{1.6}$$

of functions of the above type, and to each such function $M_k(\lambda)$, $k = 1, 2, 3, \cdots$, corresponds a normed space $\mathfrak{F}_k^\alpha$ of the above introduced type with a norm defined by (1.4); the finite number p for which $D^p f(\lambda)$ is required to exist and be continuous for any $f(\lambda) \in \mathfrak{F}_k^\alpha$ is taken to be the *same for all* k. It is easy to check that we can introduce indeed the countable union $\mathfrak{F}^\alpha$ of all the spaces $\mathfrak{F}_k^\alpha$, $k=1,2,3,\cdots$. The earlier mentioned space $\mathfrak{D}_0^\alpha$ is an illustration of such a kind of object representing the case when $p = (0, \cdots, 0)$ and

$$M_k(\lambda) = \begin{cases} 1, & |\lambda| \leq K, \\ \infty, & |\lambda| > K. \end{cases} \tag{1.7}$$

The most general type of countably normed spaces that we consider occurs when we are dealing with spaces of functions that are arbitrarily many times differentiable with respect to at least one of the arguments, i.e., $|p| = +\infty$. In that case we assume again that there is given in $\mathbf{R}^\alpha$ a sequence (1.6) of functions $M_k(\lambda)$ having the above-mentioned properties. We have $\mathfrak{F}^\alpha = \bigcap_{k=1}^\infty \mathfrak{F}_k^\alpha$, where each $\mathfrak{F}_k^\alpha$ is a complete normed space of p times differentiable functions which identically vanish together with all their derivatives on $R_M^\alpha = \{\lambda : M_k(\lambda) = +\infty, k = 1,2,\cdots\}$.[14] The norm in $\mathfrak{F}_k^\alpha$ is defined as equal

$$||f||_k = \sup_{\substack{\lambda \in \mathbf{R}^\alpha \\ |q| \leq k,\, q \leq p}} M_k(\lambda)\,|D^q f(\lambda)|. \tag{1.8}$$

---

[8] N. Bohr and L. Rosenfeld, Kgl. Danske Videnskab. Selsab., Mat.-Fys. Medd. 12, No. 8 (1933); Phys. Rev. **78**, 794 (1950).

[9] W. Heisenberg, Verhandl. Sachs. Akad Leipzig **83**, 3 (1931); **86**, 317 (1934).

[10] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics and All That* (W. A. Benjamin Company, Inc., New York, 1964).

[11] See Ref. 3, Chap. 1, Sec. 8.

[12] Assume that $\mathbf{R}^\alpha$ is an s-dimensional (Euclidean) space and $\lambda = (\lambda_1,\ldots,\lambda_s)$ represents a point in $\mathbf{R}^\alpha$. If $p = (p_1,\ldots,p_s)$, then a p-times differentiable function $f(\lambda)$ is a function which is p_k-times continuously differentiable in the argument $\lambda_k (k = 1, \ldots, s)$. If $p_k = 0$ then $f(\lambda)$ is just continuous in λ_k while $p_k = \infty$ indicates that $f(\lambda)$ is arbitrarily many times differentiable in the variable λ_k.

[13] We are, naturally, talking of a slight modification of the concept introduced in Ref. 3, Chap. 1, Sec. 8, which applied to unions of countably normed spaces.

[14] In case that we have a sequence (1.6) of functions M_k, $k = 1, 2, \cdots$, it is assumed that all of them will have the value ∞ for (and only for) a point belonging to a set R_M^α (which for some Gel'fand-type spaces can be empty).

In Ref. 3 it is shown that the so defined $\mathfrak{F}^\alpha$ is indeed a countably normed space. An example of a space of this type is the S space of faster-then-polynominal-decrease-at-infinity functions on which the tempered distributions are defined.

The definition of spaces which are countable unions of countably normed spaces proceeds along the indicated lines in an expected fashion.

We call any space of functions of one of the above-described forms a Gel'fand-type space of functions.

In order to define now the concept of the probability functional intended to replace the concept of complex probability measure, we are concerned with the properties of such functionals corresponding to the following properties of complex probability measures:

Each complex probability measure $P^\alpha(B), B \in \mathfrak{B}^\alpha$, corresponding to the ordered set α of observables is normalized in the sense that

$$P^\alpha(\mathbf{R}^\alpha) = 1. \tag{1.9}$$

Besides, due to the continuity from below of any measure, we have

$$\lim_{n\to\infty} P^\alpha(B_n) = 0 \tag{1.10}$$

if $B_1, B_2, \cdots \in \mathfrak{B}^\alpha$ is a sequence of Borel sets having the property that, for any bounded set Δ, we can find an $n_0(\Delta)$ such that all B_n, $n \geq n_0(\Delta)$, are contained in $\mathbf{R}^\alpha - \Delta$.

In order to generalize property (1.10) we introduce the concept of the *zero-approximating sequence*

$$f_1, f_2, \cdots \in \mathfrak{F}^\alpha, \qquad \alpha \in \mathfrak{O}, \tag{1.11}$$

as a sequence of uniformly bounded functions

$$|f_k(\lambda)| \leq C, \qquad \lambda \in \mathbf{R}^\alpha, \qquad k = 1, 2, \cdots, \tag{1.12}$$

having the following additional property: for any bounded subset Δ of $\mathbf{R}^\alpha$ there can be found such an $n_0(\Delta)$ that the support of all f_n, $n \geq n_0(\Delta)$, lies within $\mathbf{R}^\alpha - \Delta$.

Similarly, a uniformly bounded sequence (1.11) is called an *identity-approximating sequence* if for each closed bounded subset Δ of $\mathbf{R}^\alpha$ there can be found an $n_0(\Delta)$ such that $f_n(\lambda) = 1$ for all $\lambda \in \Delta$ and $n \geq n_0(\Delta)$ (if in addition $\Delta \cap R^\alpha_M = \varnothing$ in case that $R^\alpha_M \neq \varnothing$).

In Sec. 2 (Axiom IA) we see that in the present formalism we attach to each n-tuple $\alpha = (\alpha_1, \cdots, \alpha_n)$ of observables a direct product

$$\mathfrak{F}^{\alpha_1\times\cdots\times\alpha_n} = \mathfrak{F}^{\alpha_1} \times \cdots \times \mathfrak{F}^{\alpha_n} \tag{1.13}$$

of the spaces $\mathfrak{F}^{\alpha_1}, \cdots, \mathfrak{F}^{\alpha_n}$ attached to the individual observables $\alpha_1, \cdots, \alpha_n$, respectively. A sequence

$$f_1, f_2, \cdots \in \mathfrak{F}^{\alpha_1\times\cdots\times\alpha_n} \tag{1.14}$$

of n-tuples

$$f_k(\lambda_1, \cdots, \lambda_n) = [f_{k1}(\lambda_1), \cdots, f_{kn}(\lambda_n)],$$
$$\lambda_1 \in \mathbf{R}^{\alpha_1}, \cdots, \lambda_n \in \mathbf{R}^{\alpha_n}, \tag{1.15}$$

of functions belonging, respectively, to $\mathfrak{F}^{\alpha_1}, \cdots, \mathfrak{F}^{\alpha_n}$ is called an identity-approximating sequence in $\mathfrak{F}^{\alpha_1\times\cdots\alpha_n}$, if each of the sequences

$$f_{1j}, f_{2j}, \cdots \in \mathfrak{F}^{\alpha_j}, \qquad j = 1, \cdots, n \tag{1.16}$$

is an identity-approximating sequence in the respective space $\mathfrak{F}^{\alpha_j}$.

A probability functional in the space $\mathfrak{F}^{\alpha_1\times\cdots\times\alpha_n}$ is a functional which is linear and continuous on each of the space $\mathfrak{F}^{\alpha_1}, \cdots, \mathfrak{F}^{\alpha_n}$ separately and which has the additional properties that

$$\lim_{n\to\infty} F^\alpha(f \times h_n \times g) = 0 \tag{1.17}$$

for any zero-approximating sequence from some fixed $\mathfrak{F}^{\alpha_j}$ and

$$\lim_{n\to\infty} F^\alpha(f_n) = 1 \tag{1.18}$$

for any identity-approximating sequence (1.14).

2. THE BASIC AXIOMATIC STRUCTURE FOR THE £ FORMALISM

In formulating the basic axioms we follow to some extent the exposition of I in order to make obvious the way in which the axioms of I, Sec. 2.2 have been generalized. Thus, many of the remarks occurring in I, which are not repeated here, are easily transferable to the present case in an obvious way. But first, it is necessary to introduce some convenient notational agreements.

If $\alpha = (\alpha_1, \cdots, \alpha_m) \in \hat{\mathfrak{O}}$ and $\beta = (\beta_1, \cdots, \beta_n) \in \hat{\mathfrak{O}}$ then $\alpha \times \beta$ will denote the ordered set $(\alpha_1, \cdots, \alpha_m, \beta_1, \cdots, \beta_n) \in \hat{\mathfrak{O}}$. Similarly, if we have adjoined to $\alpha, \beta \in \mathfrak{O}$ the spaces $\mathfrak{F}^\alpha$ and $\mathfrak{F}^\beta$, and $f_1 \in \mathfrak{F}^\alpha$, $f_2 \in \mathfrak{F}^\beta$ then $f_1 \times f_2$ denotes the ordered pair of functions

$$[f_1(\lambda_1), f_2(\lambda_2)]$$
$$= f_1(\lambda_1) \times f_2(\lambda_2), \quad \lambda_1 \in \mathbf{R}^\alpha, \quad \lambda_2 \in \mathbf{R}^\beta, \tag{2.1}$$

on the spaces $\mathbf{R}^\alpha$, $\mathbf{R}^\beta$; naturally for the present there is no relation between $f_1 \times f_2$ and some space $\mathfrak{F}^{\alpha\times\beta}$, which we do not even yet require to exist.

Assume now that to each finite ordered set $\alpha \in \hat{\mathfrak{O}}$ of observables it is assigned a space $\mathfrak{F}^\alpha$ of test functions [which might be, naturally, of the form (1.13)].

Take $\alpha_1, \cdots, \alpha_n \in \hat{\mathfrak{O}}$ and assume that we have a functional $F^{\alpha_1 \times \cdots \times \alpha_n}(f)$ defined for each $f \in \mathfrak{F}^{\alpha_1, \cdots, \alpha_n}$. If $f_1(\lambda_1), \cdots, f_n(\lambda_n)$ are functions belonging to $\mathfrak{F}^{\alpha_1}, \cdots, \mathfrak{F}^{\alpha_n}$, respectively, and such that the function

$$[f_1 \times \cdots \times f_n](\lambda_1, \cdots, \lambda_n) = [f_1(\lambda_1), \cdots, f_n(\lambda_n)], \quad (\lambda_1, \cdots, \lambda_n) \in \mathbf{R}^{\alpha_1 \times \cdots \times \alpha_n}, \tag{2.2}$$

belongs to $\mathfrak{F}^{\alpha_1 \times \cdots \times \alpha_n}$, then the symbol

$$F^{\alpha_1; \cdots; \alpha_n}(f_1 \times \cdots \times f_n) \tag{2.3}$$

means that, in determining the value of $F^{\alpha_1, \cdots, \alpha_n}$ at the point $f_1 \times \cdots \times f_n$, the kth factor in $f_1 \times \cdots \times f_k \times \cdots \times f_n$ refers to the kth ordered set α_k of observables between the two *semicolons*. It is easily recognizable that this notation corresponds to the one systematically used in I and II. Notation (2.3) has to be carefully distinguished from the symbol $F^{\alpha_1 \nu \cdots \nu \alpha_n}(f_1 \times \cdots \times f_n)$ which is introduced next.

In Axiom I we introduce the set $\mathfrak{O}$ of all observables, while a physical state F will be defined as a family

$$F = \{F^\alpha : \alpha \in \hat{\mathfrak{O}}\} \tag{2.4}$$

of functionals F^α defined on $\mathfrak{F}^\alpha$. If the relation

$$F_1^{\alpha_1; \gamma; \beta_1}(f_1 \times h \times g_1) = F_2^{\alpha_2; \gamma; \beta_2}(f_2 \times h \times g_2), \quad \gamma \in \hat{\mathfrak{O}}, \quad h \in \mathfrak{F}^\gamma, \tag{2.5}$$

is true for the physical states F_1 and F_2, for given $\alpha_1, \beta_1, \alpha_2, \beta_2 \in \hat{\mathfrak{O}}$, $f_1 \in \mathfrak{F}^{\alpha_1}$, $g_1 \in \mathfrak{F}^{\beta_1}$, $f_2 \in \mathfrak{F}^{\alpha_2}$, $g_2 \in \mathfrak{F}^{\beta_2}$, and for all $\gamma \in \hat{\mathfrak{O}}$, $h \in \mathfrak{F}^\gamma$ then we write shortly

$$F_1^{\alpha_1 \nu \beta_1}(f_1 \times g_1) = F_2^{\alpha_2 \nu \beta_2}(f_2 \times g_2). \tag{2.6}$$

In general, if in a certain relation involving physical states the "variables" $v_1, \cdots, v_k$ appear, then the relation is true whenever $v_1, \cdots, v_k$ are replaced by any "constants" $\gamma_1, \cdots, \gamma_k \in \hat{\mathfrak{O}}$, and any elements of $\mathfrak{F}^{\gamma_1}, \cdots, \mathfrak{F}^{\gamma_k}$, respectively, are inserted as arguments. For example, the latter appearing relation (2.11) is equivalent to the relation

$$F^{\beta; \alpha_1; \cdots; \alpha_n; \gamma}(g \times f_1 \times \cdots \times f_n \times h) = F^{\beta; \alpha; \gamma}[g \times (f_1 \cdots f_n) \times h] \tag{2.7}$$

valid for any $\beta, \gamma \in \hat{\mathfrak{O}}$ and any $g \in \mathfrak{F}^\beta$, $h \in \mathfrak{F}^\gamma$. We might mention at this spot that if $f_1, \cdots, f_n$ are elements of $\mathfrak{F}^\alpha$ then $f_1 \cdots f_n$ denotes the function[15] ($n \geq 2$)

$$[f_1 \cdots f_n](\lambda) = f_1(\lambda) \cdots f_n(\lambda) \in \mathfrak{F}^\alpha, \quad \lambda \in \mathbf{R}^\alpha. \tag{2.8}$$

Axiom I

In order to have an $\mathfrak{L}$ formulism (or $\mathfrak{L}$ language) it is *necessary* to have a set $\mathfrak{O}$ (called the set of observables), a rule assigning to each element α belonging to the family $\hat{\mathfrak{O}}$ of all finite ordered sets of elements of $\mathfrak{O}$ a fundamental space $\mathfrak{F}^\alpha$ of test functions, and a set $\mathfrak{S}$ of families (2.4) of probability functionals F^α defined on $\mathfrak{F}^\alpha$. The following fundamental requirements have to be fulfilled:

(A) The space $\mathfrak{F}^\alpha$ associated with any finite ordered set $\alpha = (\alpha_1, \cdots, \alpha_n)$ of observables $\alpha_1, \cdots, \alpha_n$ consists of all functions

$$f_1 \times \cdots \times f_n, \quad f_1 \in \mathfrak{F}^{\alpha_1}, \cdots, f \in \mathfrak{F}^{\alpha_n}. \tag{2.9}$$

(B) If $\alpha \in \hat{\mathfrak{O}}$ and if $f_1, f_2, \cdots \in \mathfrak{F}^\alpha$ is an identity-approximating sequence then

$$\lim_{n \to \infty} F^{\nu_1 \alpha \nu_2}(f_n) = F^{\nu_1 \nu_2} \tag{2.10}$$

for any $F \in \mathfrak{S}$.

(C) If $\alpha_1, \cdots, \alpha_n$ denote the same observable α then[15]

$$F^{\nu_1 \alpha_1; \cdots; \alpha_n \nu_2}(f_1 \times \cdots \times f_n) = F^{\nu_1 \alpha \nu_2}(f_1 \cdots f_n) \tag{2.11}$$

for any $F \in \mathfrak{S}$ and any $f_1, \cdots, f_n \in \mathfrak{F}^\alpha$.

(D) For any observable α, any real nonnegative test function $f \in \mathfrak{F}^\alpha$ and any physical state F

$$F^\alpha(f) \geq 0. \tag{2.12}$$

The axiom just formulated should be compared with Axiom I in I, Sec. 2.2. The present points B, C, D, correspond to that axiom, while point A introduces a feature which makes the present formulism different in this respect from that of I. Namely, the case corresponding to the generalization of the approach adopted in I would occur if we would associate a linear space of test functions with each finite ordered set of observables. Instead, Axiom IA was chosen because it imposes a weaker demand on the introduced axiomatic structure and greatly simplifies the proof of the physical equivalence of the present formalism with the Hilbert space formalism.

It should be mentioned that Axiom IA is the most flexible part of the presently introduced axiomatic system. A number of alternatives is possible which would introduce either no change or only minor modifications in the later proof of the physical equivalence of the present formalism to the Hilbert space formalism. One of these alternatives which

[15] It is indeed easy to check on basis of properties with which $\mathfrak{F}^\alpha$ is endowed (Sec. 1) that $f_1 \ldots f_n$ belongs indeed to $\mathfrak{F}^\alpha$ if $f_1, \ldots, f_n \in \mathfrak{F}^\alpha$.

might be worth considering in order to make the notion of an observable (which is a function of other observables) more general is the following: linear Gel'fand-type spaces of test functions would be associated with finite sets of compatible observables while with an ordered set α of incompatible observables we would associate unions of direct products of the spaces corresponding to each partition of the set α into subsets of compatible observables.

This brings us to the notion of compatible observables which, from the formal point of view, is introduced by the following definition:

Definition 1: The set $\{\alpha_1, \cdots, \alpha_n\}$ is called a set of compatible observables if and only if

$$F^{\nu_1\alpha_1;\cdots;\alpha_n\nu_2}(f_1 \times \cdots \times f_n) = F^{\nu_1\alpha_{k_1};\cdots;\alpha_{k_n}\nu_2}(f_{k_1} \times \cdots \times f_{k_n}), \quad \times f_1 \in \mathfrak{F}^{\alpha_1}, \cdots, f_n \in \mathfrak{F}^{\alpha_n}, \tag{2.13}$$

for any $F \in \mathcal{S}$ and any permutation $k_1, \cdots, k_n$ of the indices $1, \cdots, n$, and if also

$$F^{\alpha_1,\cdots,\alpha_n}(f) \geq 0, \qquad F \in \mathcal{S}, \tag{2.14}$$

for any real nonnegative test function $f \in \mathfrak{F}^{\alpha_1\times\cdots\times\alpha_n}$.

If the ordered $\alpha = (\alpha_1, \cdots, \alpha_n)$ is a set of compatible observables, then we write symbolically $\{\alpha_1, \cdots, \alpha_n\} = C$ or shortly $\{\alpha\} = C$.

Proposition 1: If $\alpha_1, \cdots, \alpha_n \in \mathcal{O}$ and $\{\alpha_1, \cdots, \alpha_n\} = C$ then for any

$$0 \leq f_k(\lambda_k) \leq 1, \quad \lambda_k \in \mathbf{R}^{\alpha_k}, \quad k = 1, \cdots, n, \tag{2.15}$$

we have (according to Axiom ID the order of $\alpha_1, \cdots, \alpha_n$ irrelevant)

$$0 \leq F^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) \leq F^{\alpha_k}(f_k) \leq 1, \qquad k = 1, \cdots, n. \tag{2.16}$$

Proof: Because of (2.13) and (2.15) $F^{\alpha_1;\cdots;\alpha_n} \times (f_1 \times \cdots \times f_n) \geq 0$. A glance at the definition of $\mathfrak{F}^{\alpha}$, $\alpha \in \mathcal{O}$, shows that it is possible to find such identity-approximating sequences $f_{k1}, f_{k2}, \cdots \in \mathfrak{F}^{\alpha_k}$ $(k = 2, 3, \cdots, n)$ that $f_{k1} = f_k$. The sequence

$$f_{21} \times \cdots \times f_{n1}, \qquad f_{22} \times \cdots \times f_{n2}, \cdots, \in \mathfrak{F}^{\alpha_2\times\cdots\times\alpha_n} \tag{2.17}$$

obviously is an identity-approximating sequence in $\mathfrak{F}^{\alpha_1\times\cdots\times\alpha_n}$. Standard methods[3,4] would enable us to construct $f_{k1}, f_{k2}, \cdots$ $(k = 2, 3, \cdots, n)$ so that

$$f_{kl}(\lambda_k) \leq f_{kl+1}(\lambda_k), \qquad \lambda_k \in \mathbf{R}^{\alpha_k}, \quad l = 1, 2, \cdots. \tag{2.18}$$

From (2.13), (2.17) and an application of Axiom IB it follows that $F^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) \leq F^{\alpha_1}(f_1)$. A similar procedure yields that $F^{\alpha_1}(f_1) \leq 1$. Q.E.D.

We can now *define* the concept of equality of elements of $\mathcal{S}$ and $\mathcal{O}$ in the same way as it was done in I (Axioms II and III).

Definition 2: Two physical states F_1 and F_2 are said to be equal, symbolically written $F_1 = F_2$, if and only if

$$F_1^{\alpha}(f) = F_2^{\alpha}(f) \tag{2.19}$$

for all $\alpha \in \hat{\mathcal{O}}$, $f \in \mathfrak{F}^{\alpha}$.

Definition 3: Two observables α, β are said to be equal if and only if

$$\mathfrak{F}^{\alpha} \equiv \mathfrak{F}^{\beta} \tag{2.20}$$

and if also

$$F^{\nu_1\alpha\nu_2}(f) = F^{\nu_1\beta\nu_2}(f) \tag{2.21}$$

for all $f \in \mathfrak{F}^{\alpha} \equiv \mathfrak{F}^{\beta}$.

We note that, among other things, relation (2.20) implies that if $\alpha = \beta$ then both observables are of the same dimension.

It is obvious that the following relations (2.22) and (2.23) are equivalent ways of writing, respectively, relations (2.19) and (2.21):

$$F_1^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) = F_2^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n), \quad \alpha_1, \cdots, \alpha_n \in \mathcal{O}, \quad f_1 \in \mathfrak{F}^{\alpha_1}, \cdots, f_n \in \mathfrak{F}^{\alpha_n}, \quad n = 1, 2, \cdots, \tag{2.22}$$

$$F^{\gamma_1';\cdots;\gamma_m';\alpha;\gamma_1'';\cdots;\gamma_n''} \times (f_1 \times \cdots \times f_m \times h \times g_1 \times \cdots \times g_n) = F^{\gamma_1';\cdots;\gamma_m';\beta;\gamma_1'';\cdots;\gamma_n''} \times (f_1 \times \cdots \times f_m \times h \times g_1 \times \cdots \times g_n),$$
$$F \in \mathcal{S}, \quad h \in \mathfrak{F}^{\alpha} \equiv \mathfrak{F}^{\beta}, \quad \gamma_1', \cdots, \gamma_n'' \in \mathcal{O}, \quad f_1 \in \mathfrak{F}^{\gamma_1'}, \cdots, g_n \in \mathfrak{F}^{\gamma_n''}, \quad m, n = 1, 2, \cdots. \tag{2.23}$$

When Axiom IA is replaced with a more involved assumption, it is important to do that in such a fashion that (2.22) and (2.23) are sufficient and necessary conditions for the equality of, respectively, two physical states and two observables.

We can now introduce a topological structure in the space $\mathcal{S}$ of all physical states. A subset of $\mathcal{S}$ of the form

$$W(F_0; \alpha; f_1, \cdots, f_n; \epsilon) = \{F: |F^\alpha(f_1) - F_0^\alpha(f_1)| < \epsilon, \cdots, |F^\alpha(f_n) - F_0^\alpha(f_n)| < \epsilon\} \quad (2.24)$$

is called a weak neighborhood of the physical state F_0. The topology in which a neighborhood basis of an element F_0 of $\mathcal{S}$ consists of all sets (2.24) corresponding to all $\alpha \in \hat{\mathcal{O}}, f_1, \cdots, f_n \in \mathfrak{F}^\alpha$ $(n=1, 2, \cdots)$ and all $\epsilon > 0$ is called the weak topology in $\mathcal{S}$.

Our next axiom endows $\mathcal{S}$ with some structure by making it a convex and weakly closed topological space.

Axiom II

(A) If F_1, F_2 are any two elements of $\mathcal{S}$ and $0 \leq t \leq 1$ then the set of all probability functionals

$$tF_1^\alpha(f) + (1 - t)F_2^\alpha(f), \qquad f \in \mathfrak{F}^\alpha, \quad (2.25)$$

corresponding to all $\alpha \in \hat{\mathcal{O}}$ determines a (unique) physical state F, i.e., $\mathcal{S}$ is a convex set.

(B) If $F_1, F_2, \cdots$ is a Cauchy sequence in the weak topology of $\mathcal{S}$, then it has a limit belonging to $\mathcal{S}$, i.e., $\mathcal{S}$ is closed in the weak topology.

We would like to remark concerning Axiom IIB, that the limit of a Cauchy sequence of continuous linear functionals on $\mathfrak{F}^\alpha$, $\alpha \in \mathcal{O}$, is again a continuous bounded functional.[16] As the weak neighborhoods of physical states have a more direct empirical significance than the states themselves,[17] Axiom IIB represents more a convenience than a serious restriction on the structure of $\mathcal{S}$.

In order to define the concept of spectrum, we introduce the concept of a *zero modulo* $\mathcal{S}$ *test function* $f \in \mathfrak{F}^\alpha$ [symbolically $f = 0 \pmod{\mathcal{S}}$]; it denotes a test function for which $F^\alpha(f) = 0$ for all $F \in \mathcal{S}$. A subset B of $\mathbf{R}^\alpha$ is called a *zero subset of* $\mathbf{R}^\alpha$ if all test functions in $\mathfrak{F}^\alpha$ with a support entirely within B are equal to zero modulo $\mathcal{S}$.

Definition 4: A point λ in $\mathbf{R}^\alpha$, $\alpha \in \hat{\mathcal{O}}$, is said to be a point of the spectrum of α if all intervals in $\mathbf{R}^\alpha$ containing λ are nonzero subsets of $\mathbf{R}^\alpha$. The set of all such points is called the spectrum $\mathbf{S}^\alpha$ of α.

It is obvious that:

Proposition 2: The spectrum $\mathbf{S}^\alpha$ of any $\alpha \in \hat{\mathcal{O}}$ is a closed subset of $\mathbf{R}^\alpha$.

Our next axiom imposes some restrictions on the structure of the spectra.

Axiom III

(A) If $\alpha_1, \cdots, \alpha_n$ are any n observables $(n = 1, 2, \cdots)$ then

$$\mathbf{S}^{\alpha_1 \times \cdots \times \alpha_n} \subset \mathbf{S}^{\alpha_1} \times \cdots \times \mathbf{S}^{\alpha_n}. \quad (2.26)$$

(B) If $\{\alpha\} = C$, $\alpha \in \hat{\mathcal{O}}$, and I is an interval in $\mathbf{R}^\alpha$ containing a point of the spectrum $\mathbf{S}^\alpha$ of α and such that $I \cap R_M^\alpha = \varnothing$,[14] then for each $f \in \mathfrak{F}^\alpha$ having the property

$$f(\lambda) = 1, \qquad \lambda \in I, \quad (2.27)$$

there is such an $F \in \mathcal{S}$ that

$$F^\alpha(f) = 1. \quad (2.28)$$

It has to be remarked that in case $\{\alpha_1, \cdots, \alpha_n\} = C$, Axiom III(A) is a consequence of the Definition 1a and of Proposition 1; this can be proved in much the same manner as Proposition 6 in I, Sec. 2.2, has been proved.

We have reached the stage where we are able to introduce the concept of a function of observables. In order to find a suitable definition of this concept, we turn first to the definition of this concept in I, Sec. 2.2. In that case if $\alpha \in \mathcal{O}$ and $\varphi(\lambda)$ is a Borel-measurable function on $\mathbf{R}^\alpha$, we define $\beta = \varphi(\alpha)$ as that observable for which

$$P^\beta(B) = P^\alpha[\varphi^{-1}(B)], \quad B \in \mathfrak{B}^\alpha, \quad P \in \mathcal{S}. \quad (2.29)$$

To see how this definition can be formulated when the complex probability measures $P^\alpha(B)$ are replaced by probability functionals $F^\alpha(f)$, $f \in \mathfrak{D}_0^\alpha$, we write according to (1.2)

$$F^\beta(f) = \int f(\lambda)\, dP^\beta(\lambda) = \int f(\lambda)\, dP^{\varphi(\alpha)}(\lambda) = \int f(\lambda)\, dP^\alpha(\varphi^{-1}(\lambda)) = \int f[\varphi(\lambda)]\, dP^\alpha(\lambda), \quad (2.30)$$

where we used the notation

$$P^\alpha(\lambda) = P^\alpha(B_\lambda), \qquad \varphi^{-1}(\lambda) = \varphi^{-1}(B_\lambda), \qquad B_\lambda = (-\infty, \lambda). \quad (2.31)$$

We have to take into consideration that, in order that

$$\int f[\varphi(\lambda)]\, dP^\alpha(\lambda) = F^\alpha(f[\varphi(\lambda)]), \quad (2.32)$$

we must require that $f[\varphi(\lambda)] \in \mathfrak{F}^\alpha$.

Now that we have the clue, in order to make our definition perfectly general and meaningful, we have to introduce the following auxiliary spaces of functions.

We denote by $\mathfrak{F}^\alpha(\beta)$ the set of all functions

[16] Ref. 3, Chap. I, Sec. 5.6.

[17] E. Prugovečki, *"On a Theory of Measurement of Incompatible Observables in Quantum Mechanics"* (to be published).

$\varphi(\lambda)$, $\lambda \in \mathbf{R}^\alpha$, with domain of definition $\mathbf{R}^\alpha$ and domain of values $\mathbf{R}^\beta$ which are such that all functions

$$f[\varphi(\lambda)], \qquad \lambda \in \mathbf{R}^\alpha, \tag{2.33}$$

belong to $\mathfrak{F}^\alpha$ for a set of functions $f \in \mathfrak{F}^\beta$ everywhere dense in $\mathfrak{F}^\beta$.

Definition 5: We say that the observable β is a function of the observable α if there exists a function $\varphi(\lambda)$, $\lambda \in \mathbf{R}^\alpha$, belonging to $\mathfrak{F}^\beta(\alpha)$ which is such that

$$F^\beta(f) = F^\alpha(f[\varphi]), \qquad f[\varphi] \in \mathfrak{F}^\alpha, \tag{2.34}$$

for all $F \in \mathcal{S}$; here the $f[\varphi]$'s denote the functions (2.33) belonging to $\mathfrak{F}^\alpha$. In that case we write $\beta = \varphi(\alpha), \varphi \in \mathfrak{F}^\beta(\alpha)$.

It might naturally happen that, for given spaces $\mathfrak{F}^\alpha$ and $\mathfrak{F}^\beta$, the space $\mathfrak{F}^\beta(\alpha)$ is empty due to the fact that the spaces $\mathfrak{F}^\alpha$ and $\mathfrak{F}^\beta$ might be of a different nature. This feature of the formulated framework which makes the explicit giving of the space $\mathfrak{F}^\alpha$ for each $\alpha \in \mathfrak{O}$ essential prevents us to introduce an axiom like Axiom IV in I, Sec. 2.2. Namely, for a given fundamental set of observables, the definition of all possible versions of functions of these observables would present a cumbersome task of a completely formal character and not serving any real purpose or introducing any simplifications. A weaker version is introduced instead.

The following convenient notation is introduced:

If α denotes the n-tuple $(\alpha_1, \cdots, \alpha_n) \in \hat{\mathfrak{O}}$ then we denote with α^* the n-tuple $(\alpha_n, \cdots, \alpha_1) \in \hat{\mathfrak{O}}$. Similarly, λ^* denotes the point $(\lambda_n, \cdots, \lambda_1)$ of $\mathbf{R}^{\alpha^*}$ if λ stood for the point $(\lambda_1, \cdots, \lambda_n)$ of $\mathbf{R}^\alpha$. If $f(\lambda)$, $\lambda \in \mathbf{R}^\alpha$, is a function belonging to $\mathfrak{F}^\alpha$ and if f stands for it as an element of $\mathfrak{F}^\alpha$, then f^* denotes the element of $\mathfrak{F}^{\alpha^*}$ which is equal to the function $\overline{f(\lambda^*)}$, $\lambda^* \in \mathbf{R}^{\alpha^*}$.

Axiom IV

For any $F \in \mathcal{S}$ and any $\beta \in \hat{\mathfrak{O}}$ the expression

$$M = \sum_{i,j=1}^{n} \bar{a}_i a_j F^{\beta^*;\beta}(g_i^* \times g_j) \tag{2.35}$$

is nonnegative for arbitrary $g_1, \cdots, g_n \in \mathfrak{F}^\beta$ and arbitrary complex numbers $a_1, \cdots, a_n$ $(n = 1, 2, \cdots)$. If (2.35) is positive for some choice of $g_1, \cdots, g_n \in \mathfrak{F}^\beta$ and $a_1, \cdots, a_n \in \mathbf{C}^1$, then there exists a physical state F_1 for which

$$F_1^\alpha(f) = M^{-1} \sum_{i,j=1}^{n} \bar{a}_i a_j F^{\beta^*;\alpha;\beta}(g_i^* \times f \times g_j) \tag{2.36}$$

for any $\alpha \in \hat{\mathfrak{O}}$, $f \in \mathfrak{F}^\alpha$.

It is easy to see that the following proposition represents in fact an equivalent of Axiom IV.

Proposition 3: For any $F \in \mathcal{S}$ and any $\beta_1, \cdots, \beta_n \in \hat{\mathfrak{O}}$

$$M = \sum_{i,j=1}^{n} \bar{a}_i a_j F^{\beta_i^*;\beta_j}(h_i^* \times h_j) \geq 0 \tag{2.37}$$

for all $h_1 \in \mathfrak{F}^{\beta_1}, \cdots, h_n \in \mathfrak{F}^{\beta_n}$ and all complex numbers $a_1, \cdots, a_n$. If $M > 0$ for some choice of $h_1, \cdots, h_n$ and $a_1, \cdots, a_n$ then there exists an $F_1 \in \mathcal{S}$ for which

$$F_1^\alpha(f) = M^{-1} \sum_{i,j=1}^{n} \bar{a}_i a_j F^{\beta_i^*;\alpha;\beta_j}(h_i^* \times f \times h_j) \tag{2.38}$$

for all $\alpha \in \hat{\mathfrak{O}}$, $f \in \mathfrak{F}^\alpha$. For all $\alpha \in \hat{\mathfrak{O}}$, $f \in \mathfrak{F}^\alpha$ and all $F \in \mathcal{S}$ it is also true that

$$F^\alpha(f) = \overline{F^{\alpha^*}(f^*)}. \tag{2.39}$$

To see that the above proposition is true we have to write in (2.35) and (2.36) $\beta = (\beta_1, \cdots, \beta_n)$ and insert

$$g_{ik} = g_{1k}^{(i)} \times \cdots \times g_{i-1k}^{(i)} \times h_i \times g_{i+1k}^{(i)} \times \cdots \times g_{nk}^{(i)}, \qquad k = 1, 2, \cdots \tag{2.40}$$

instead of g_i into (2.35) and (2.36); here each sequence

$$g_{i1}^{(i)}, g_{i2}^{(i)}, \cdots \in \mathfrak{F}^{\beta_i}, \qquad i, j = 1, \cdots, n \tag{2.41}$$

is taken to be an identity-approximating sequence. We obtain thus (2.37) and (2.38) as a limit from (2.35) and (2.36), respectively, by using Axiom IB.

By choosing $n = 2$, taking $\beta_1 \in \hat{\mathfrak{O}}$ arbitrary, $\beta_2 \in \mathfrak{O}$, and making use of an identity-approximating sequence to replace $h_2 \in \mathfrak{F}^{\beta_2}$, it is easy to derive that

$$\bar{a}_1 a_1 F^{\beta_1^*;\beta_1}(h_1^* \times h_1) + \bar{a}_1 F^{\beta_1^*}(h_1^*) + a_1 F^{\beta_1}(h_1) + 1 \geq 0. \tag{2.42}$$

From this (2.39) follows directly if we denote β_1 with α and h_1 with f.

3. CONCEPT OF A HILBERT-SPACE REPRESENTATION OF THE £ FORMALISM

We now investigate the relation of the introduced £ formalism obeying Axioms I to IV in Sec. 2 to the Hilbert-space formalism. We do that by introducing the concept of the Hilbert-space representation of the £ formalism in a manner similar to that of II, Sec. 1. However, our generalization of the formalism given in I and II has some new peculiarities which have to be analyzed in order to clarify the meaning of a Hilbert-space representation.

First we have to introduce, as in I and II, the set $\mathfrak{O}_b$ of all bounded observables. In this context an observable α is called bounded if its spectrum

$\mathbf{S}^\alpha$ is a bounded set in $\mathbf{R}^\alpha$. If α is a one-dimensional observable then we associate with it the bound

$$||\alpha|| = \sup_{\lambda \in \mathbf{S}^\alpha} |\lambda|. \tag{3.1}$$

If α is an s-dimensional observable then the bound of α cannot be introduced in a meaningful and useful way. Under certain circumstances it might be convenient to denote by $||\alpha||$ the following s-tuple:

$$||\alpha|| = (||\alpha||_{(1)}, \cdots, ||\alpha||_{(s)}), \tag{3.2}$$

where $||\alpha||_{(k)}$, $k = 1, \cdots, s$, is defined as

$$||\alpha||_{(k)} = \sup_{\lambda \in \mathbf{S}^\alpha} |\lambda_k|, \quad \lambda = (\lambda_1, \cdots, \lambda_s) \in \mathbf{R}^\alpha. \tag{3.3}$$

The reason for this definition of $||\alpha||$ can be illustrated with an example. Consider a quantum-mechanical system of n particles and denote by $\chi_{(1)}, \cdots, \chi_{(n)}$ the x coordinate operators corresponding to the particles 1 to n, respectively. If the spectral decompositions of these operators are

$$X_{(1)} = \int \lambda_1 \, dE_{(1)}(\lambda_1), \cdots, X_{(n)} = \int \lambda_n \, dE_{(n)}(\lambda_n), \tag{3.4}$$

then the operator $E_{(k)}(I_k)$ is associated with the probability $\langle \Psi | \, E_{(k)}(I_k) \, | \Psi \rangle$ that, in the physical state represented by the normalized vector Ψ, the k particle has its x coordinate within the interval I_k regardless of the x coordinates of the rest of the particles. The probability that the first, second, etc. particle has its x coordinate within, respectively, $I_1, I_2, \cdots$, is

$$\langle \Psi | \, E_{(1)}(I_1) \cdots E_{(n)}(I_n) \, | \Psi \rangle, \tag{3.5}$$

and is therefore associated with the operator

$$\int_{I_1 \times \cdots \times I_n} \cdots \int \lambda_1 \cdots \lambda_n \, dE_{(1)}(\lambda_1) \cdots dE_{(n)}(\lambda_n). \tag{3.6}$$

It is, however, clear that the norm of the operator (3.6) does not convey any relevant information; it is rather the norms of $\int_{I_k} \lambda_k \, dE_{(k)}(\lambda_k)$, $k = 1, \cdots, n$, that are of significance.

A set $\mathcal{C}$ of observables is called a *set of compatible observables* if any finite subset of $\mathcal{C}$ consists of compatible observables. We denote by $\mathfrak{M}_c$ the family of all maximal sets of compatible bounded observables. In this context we call a set $\mathcal{C}$ of compatible bounded observables maximal if any other set of compatible bounded observables containing $\mathcal{C}$ is identical with $\mathcal{C}$.

The following proposition can now be proved in precisely the same manner as Propositions 1 and 3 in I, Sec. 3.2, was proved.

Proposition 1: Each bounded observable belongs to at least one maximal set of compatible bounded observables and consequently $\mathcal{O}_b = \bigcup_{\mathcal{C} \in \mathfrak{M}_c} \mathcal{C}$.

Due to the fact that we have not included among our axioms the equivalent of Axiom IV in I, Sec. 2.2, we cannot expect that the sets $\mathcal{C}$ can be made in general into normed real algebras. Consequently, we have to devise a new approach to the problem of defining a Hilbert space representation of our $\mathfrak{L}$ formalism. We do that in such a manner as to comply with the customary attitude adopted in quantum field theory.

Definition 1: A smeared observable is, by definition, any ordered pair (α, f) consisting of an observable $\alpha \in \mathcal{O}$, and a test function $f \in \mathcal{F}^\alpha$. The set of all smeared observables

$$(\alpha, f), \qquad \alpha \in \mathcal{O}, \qquad f \in \mathcal{F}^\alpha, \tag{3.7}$$

is denoted by $\mathfrak{z}$.

Definition 2: Two smeared observables (α, f) and (β, g) are called equal if and only if

$$F^{\nu_1 \alpha \nu_2}(f) = F^{\nu_1 \beta \nu_2}(g) \tag{3.8}$$

for all $F \in \mathcal{S}$.

It has to be remarked that a smeared observable is not an observable at all in the customary sense. For example, if f is a complex function, we cannot associate with (α, f), $\alpha \in \mathcal{O}$, a real spectrum. Otherwise, it would be convenient to define physical states on

$$[\alpha_1(f_1), \cdots \alpha_n(f_n)] \in \hat{\mathfrak{z}}, \qquad n = 1, 2, \cdots, \tag{3.9}$$

by writing

$$F^{\alpha_1(f_1), \cdots \alpha_n(f_n)}(g) = F^{\alpha_1, \cdots \alpha_n}[(f_1 \times \cdots \times f_n)g], \qquad g \in \mathcal{F}^{\alpha_1 \times \cdots \times \alpha_n}, \tag{3.10}$$

where $f_1 \times \cdots \times f_n$ is defined by (2.2). Namely due to Axiom IA, $f_1 \times \cdots \times f_n \in \mathcal{F}^{\alpha_1 \times \cdots \times \alpha_n}$ and, therefore,

$$(f_1 \times \cdots \times f_n)g = f_1 g_1 \times \cdots \times f_n g_n \in \mathcal{F}^{\alpha_1 \times \cdots \times \alpha_n}, \quad \lambda_1 \in \mathbf{R}^{\alpha_1}, \cdots, \lambda_n \in \mathbf{R}^{\alpha_n}, \tag{3.11}$$

if $g = g_1 \times \cdots \times g_n$, where $g_1 \in \mathcal{F}^{\alpha_1}, \cdots, g_n \in \mathcal{F}^{\alpha_n}$.

We can associate with $\alpha(f) \in \mathfrak{z}$ an s-dimensional complex Euclidean space $\mathbf{C}^{\alpha(f)} = \mathbf{C}^1 \times \cdots \times \mathbf{C}^1$ (s times) in case that α is an s-dimensional observable.

Definition 3: A point $\lambda \in \mathbf{C}^{\alpha(f)}$ is said to belong to the spectrum $\mathbf{S}^{\alpha(f)}$ of $\alpha(f)$ if for any interval containing λ there exists a test-function with support in that interval for which $F^{\alpha(f)}(g) = F^{\alpha}(fg) \neq 0$.

A smeared observable $\alpha(f)$ is called bounded if its spectrum $\mathbf{S}^{\alpha(f)}$ is a bounded set in $\mathbf{C}^{\alpha(f)}$.

It is obviously true that

Proposition 2: If $g_1, g_2, \cdots, \in \mathfrak{F}^{\alpha_1 \times \cdots \times \alpha_n}$ is an identity-approximating sequence in $\mathfrak{F}^{\alpha_1 \times \cdots \times \alpha_n}$ and $\alpha_1(f_1), \cdots, \alpha_n(f_n)$ are bounded smeared observables then

$$F^{\nu_1 \alpha_1; \cdots; \alpha_n \nu_2}(f_1 \times \cdots \times f_n) = \lim_{k \to \infty} F^{\nu_1 \alpha_1(f_1), \cdots, \alpha_n(f_n) \nu_2}(g_k). \tag{3.12}$$

We denote by $\mathfrak{z}_b$ the set of all bounded smeared observables $\alpha(f)$ corresponding to all $\alpha \in \mathfrak{O}$, $\{\alpha\} = C$, and all $f \in \mathfrak{F}^{\alpha}$ with bounded support in $\mathbf{R}^{\alpha}$.

Proposition 3: If we have for $F_1, F_2 \in \mathcal{S}$

$$F_1^{\alpha}(f) = F_2^{\alpha}(f) \tag{3.13}$$

for all $\alpha(f) \in \hat{\mathfrak{z}}_b$ then $F_1 = F_2$.

Take α to be a single observable. In order to prove the above proposition, we have to make use of a decomposition of the identity of the following form (which exists as is shown in the first volume of Ref. 3):

$$e_1(\lambda) + e_2(\lambda) + \cdots = 1, \qquad \lambda \in \mathbf{R}^{\alpha}, \tag{3.14}$$

where $e_1(\lambda), e_2(\lambda), \cdots$ are arbitrarily many times differentiable functions of compact support, having the additional property that

$$0 \leq e_k(\lambda) \leq 1, \quad \lambda \in \mathbf{R}^{\alpha}, \quad k = 1, 2, \cdots. \tag{3.15}$$

By consulting the type of spaces introduced in Sec. 1 we readily establish that $e_k f \in \mathfrak{F}^{\alpha}$ if $f \in \mathfrak{F}^{\alpha}$. Due to the fact that we can choose

$$\sum_{k=n}^{\infty} e_k(\lambda), \qquad n = 1, 2, \cdots,$$

to constitute a zero-approximating sequence and that the functionals F_1^{α} and F_2^{α} are linear we can write (3.13) in the form

$$\sum_{k=1}^{\infty} F_1^{\alpha}(e_k f) = \sum_{k=1}^{\infty} F_2^{\alpha}(e_k f) \tag{3.16}$$

and the proof is completed for the case $\alpha \in \mathfrak{O}$. The case when α denotes an ordered set of observables can be treated by generalizing the above procedure in a straightforward manner.

By using precisely the same technique we can prove also the following proposition.

Proposition 4: Two smeared observables $\alpha(f), \beta(g)$ are equal if

$$F^{\gamma_1; \alpha; \gamma_2}(h_1 \times f \times h_2) = F^{\gamma_1; \beta; \gamma_2}(h_1 \times g \times h_2) \tag{3.17}$$

for all $F \in \mathcal{S}$ and for all $\gamma_1(h_1), \gamma_2(h_2) \in \hat{\mathfrak{z}}_b$.

Definition 4: We say that a Hilbert space $\mathfrak{K}$ provides a representation for a given $\mathfrak{L}$ language if there exists an injective mapping

$$\alpha(f) = (\alpha, f) \to A_{\alpha}(f), \qquad \alpha(f) \in \hat{\mathfrak{z}}_b, \quad A_{\alpha}(f) \in \mathfrak{B}(\mathfrak{K}), \tag{3.18}$$

of the set $\hat{\mathfrak{z}}_b$ into the C^* algebra $\mathfrak{B}(\mathfrak{K})$ of all bounded operators on $\mathfrak{K}$, as well as an injective mapping

$$F \to \{a\Psi_F, a \in \mathbf{C}^1\}, \qquad F \in \mathcal{S}_0, \quad \Psi_F \in \mathfrak{K}, \quad ||\Psi_F|| = 1, \tag{3.19}$$

of the set $\mathcal{S}_0$ of pure physical states into the set of all rays characterized by normalized vectors $\Psi_F \in \mathfrak{K}$, and having the following properties:

(1) $A_{\alpha}(\bar{f}) = [A_{\alpha}(f)]^*, \qquad \alpha(f) \in \mathfrak{z}_b.$ (3.20)

(2) If $\alpha(f) \in \hat{\mathfrak{z}}_b$ is of the form

$$\alpha(f) = [\alpha_1(f), \cdots, \alpha_m(f_m)], \qquad \alpha_1(f_1), \cdots, \alpha_m(f_m) \in \mathfrak{z}_b \tag{3.21}$$

then $\alpha(f)$ is mapped by (3.18) into the product

$$A_{\alpha_1}(f_1) \cdots A_{\alpha_m}(f_m). \tag{3.22}$$

This mapping is such, that if for some $\beta(g) = [\beta_1(g_1), \cdots, \beta_n(g_n)] \in \hat{\mathfrak{z}}_b$

$$F^{\nu_1 \alpha_1; \cdots; \alpha_m \nu_2}(f_1 \times \cdots \times f_m) = F^{\nu_1 \beta_1; \cdots; \beta_n \nu_2}(g_1 \times \cdots \times g_n), \tag{3.23}$$

then

$$A_{\alpha_1}(f_1) \cdots A_{\alpha_m}(f_m) = A_{\beta_1}(g_1) \cdots A_{\beta_n}(g_n) \tag{3.24}$$

[this having as a consequence that smeared observables corresponding to compatible observables are mapped by (3.18) into commuting operators].

(3) Mapping (3.19) is such that

$$F^{\alpha_1; \cdots; \alpha_n}(f_1 \times \cdots \times f_n) = \langle \Psi_F | A_{\alpha_1}(f_1) \cdots A_{\alpha_n}(f_n) | \Psi_F \rangle \tag{3.25}$$

for any $\alpha_1(f_1), \cdots, \alpha_n(f_n) \in \mathfrak{z}_b$.

(4) The closure of the linear manifold spanned

by the image of $\mathcal{S}_0$ into $\mathfrak{H}$ is identical with the entire Hilbert space $\mathfrak{H}$.

We note that from the physical point of view, it would be sufficient to require in (3) that the linear manifold spanned by the image of $\mathcal{S}_0$ is everywhere dense in $\mathfrak{H}$ in the weak topology induced by (2.24).

4. CONSTRUCTION OF THE B^* ALGEBRA $\mathfrak{A}$

In order to find a Hilbert-space representation of the introduced $\mathfrak{L}$ formalism, we follow the general pattern applied in II to the quantum-mechanical language formulated in I. We give, however, a much simplified version of that construction. This is due on the one hand to the mathematically more convenient formulation in terms of probability functionals instead of complex probability measures (which renders some techanical details trivial) and, on the other hand, to the elimination of some restrictions (discussed in the next section), which are introduced in II.

We begin our construction by considering the set $\hat{\mathfrak{F}}_b$ of finite ordered sets of elements from $\mathfrak{F}_b$, and by introducing equivalence classes in $\hat{\mathfrak{F}}_b$ in the following way.

Rule 1: Two elements $[\alpha_1(f_1), \cdots \alpha_m(f_m)]$ and $[\beta_1(g_1), \cdots \beta_n(g_n)]$ of $\hat{\mathfrak{F}}_b$ are equivalent if and only if

$$F^{\nu_1\alpha_1;\cdots;\alpha_m\nu_s}(f_1 \times \cdots \times f_m) = F^{\nu_1\beta_1;\cdots;\beta_n\nu_s}(g_1 \times \cdots \times g_n) \tag{4.1}$$

for all $F \in \mathcal{S}$.

It is indeed very easy to check that the relation introduced by this rule is an equivalence relation. We denote by $\mathcal{P}$ the set of all the equivalence classes of elements of $\mathfrak{F}_b$.

Consider now the set $\mathbf{C}^1 \times \mathcal{P}$ of ordered pairs (a, ρ) consisting of a complex number a and an element ρ of $\mathcal{P}$. We denote by $\mathfrak{R}$ the set of all equivalence classes of elements of $\mathbf{C}^1 \times \mathcal{P}$ defined by means of an equivalence relation given by the following rule.

Rule 2: Two elements (a, ρ) and (b, σ) of $\mathbf{C}^1 \times \mathcal{P}$, where

$$\rho = [\alpha_1(f_1), \cdots, \alpha_m(f_m)], \tag{4.2}$$

$$\sigma = [\beta_1(g_1), \cdots, \beta_n(g_n)], \tag{4.3}$$

$\alpha(f_1), \cdots, \beta_n(g_n) \in \mathfrak{F}_b$, are equivalent if and only if

$$aF^{\nu_1\alpha_1;\cdots;\alpha_m\nu_s}(f_1 \times \cdots \times f_m) = bF^{\nu_1\beta_1;\cdots;\beta_n\nu_s}(g_1 \times \cdots \times g_n) \tag{4.4}$$

for all $F \in \mathcal{S}$.

We denote the equivalence class containing the element $(a, \rho) \in \mathbf{C}^1 \times \mathcal{P}$ symbolically by $a\rho$. In case that $\rho \in \mathcal{P}$ contains the element $[\alpha_1(f_1), \cdots \alpha_n(f_n)]$ of $\hat{\mathfrak{F}}_b$, we write

$$a\rho = a\alpha_1(f_1) \circ \cdots \circ \alpha_n(f_n) \tag{4.5}$$

to denote the corresponding element of $\mathfrak{R}$. We call expressions of the form (4.5) *ordered products* on $\mathfrak{F}_b$.

Definition 2: The zero equivalence class of $\mathbf{C}^1 \times \mathcal{P}$ is the equivalence class of all $(a, \rho) \in \mathbf{C}^1 \times \mathcal{P}$ for which, if ρ is of form (4.2),

$$aF^{\nu_1\alpha_1;\cdots;\alpha_m\nu_s}(f_1 \times \cdots \times f_m) = 0 \tag{4.6}$$

for all $F \in \mathcal{S}$.

Consider now the family $\tilde{\mathfrak{R}}$ of finite subsets

$$\tilde{p} = \{a_1\rho_1, \cdots, a_n\rho_n\} \tag{4.7}$$

of $\mathfrak{R}$. As each $\tilde{p} \in \tilde{\mathfrak{R}}$ is a finite set, we can write (4.7) in the form

$$\tilde{p} = \{a_1^{(1)}\rho^{(1)}, \cdots, a_{k_1}^{(1)}\rho^{(1)}, \cdots, \times a_1^{(s)}\rho^{(s)}, \cdots, a_{k_s}^{(s)}\rho^{(s)}\}, \tag{4.8}$$

where, in the sense of inequality of elements of $\mathfrak{R}$,

$$1\rho^{(i)} \neq 1\rho^{(j)} \quad \text{for} \quad i \neq j, \quad i, j = 1, \cdots, s. \tag{4.9}$$

The element

$$\tilde{p} = \{a^{(1)}\rho^{(1)}, \cdots, a^{(s)}\rho^{(s)}\}, \quad a^{(1)} = a_1^{(1)} + \cdots + a_{k_1}^{(1)}, \cdots, \quad a^{(s)} = a_1^{(s)} + \cdots + a_{k_s}^{(s)}, \tag{4.10}$$

of $\tilde{\mathfrak{R}}$ is called the *stardand form* of (4.7).

Rule 3: Two elements of $\tilde{\mathfrak{R}}$ are equivalent if and only if their corresponding standard forms are equal elements of $\tilde{\mathfrak{R}}$, i.e., identical finite sets of elements of $\mathfrak{R}$.

It is obvious that Rule 3 defines equivalence classes of elements of $\tilde{\mathfrak{R}}$. We denote by $\mathfrak{A}_1$ the set of all this equivalence classes. If ξ is an element of $\mathfrak{A}_1$ containing an element of $\tilde{\mathfrak{R}}$ of form (4.7), then we write symbolically

$$\xi = a_1\rho_1 \oplus \cdots \oplus a_n\rho_n, \quad \rho_1, \cdots, \rho_n \in \mathfrak{R}. \tag{4.11}$$

Expressions of the form

$$a_1\alpha_{11}(f_{11}) \circ \cdots \circ \alpha_{1k_1}(f_{1k_1}) \oplus \cdots \oplus a_n\alpha_{n1}(f_{n1}) \circ \cdots \circ \alpha_{nk_n}(f_{nk_n}) \tag{4.12}$$

will be called *polynomial forms over* $\mathfrak{F}_b$.

Lemma 1. Operations of addition, multiplication, and multiplication with a c number can be defined

in $\mathfrak{A}_1$ in such a manner that $\mathfrak{A}_1$ becomes an algebra.

The zero element **0** and the identity element **1** of $\mathfrak{A}_1$ are those elements of $\mathfrak{A}_1$ which are equivalence classes of elements of $\tilde{\mathfrak{R}}$ having as a standard form a set of elements of $\mathfrak{R}$ consisting only of, respectively, the zero equivalence class of $\mathbf{C}^1 \times \mathcal{P}$ in case of the zero element, and of any $a\alpha(f) \in \mathbf{C}^1 \times \mathcal{P}$ for which

$$F^{\beta;\alpha;\gamma}(g \times f \times h) = F^{\beta;\gamma}(g \times h), \qquad F^{\alpha}(f) = 1,$$

for all $\beta, \gamma \in \hat{\mathfrak{O}}$, $g \in \mathfrak{F}^{\beta}$, $h \in \mathfrak{F}^{\gamma}$, and all $F \in \mathfrak{S}$—in the case of the identity element.

If $\xi_1, \xi_2 \in \mathfrak{A}_1$ and $a \in \mathbf{C}^1$, where

$$\begin{aligned} \xi_1 &= a_1\rho_1 \oplus \cdots \oplus a_m\rho_m, \\ \xi_2 &= b_1\sigma_1 \oplus \cdots \oplus b_n\sigma_n, \end{aligned} \tag{4.13}$$

$$a_1, \cdots, b_n \in \mathbf{C}^1, \qquad \rho_1, \cdots \sigma_n \in \mathfrak{R},$$

then the definitions are:

Addition. The sum $\xi_1 + \xi_2$ of ξ_1, ξ_2 is defined as

$$\begin{aligned} \xi_1 + \xi_2 = {} & a_1\rho_1 \oplus \cdots \oplus a_m\rho_m \\ & \oplus b_1\sigma_1 \oplus \cdots \oplus b_n\sigma_n. \end{aligned} \tag{4.14}$$

Multiplication. The product $\xi_1\xi_2$ of ξ_1, ξ_2 is defined as

$$\begin{aligned} \xi_1\xi_2 = {} & a_1b_1\rho_1 \circ \sigma_1 \oplus \cdots \oplus a_1b_n\rho_1 \circ \sigma_n \oplus \cdots \\ & \oplus a_mb_1\rho_m \circ \sigma_1 \oplus \cdots \oplus a_mb_n\rho_m \circ \sigma_n, \end{aligned} \tag{4.15}$$

where we introduce the notation

$$\begin{aligned} \rho \circ \sigma &= \alpha_1(f_1) \circ \cdots \circ \alpha_r(f_r) \circ \beta_1(g_1) \circ \cdots \circ \beta_s(g_s), \\ & \alpha_1(f_1), \cdots \beta_s(g_s) \in \mathfrak{z}_b, \end{aligned} \tag{4.16}$$

in case that

$$\begin{aligned} \rho &= \alpha_1(f_1) \circ \cdots \circ \alpha_r(f_r), \\ \sigma &= \beta_1(g_1) \circ \cdots \circ \beta_s(g_s). \end{aligned} \tag{4.17}$$

c-number multiplication. The product $a\xi_1$ of ξ_1 with the complex number a is defined as

$$a\xi_1 = (aa_1)\rho_1 \oplus \cdots \oplus (aa_m)\rho_m. \tag{4.18}$$

The proof of this lemma proceeds in the same way as the proof of Lemma 1 in II, Sec. 2.1, and consequently it is not reproduced.

A * operation can be introduced in the algebra $\mathfrak{A}_1$ by a method similar to that of Sec. 2.2 in II. First, for any element

$$q = [\alpha_1(f_1), \cdots, \alpha_n(f_n)] \tag{4.19}$$

of $\hat{\mathfrak{z}}_b$ we define ($\bar{f} \in \mathfrak{F}^{\alpha}$ denotes the complex conjugate of $f \in \mathfrak{F}^{\alpha}$)

$$q^* = [\alpha_n(\bar{f}_n), \cdots, \alpha_1(\bar{f}_1)] \tag{4.20}$$

and then, proceeding in much the same fashion as in proving Lemmas 3 and 4 in II, Sec. 2.2, we can show:

Lemma 2. A * operation can be introduced in $\mathfrak{A}_1$ in such a fashion that we obtain a * algebra $\mathfrak{A}_2$. This * operation maps the element (of the most general form)

$$\begin{aligned} \xi = {} & a_1\alpha_{11}(f_{11}) \circ \cdots \circ \alpha_{1k_1}(f_{1k_1}) \oplus \cdots \\ & \oplus a_n\alpha_{n1}(f_{n1}) \circ \cdots \circ \alpha_{nk_n}(f_{nk_n}) \end{aligned} \tag{4.21}$$

of $\mathfrak{A}_2$ into the element

$$\begin{aligned} \xi^* = {} & \bar{a}_1\alpha_{1k_1}(\overline{f_{1k_1}}) \circ \cdots \circ \alpha_{11}(\overline{f_{11}}) \oplus \cdots \\ & \oplus \bar{a}_n\alpha_{nk_n}(\overline{f_{nk_n}}) \circ \cdots \circ \alpha_{n1}(\overline{f_{n1}}). \end{aligned} \tag{4.22}$$

As the next step we can introduce in $\mathfrak{A}_2$ a convex functional by means of which we will get out of $\mathfrak{A}_2$ a normed * algebra $\mathfrak{A}_3$. To that purpose we first have to introduce a class of linear functionals on $\mathfrak{A}_2$.

Lemma 3. Corresponding to each $F \in \mathfrak{S}$, a linear normalized functional $\langle\xi\rangle_F$, $\xi \in \mathfrak{A}_2$, can be introduced, which, on an element ξ of $\mathfrak{A}_2$ represented by a polynomial form over $\mathfrak{z}_b$ having the general form (4.21), has the value

$$\begin{aligned} \langle\xi\rangle_F = {} & a_1F^{\alpha_{11};\cdots;\alpha_{1k_1}}(f_{11} \times \cdots \times f_{1k_1}) + \cdots \\ & + a_mF^{\alpha_{n1};\cdots;\alpha_{nk_n}}(f_{n1} \times \cdots \times f_{nk_n}). \end{aligned} \tag{4.23}$$

The proof of this lemma proceeds in the obvious way by showing that the functional

$$\begin{aligned} & \langle[\alpha_1(f_1), \cdots \alpha_n(f_n)]\rangle_F \\ & \qquad = F^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) \end{aligned} \tag{4.24}$$

has the same value on all elements (4.6) of $\mathfrak{z}_b$ belonging to the same class of $\mathcal{P}$. The same is then shown for equivalence classes on $\mathbf{C}^1 \times \mathcal{P}$, i.e., for elements of $\mathfrak{R}$, in relation to the functional

$$\begin{aligned} & \langle a\alpha_1(f_1) \circ \cdots \circ \alpha_n(f_n)\rangle_F \\ & \qquad = aF^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n); \end{aligned} \tag{4.25}$$

it is easy to see that besides the functional (4.25) is equal to zero on the zero equivalence class of $\mathbf{C}^1 \times \mathcal{P}$. By establishing that the functional (4.23) defined on $\tilde{\mathfrak{R}}$ has the same value on any equivalence class of elements from $\mathfrak{R}$ we establish that (4.23) defines a functional on $\tilde{\mathfrak{R}}$ which assigns a unique value to each element of $\tilde{\mathfrak{R}}$, i.e., of $\mathfrak{A}_1$. It is then easy to prove that this functional is linear. The fact that it is normalized, i.e., $\langle\mathbf{1}\rangle_F = 1$, follows directly from (4.23) and the definition of the unit element of $\mathfrak{A}_1$.

Lemma 4. The functional

$$\Phi(\xi) = \sup_{F \in \mathcal{S}} (\langle \xi^* \xi \rangle_F)^{\frac{1}{2}} \tag{4.26}$$

is convex, normalized, and everywhere finite on $\mathfrak{A}_2$. It also has the following properties:

$$\Phi(\xi\eta) \leq \Phi(\xi)\Phi(\eta), \qquad \xi, \eta \in \mathfrak{A}_2, \tag{4.27}$$

$$\Phi(\xi^*) = \Phi(\xi), \qquad \xi \in \mathfrak{A}_2, \tag{4.28}$$

$$\Phi(\xi^*\xi) = [\Phi(\xi)]^2, \qquad \xi \in \mathfrak{A}_2. \tag{4.29}$$

Proof: In order to prove this lemma we first have to establish the fact that the functional (4.26) is everywhere finite. In the course of doing this we also show that the linear functionals $\langle \xi \rangle_F$ are positive functionals for any $F \in \mathcal{S}$, i.e.,

$$\langle \xi^* \xi \rangle_F \geq 0, \qquad \xi \in \mathfrak{A}_2, \tag{4.30}$$

for all $F \in \mathcal{S}$.

For any $\alpha \in \mathfrak{O}$ and any $f \in \mathfrak{F}^\alpha$ satisfying the condition

$$|f(\lambda)| \leq C, \qquad \lambda \in \mathbf{R}^\alpha, \tag{4.31}$$

it follows from Proposition 1, Sec. 2., that

$$0 \leq F^\alpha(\bar{f}f) \leq C^2. \tag{4.32}$$

Now, for any ξ written in the general form (4.21) we have according to Proposition 3, Sec. 2

$$\langle \xi^* \xi \rangle_F = \sum_{i,j=1}^{n} \bar{a}_i a_j F^{\alpha_i^* ; \alpha_j}(f_i^* \times f_j) \geq 0, \tag{4.33}$$

where

$$\begin{aligned} \alpha_i &= \alpha_{i1} \times \cdots \times \alpha_{ik_i}, \\ f_i &= f_{i1} \times \cdots \times f_{ik_i} \in \mathfrak{F}^{\alpha_i}. \end{aligned} \tag{4.34}$$

Consequently $\langle \xi \rangle_F$ is a positive functional and we can apply[19] Cauchy–Schwartz inequality

$$|\langle \xi^* \eta \rangle_F|^2 \leq \langle \xi^* \xi \rangle_F \langle \eta^* \eta \rangle_F, \qquad \xi, \eta \in \mathfrak{A}_2, \tag{4.35}$$

in order to obtain

$$\begin{aligned} &|\langle \alpha_1(f_1) \circ \cdots \circ \alpha_n(f_n) \rangle_F|^2 \\ &\leq \langle \alpha_n(\bar{f}_n) \circ \cdots \circ \alpha_1(\bar{f}_1) \circ \alpha_1(f_1) \circ \cdots \circ \alpha_n(f_n) \rangle_F. \end{aligned} \tag{4.36}$$

We can derive, however, from Axiom IV (see also Hypotheses 2A and 2B in II, Sec. 2.3) that there exists such an $F_1 \in \mathcal{S}$ that

$$\begin{aligned} &\langle \alpha_n(\bar{f}_n) \circ \cdots \circ \alpha_1(\bar{f}_1) \circ \alpha_1(f_1) \circ \cdots \circ \alpha_n(f_n) \rangle_F \\ &\leq \langle \alpha_1(\bar{f}_1) \circ \alpha_1(f_1) \rangle_{F_1} \langle \alpha_n(\bar{f}_n) \circ \cdots \circ \alpha_2(\bar{f}_2) \\ &\quad \circ \alpha_2(f_2) \circ \cdots \circ \alpha_n(f_n) \rangle_F. \end{aligned} \tag{4.37}$$

After $n - 2$ further similar steps we get, by making use also of Axiom IC, Eq. (2.11),

$$\begin{aligned} &\langle \alpha_n(\bar{f}_n) \circ \cdots \circ \alpha_1(\bar{f}_1) \circ \alpha_1(f_1) \circ \cdots \circ \alpha_n(f_n) \rangle_F \\ &\leq \langle \alpha_1(\bar{f}_1) \circ \alpha_1(f_1) \rangle_{F_1} \cdots \langle \alpha_{n-1}(\bar{f}_{n-1}) \\ &\quad \circ \alpha_{n-1}(f_{n-1}) \rangle_{F_{n-1}} \langle \alpha_n(\bar{f}_n) \circ \alpha_n(f_n) \rangle_F \\ &= F_1^{\alpha_1}(\bar{f}_1 f_1) \cdots F_{n-1}^{\alpha_{n-1}}(\bar{f}_{n-1} f_{n-1}) F^{\alpha_n}(\bar{f}_n f_n). \end{aligned} \tag{4.38}$$

Taking into consideration that $\alpha_1(f_1), \cdots, \alpha_n(f_n) \in \mathfrak{d}_b$, we must have

$$\begin{aligned} &|f_1(\lambda_1)| \leq C_1, \cdots, |f_n(\lambda_n)| \leq C_n, \\ &\lambda_1 \in \mathbf{R}^{\alpha_1}, \cdots, \lambda_n \in \mathbf{R}^{\alpha_n}. \end{aligned} \tag{4.39}$$

Applying (4.32), we finally derive from (4.36), (4.37), and (4.38)

$$|\langle \alpha_1(f_1) \circ \cdots \circ \alpha_n(f_n) \rangle_F| \leq C_1 \cdots C_n \tag{4.40}$$

and thus we have that the functional (4.26) is finite at every point $\xi \in \mathfrak{A}_2$.

The fact that $\Phi(1) = 1$ is a direct consequence of the definition (4.23) and of the normalization of all functionals $\langle \xi \rangle_F$, $F \in \mathcal{S}$.

For the proof of (4.27)–(4.29) and of the convexity of this functional, the proofs of the Lemmas 7 and 8 in II, Sec. 2.3, can be consulted. Q.E.D.

We note that the functional (4.26) satisfies all the conditions for being a norm with the exception of the following one: The relation $\Phi(\xi) = 0$ does not imply that ξ is the zero element of $\mathfrak{A}_2$. Therefore, we need the following lemma.

Lemma 5. The set

$$I = \{\xi : \Phi(\xi) = 0, \xi \in \mathfrak{A}_2\} \tag{4.41}$$

is a two-sided ideal of the * algebra $\mathfrak{A}_2$. The set $\mathfrak{A}_3 = \mathfrak{A}_2/I$ of all residue classes modulo I is a * algebra, whose elements are denoted by the letters $X, Y, Z, \cdots$. A norm can be introduced in $\mathfrak{A}_3$ by assigning to each $X \in \mathfrak{A}_3$ the norm

$$||X|| = \Phi(\xi), \quad X \in \mathfrak{A}_3, \quad \xi \in X \cap \mathfrak{A}_2, \tag{4.42}$$

where ξ is any of the elements of $\mathfrak{A}_2$ belonging to the residue class X. With the norm (4.42) $\mathfrak{A}_3$ becomes a normed * algebra.

Proof: In order to prove that I is a two-sided ideal we have to show that I is a linear space and that, if $\xi \in I$ then $\eta\xi \in I$ and $\xi\eta \in I$ for all $\eta \in \mathfrak{A}_2$

[18] Unity-equivalent smeared observables might occur when observables with discrete spectrum, like spin components, are present.

[19] M. A. Naimark, *Normed Rings*, translated by L. F. Boron (P. Noordhoff Ltd., Groningen, The Netherlands, 1959), p. 187.

(besides, I certainly does not coincide with $\mathfrak{A}_2$ because already the identity of $\mathfrak{A}_2$ does not belong to I). As the functional $\Phi(\xi)$ is convex we obtain for any $\xi_1, \xi_2 \in I$ and any complex numbers a_1, a_2.

$$0 \leq \Phi(a_1\xi_1 + a_2\xi_2) \leq |a_1|\,\Phi(\xi_1) + |a_2|\,\Phi(\xi_2) = 0. \tag{4.43}$$

This shows that I is a linear manifold. Furthermore, applying Schwartz–Cauchy inequality (4.35) we get for any $\xi \in I$ and any $\eta \in \mathfrak{A}_2$

$$|\langle(\eta\xi)^*\eta\xi\rangle_F|^2 \leq \langle\xi^*\xi\rangle_F\langle(\eta^*\eta\xi)^*\eta^*\eta\xi\rangle_F = 0. \tag{4.44}$$

This shows that I is a left ideal. To prove that I is also a right ideal we must make use of Proposition 3, Sec. 2. We can write $\xi^*\xi$ and η in the form [consider the general form (4.21) of an element of $\mathfrak{A}_2$]:

$$\xi^*\xi = \sum_{k=1}^{m} a_k\xi_k, \quad \xi_k = \alpha_1^{(k)}(f_1^{(k)}) \circ \cdots \circ \alpha_{m_k}^{(k)}(f_{m_k}^{(k)}), \tag{4.45}$$

$$\eta = \sum_{i=1}^{n} b_i\eta_i, \quad \eta_i = \beta_1^{(i)}(g_1^{(i)}) \circ \cdots \circ \beta_{n_i}^{(i)}(g_{n_i}^{(i)}). \tag{4.46}$$

Introduce the notation

$$\begin{aligned} \alpha_k &= \alpha_1^{(k)} \times \cdots \times \alpha_{m_k}^{(k)}, \\ f_k &= f_1^{(k)} \times \cdots \times f_{m_k}^{(k)} \in \mathfrak{F}^{\alpha_k}, \end{aligned} \tag{4.47}$$

$$\begin{aligned} \beta_i &= \beta_1^{(i)} \times \cdots \times \beta_{n_i}^{(i)}, \\ g_i &= g_1^{(i)} \times \cdots \times g_{n_i}^{(i)} \in \mathfrak{F}^{\beta_i}. \end{aligned} \tag{4.48}$$

We can now state on basis of Proposition 3, Sec. 2, that, in case

$$\langle\eta^*\eta\rangle_F = \sum_{i,j=1}^{n} \bar{b}_i b_j \langle\eta_i^*\eta_j\rangle_F = \sum_{i,j=1}^{n} \bar{b}_i b_j F^{\beta_i^*;\beta_j}(g_i^* \times g_j) \tag{4.49}$$

is greater than zero, there exists such an $F_1 \in \mathfrak{S}$ that

$$\begin{aligned} \langle\eta^*\xi^*\xi\eta\rangle_F &= \sum_{i,j=1}^{n}\sum_{k=1}^{m} \bar{b}_i b_j a_k \langle\eta_i^*\xi_k\eta_j\rangle_F \\ &= \sum_{i,j=1}^{n} \bar{b}_i b_j \sum_{k=1}^{m} a_k F^{\beta_i^*;\alpha_k;\beta_j}(g_i^* \times f_k \times g_j) \\ &= \left[\sum_{i,j=1}^{n} \bar{b}_i b_j F^{\beta_i^*;\beta_j}(g_i^* \times g_j)\right] \\ &\quad \times \sum_{k=1}^{m} a_k F_1^{\alpha_k}(f_k) = \langle\eta^*\eta\rangle_F\langle\xi^*\xi\rangle_{F_1}. \end{aligned} \tag{4.50}$$

As we assumed that $\xi \in I$, we have according to (4.26)

$$\langle\xi^*\xi\rangle_F = 0. \tag{4.51}$$

Consequently $\langle(\xi\eta)^*\xi\eta\rangle_F = 0$. In case that (4.49) is equal to zero and Proposition 3, Sec. 2, cannot be used, we can employ (4.35) to derive

$$|\langle\eta^*\xi^*\xi\eta\rangle_F|^2 \leq \langle\eta^*\eta\rangle_F\langle(\xi^*\xi\eta)^*\xi^*\xi\eta\rangle_F = 0. \tag{4.52}$$

Thus in general we can say that

$$\langle(\xi\eta)^*\xi\eta\rangle_F = 0 \tag{4.53}$$

for all $F \in \mathfrak{S}$. This means that $\Phi(\xi\eta) = 0$ and, therefore, $\xi\eta \in I$, i.e., I is a right ideal too.

We can define now the sum $X + Y$ and product XY of two elements X, Y of $\mathfrak{A}_3$ as the residue class modulo I containing $\xi + \eta$ and $\xi\eta$, respectively, where ξ, η are any elements belonging to the residue classes X, Y, respectively. It is mainly a standard task to show that the residue classes $X + Y$ and XY are independent of the particular choices of $\xi \in X$ and $\eta \in X$. The product aX of $X \in \mathfrak{A}_3$ with a c-number a is defined analogously as the residue class of all $a\xi$ where $\xi \in X$.

Due to the convexity of the functional $\Phi(\xi)$, $\xi \in \mathfrak{A}_2$, the value $||X||$ in (4.42) is also independent of the choice of $\xi \in X$. Namely if $\xi_1, \xi_2 \in X$ then

$$\Phi(\xi_1) \leq \Phi(\xi_1 - \xi_2) + \Phi(\xi_2) = \Phi(\xi_2), \tag{4.54}$$

and by reversing the roles of ξ_1 and ξ_2 in (4.54) we derive that $\Phi(\xi_1) = \Phi(\xi_2)$. Due to the construction of $\mathfrak{A}_3$, $||X|| = 0$ if and only if $X = I$. Thus the set I will play the role of the zero element of $\mathfrak{A}_3$ and as such it is denoted from now on with the symbol $\mathbf{0}$. Q.E.D.

We denote by $\mathfrak{A}$ the B^* algebra which is the completion in the norm (4.42) of the algebra $\mathfrak{A}_3$. We denote the elements of $\mathfrak{A}$ with the letters $x, y, z, \cdots$.

Rule 4: To each $\alpha(f) \in \mathfrak{F}_b$ assign an element $\xi_\alpha(f)$ of $\mathfrak{A}_2$ which contains $1\alpha(f)$ as a standard polynominal form corresponding to the equivalence class of elements of $\tilde{\mathfrak{R}}$ represented by it. The corresponding elements of $\mathfrak{A}_3$ and $\mathfrak{A}$ are denoted by $X_\alpha(f)$ and $x_\alpha(f)$.

We note that

$$\langle\xi_\alpha(f)\rangle_F = F^\alpha(f), \tag{4.55}$$

and that due to Lemma 3 in general

$$\langle\xi_{\alpha_1}(f_1) \cdots \xi_{\alpha_n}(f_n)\rangle_F = F^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n). \tag{4.56}$$

From (4.56) we can derive that the mapping

$$\alpha(f) \to X_\alpha(f), \quad \alpha(f) \in \mathfrak{F}_b, \quad X_\alpha(f) \in \mathfrak{A}_3, \tag{4.57}$$

is an injective mapping of $\mathfrak{z}_b$ into $\mathfrak{A}_3$ or $\mathfrak{A}$, respectively, i.e., that $\alpha(f) \neq \beta(g)$ implies $X_\alpha(f) \neq Y_\beta(g)$. Namely, assume that $X_\alpha(f) = Y_\beta(g)$. This implies that

$$\xi_\alpha(f) - \xi_\beta(g) \in I, \tag{4.58}$$

where I is the ideal (4.41).

Take now arbitrary

$$\begin{aligned} \gamma_1 &= \gamma_1^{(1)} \times \cdots \times \gamma_m^{(1)} \in \hat{\mathfrak{O}}, \\ \gamma_2 &= \gamma_1^{(2)} \times \cdots \times \gamma_n^{(2)} \in \hat{\mathfrak{O}}, \\ h_1 &= h_1^{(1)} \times \cdots \times h_m^{(1)} \in \mathfrak{F}^{\gamma_1}, \\ h_2 &= h_2^{(2)} \times \cdots \times h_n^{(2)} \in \mathfrak{F}^{\gamma_2}, \end{aligned} \tag{4.59}$$

which are such that

$$\gamma_1^{(1)}(h_1^{(1)}), \cdots, \gamma_n^{(2)}(h_n^{(2)}) \in \mathfrak{z}_b. \tag{4.60}$$

Introduce the shorter notation

$$\begin{aligned} \xi_{\gamma_1}(h_1) &= \xi_{\gamma_1(1)}(h_1^{(1)}) \cdots \xi_{\gamma_m(1)}(h_m^{(1)}), \\ \xi_{\gamma_2}(h_2) &= \xi_{\gamma_1(2)}(h_1^{(2)}) \cdots \xi_{\gamma_n(2)}(h_n^{(2)}). \end{aligned} \tag{4.61}$$

According to (4.56) we can write

$$F^{\gamma_1;\alpha;\gamma_2}(h_1 \times f \times h_2) - F^{\gamma_1;\beta;\gamma_2}(h_1 \times g \times h_2)$$

$$= \langle \xi_{\gamma_1}(h_1)\xi_\alpha(f)\xi_{\gamma_2}(h_2)\rangle_F - \langle \xi_{\gamma_1}(h_1)\xi_\beta(g)\xi_{\gamma_2}(h_2)\rangle_F. \tag{4.62}$$

But due to the definition (4.26) and the relation (4.27) and (4.28) we can derive by making use at the first step of the Cauchy–Schwartz inequality (4.35):

$$\begin{aligned} &|\langle \xi_{\gamma_1}(h_1)[\xi_\alpha(f) - \xi_\beta(g)]\xi_{\gamma_2}(h_2)\rangle_F| \\ &\leq (\langle \{[\xi_\alpha(f) - \xi_\beta(g)]\xi_{\gamma_2}(h_2)\}^*[\xi_\alpha(f) - \xi_\beta(g)]\xi_{\gamma_2}(h_2)\rangle_F \\ &\quad \times \langle \xi_{\gamma_1}(h_1)\xi_{\gamma_1}(h_1)^*\rangle_F)^{\frac{1}{2}} \leq \Phi[\xi_\alpha(f) - \xi_\beta(g)] \\ &\quad \times \Phi[\xi_{\gamma_1}(h_1)]\Phi[\xi_{\gamma_2}(h_2)] = 0. \end{aligned} \tag{4.63}$$

The right-hand side of (4.63) is equal to zero because $\Phi[\xi_\alpha(f) - \xi_\beta(g)] = 0$ due to (4.58). From (4.62) and (4.63) we get

$$F^{\gamma_1;\alpha;\gamma_2}(h_1 \times f \times h_2) = F^{\gamma_1;\beta;\gamma_2}(h_1 \times g \times h_2),$$

$$\gamma_1, \gamma_2 \in \hat{\mathfrak{O}}, \quad f \in \mathfrak{F}^{\gamma_1}, \quad g \in \mathfrak{F}^{\gamma_2}, \tag{4.64}$$

and this according to Proposition 5, Sec. 3, means that $\alpha(f) = \beta(g)$. Therefore, the mapping (4.57) is injective and we can state the following.

Theorem 1. There exists a B^* algebra $\mathfrak{A}$ (see Lemmas 1–5, this section) and an injective mapping

$$\alpha(f) \to x_\alpha(f), \qquad \alpha(f) \in \mathfrak{z}_b, \qquad x_\alpha(f) \in \mathfrak{A}, \tag{4.65}$$

of the set $\mathfrak{z}_b$ into $\mathfrak{A}$. The smallest closed * subalgebra of $\mathfrak{A}$ containing the image of $\mathfrak{z}_b$ under the mapping (4.65) is $\mathfrak{A}$ itself.

Each of the functionals $\langle \xi\rangle_F$, $\xi \in \mathfrak{A}_2$, induces a functional $\langle X\rangle_F$, $X \in \mathfrak{A}_3$, on $\mathfrak{A}_3$ which can be defined as

$$\langle X\rangle_F = \langle \xi\rangle_F, \qquad \xi \in X \cap \mathfrak{A}_2. \tag{4.66}$$

We can easily establish that, for given F, (4.66) assigns a unique value to each $X \in \mathfrak{A}_3$. Namely, if $\xi, \eta \in X$ then

$$\begin{aligned} |\langle \xi\rangle_F - \langle \eta\rangle_F|^2 &= \langle \xi - \eta\rangle_F^2 \\ &\leq \langle (\xi - \eta)^*(\xi - \eta)\rangle_F = 0, \end{aligned} \tag{4.67}$$

where the last step is a consequence of the fact that $\xi - \eta \in I$.

We note now that from the Schwartz–Cauchy inequality (4.35) it follows that for any F

$$|\langle X\rangle_F|^2 \leq \langle 1^*1\rangle_F \langle X^*X\rangle_F, \qquad X \in \mathfrak{A}_3. \tag{4.68}$$

Hence each functional $\langle X\rangle_F$, $X \in \mathfrak{A}_3$, is uniquely extendible[20] to a positive bounded linear functional $\langle x\rangle_F$, $x \in \mathfrak{A}$. Consequently, with the help of (4.56) we arrive at the conclusion that the following is true:

Theorem 2. To each $F \in \mathfrak{S}$ can be uniquely assigned a continuous positive linear functional on $\mathfrak{A}$ having the property that

$$\langle x_{\alpha_1}(f_1) \cdots x_{\alpha_n}(f_n)\rangle_F$$

$$= F^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) \tag{4.69}$$

for any $\alpha_1(f_1), \cdots, \alpha_n(f_n) \in \mathfrak{z}_b$.

Theorem 3. The mapping (4.65) can be extended to $\hat{\mathfrak{z}}_b$ by assigning to each $[\alpha(f_1), \cdots \alpha_n(f_n)] \in \hat{\mathfrak{z}}_b$ the element $x_{\alpha_1}(f_1) \cdots x_{\alpha_n}(f_n) \in \mathfrak{A}$. If $\mathfrak{A}_4$ denotes the linear manifold spanned by the image of $\mathfrak{z}_b$, then $\mathfrak{A}_4$ is a * subalgebra of $\mathfrak{A}$ which is everywhere dense in $\mathfrak{A}$ in the norm of $\mathfrak{A}$.

Clearly, the validity of Theorem 3 is due to the construction of $\mathfrak{A} - \mathfrak{A}_4$ being the image in $\mathfrak{A}$ of the algebra $\mathfrak{A}_3$ whose completion is $\mathfrak{A}$ itself.

5. THE HILBERT-SPACE REPRESENTATION OF THE £ FORMALISM

We have seen in the last section that to each physical state F corresponds uniquely a positive continuous functional $\langle x\rangle_F$ on $\mathfrak{A}$. It is easy to see that due to (4.69) and Proposition 3 in Sec. 3, the following lemma is true.

[20] Ref. 19, p. 258.

Lemma 1. There exists an injective mapping

$$F \to \langle x\rangle_F, \qquad F \in \mathfrak{s}, \tag{5.1}$$

of the set $\mathfrak{s}$ into the set $\mathfrak{A}'$ of all continuous linear functionals on $\mathfrak{A}$ which assignes a positive functional to each $F \in \mathfrak{s}$. Each of these functionals is normalized, i.e.,

$$|\langle x\rangle_F| \leq ||x||. \tag{5.2}$$

We denote by $\mathfrak{K}$ the image of $\mathfrak{s}$ in $\mathfrak{A}'$.

Lemma 2. The set $\mathfrak{K}$ is a convex subset of $\mathfrak{A}'$, closed in the weak topology of $\mathfrak{A}'$.

Proof: If we have a sequence $\langle x\rangle_{F_1}, \langle x\rangle_{F_2}, \cdots$ of elements of $\mathfrak{K}$ which is a Cauchy sequence in the weak topology of $\mathfrak{A}'$, then for any $\epsilon > 0$ there exists an $N(\epsilon)$ such that

$$|\langle x\rangle_{F_m} - \langle x\rangle_{F_n}| < \epsilon \quad \text{for} \quad m, n \geq N(\epsilon). \tag{5.3}$$

Inserting any $x_{\alpha_1}(f_1) \cdots x_{\alpha_n}(f_n)$ in (5.3) we get from (4.69)

$$|F_m^{\alpha_1;\cdots;\alpha_k}(f_1 \times \cdots \times f_k) - F_n^{\alpha_1;\cdots;\alpha_k} \times (f_1 \times \cdots \times f_k)| < \epsilon, \qquad m, n \geq N(\epsilon). \tag{5.4}$$

Consequently $F_1, F_2, \cdots$ is a Cauchy sequence in the weak topology of $\mathfrak{s}$; thus according to Axiom IIB there exists a limit $F \in \mathfrak{s}$. Due to Theorem 2, Sec. 4, we have

$$\langle x_{\alpha_1}(f_1) \cdots x_{\alpha_k}(f_k)\rangle_F = \lim_{n\to\infty} \langle x_{\alpha_1}(f_1) \cdots x_{\alpha_n}(f_n)\rangle_{F_n} \tag{5.5}$$

for any $\alpha_1(f_1), \cdots \alpha_k(f_k) \in \mathfrak{d}_b$. The linearity of any functional belonging to $\mathfrak{K}$ in combination with Theorem 3, Sec. 4, allows us to write

$$\langle x\rangle_F = \lim_{n\to\infty} \langle x\rangle_{F_n}, \qquad x \in \mathfrak{A}_4. \tag{5.6}$$

As $\mathfrak{A}_4$ is everywhere dense in $\mathfrak{A}$ in the norm (Theorem 3, Sec. 4) we can choose for any element x of $\mathfrak{A}$ a sequence $x_1, x_2, \cdots \in \mathfrak{A}_4$ converging in the norm to x, i.e., for $\epsilon > 0$:

$$||x - x_k|| < \epsilon, \qquad k \geq n_0(\epsilon). \tag{5.7}$$

Therefore, with the help of (5.2) we get

$$\begin{aligned} |\langle x\rangle_F - \langle x\rangle_{F_n}| &\leq |\langle x\rangle_F - \langle x_k\rangle_F| \\ &\quad + |\langle x_k\rangle_F - \langle x_k\rangle_{F_n}| + |\langle x_k\rangle_{F_n} - \langle x\rangle_F| \\ &\leq 2\,||x - x_k|| + |\langle x_k\rangle_F - \langle x_k\rangle_{F_n}|. \end{aligned} \tag{5.8}$$

For given $\epsilon > 0$ we can choose, according to (5.7), such a k_0 that

$$||x - x_{k_0}|| < \tfrac{1}{4}\epsilon. \tag{5.9}$$

Due to the fact that $x_{k_0} \in \mathfrak{A}_4$ there exists, according to (5.6), such an $N(k_0, \frac{1}{2}\epsilon)$ that

$$|\langle x_{k_0}\rangle_F - \langle x_{k_0}\rangle_{F_n}| < \tfrac{1}{2}\epsilon, \qquad n \geq N(k_0, \tfrac{1}{2}\epsilon), \tag{5.10}$$

Therefore, from (5.8) we get

$$|\langle x\rangle_F - \langle x\rangle_{F_n}| < \epsilon, \qquad n \geq N(k_0, \tfrac{1}{2}\epsilon). \tag{5.11}$$

Consequently we have proved that $\mathfrak{K}$ is closed in the weak topology of $\mathfrak{A}'$.

The convexity of $\mathfrak{K}$ is a consequence of Axiom IIA. Namely, due to Theorem 3, Sec. 4, we can write, by choosing

$$x = x_{\alpha_1}(f_1) \cdots x_{\alpha_n}(f_n), \ \alpha_1(f_1), \cdots \alpha_n(f_n) \in \mathfrak{d}_b, \tag{5.12}$$

for any given $F_1, F_2 \in \mathfrak{s}$ and $0 \leq t \leq 1$:

$$\begin{aligned} t\langle x\rangle_{F_1} + (1-t)\langle x\rangle_{F_2} &= tF_1^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) \\ &\quad + (1-t)F_2^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) \\ &= F^{\alpha_1;\cdots;\alpha_n}(f_1 \times \cdots \times f_n) = \langle x\rangle_F, \end{aligned} \tag{5.13}$$

where F exists according to Axiom IIA. Due to the same theorem relation (5.13) is true for any $x \in \mathfrak{A}$ and therefore, the convexity of $\mathfrak{K}$ is established. Q.E.D.

Applying Lemmas 1 and 2 as well as Krein–Milman theorem[21] we can state

Theorem 1. The set $\mathfrak{K}$ which is the image of $\mathfrak{s}$ in $\mathfrak{A}'$, is a closed subset of the set $\mathfrak{N}$ of all functionals belonging to $\mathfrak{A}'$ with a norm[22] smaller than one. Hence it is a convex, bicompact[23] subset of $\mathfrak{A}'$ in the weak topology of $\mathfrak{A}'$. Consequently $\mathfrak{K}$ contains extremal points and it is the smallest convex closed set (in the weak topology of $\mathfrak{A}'$) containing all the extremal points of $\mathfrak{K}$.

In this context a functional $\mathfrak{f}(x) \in \mathfrak{A}'$ is called an extremal point of a subset S of $\mathfrak{A}'$ if it belongs to S and if

$$\mathfrak{f}(x) = t\mathfrak{f}_1(x) + (1-t)\mathfrak{f}_2(x), \qquad x \in \mathfrak{A}, \tag{5.14}$$

for some $0 \leq t \leq 1$ and $\mathfrak{f}_1, \mathfrak{f}_2 \in S$ implies that $\mathfrak{f}_1(x) \equiv \mathfrak{f}_2(x) \equiv \mathfrak{f}(x)$.

Theorem 2. A functional $\langle x\rangle_F \in \mathfrak{K}$ is an extremal point of $\mathfrak{K}$ if and only if the physical state F it represents is a pure physical state. Each indecomposable on $\mathfrak{K}$ functional $\langle x\rangle_F \in \mathfrak{K}$ is an extremal point of $\mathfrak{K}$ and, therefore, it corresponds to a pure physical state.

[21] Ref. 19, p. 62.

[22] If $\mathfrak{f}(x) \in \mathfrak{A}^1$ then $|\mathfrak{f}(x)| \leq C\,||x||$, $x \in \mathfrak{A}$, for some positive numbers C. The norm $||\mathfrak{f}||$, $\mathfrak{f} \in \mathfrak{A}'$, of $\mathfrak{f}(x)$ is defined as the greatest lower bound of all such C.

[23] Ref. 19, p. 56, Proposition III.

The proof of this theorem is not reproduced. Namely, the first part of the theorem can be proved by the methods used in the proof of Lemma 2; the proof of Proposition 4, Sec. 3.2, in II can be also consulted. We should now mention that a functional $\mathfrak{f}(x)$ belonging to a subset S of positive functionals from $\mathfrak{A}'$ is called *indecomposable on* S if every (positive) functional $\mathfrak{f}_1(x) \in S$ dominated by $\mathfrak{f}(x)$,

$$\mathfrak{f}_1(x^*x) \leq \text{const}\, \mathfrak{f}(x^*x), \tag{5.15}$$

is a multiple of $\mathfrak{f}(x)$, i.e., if (5.15) implies

$$\mathfrak{f}_1(x) = \lambda \mathfrak{f}(x), \qquad x \in \mathfrak{A}, \tag{5.16}$$

for some $\lambda \geq 0$. For the proof of the second part of Theorem 2, the proof of Proposition 5, Sec. 3.2, in II can be consulted. It has to be noticed that due to the fact that we have not postulated the equivalent of Axiom VI, 2 in I, Sec. 2.2, we cannot assert that each extremal point of $\mathfrak{K}$ is a functional indecomposable on $\mathfrak{K}$.

By following precisely the same reasoning as in II, Sec. 3.3, we can prove the following theorem:

Theorem 3. For any given $\mathfrak{L}$ formalism there exists a Hilbert space $\mathfrak{H}$ which provides a Hilbert space representation (Def. 4, Sec. 3) of that $\mathfrak{L}$ formalism.

The above theorem is proved by establishing that there is a complete isomorphism between the algebra $\mathfrak{A}$ and a C^* subalgebra $\mathfrak{A}_0$ of the algebra $\mathfrak{B}(\mathfrak{H})$ of all bounded operators on $\mathfrak{H}$, while the set $\mathfrak{K}_0$ of all extremal functionals of $\mathfrak{K}$ (corresponding to pure physical states) is injectively mapped into the set of all rays on $\mathfrak{H}$. For details the proof of Theorem 3, in II, Sec. 3.3, which is completely analogous in form and content with the above theorem, should be consulted.

Theorem 4. To each one-dimensional observable α can be attached a bounded self-adjoined operator A_α in $\mathfrak{H}$ which is such that

$$||A_\alpha|| = ||\alpha|| \tag{5.17}$$

and for any $\beta(g)$, $\gamma(h) \in \mathfrak{d}_b$

$$F^{\beta;\alpha;\gamma}(g \times f_\lambda \times h) = \langle \Psi_F |\, A_\beta(g) A_\alpha A_\gamma(h)\, | \Psi_F \rangle \tag{5.18}$$

if $f_\lambda \in \mathfrak{F}^\alpha$ is so chosen that[24]

$$f_\lambda(\lambda) = \lambda, \qquad -||\alpha|| \leq \lambda \leq +||\alpha||. \tag{5.19}$$

(Ψ_F is any normalized vector determining the ray on the Hilbert space into which F is mapped.)

Proof: Functions $f_\lambda \in \mathfrak{F}^\alpha$ satisfying condition (5.19) can be found in any of the Gel'fand-type spaces of functions on $\mathbf{R}^1$ introduced in Sec. 1. It follows directly from the definition of the spectrum $\mathbf{S}^\alpha$ of α (Def. 3, Sec. 2) that in case the support of a function $f \in \mathfrak{F}^\alpha$ is contained outside $\mathbf{S}^\alpha$ the smeared observable $\alpha(f)$ is equal (in the sense of Def. 2, Sec. 3) to a zero smeared observable, i.e.,

$$F^{\alpha_1;\alpha_2}(f) = 0, \qquad F \in \mathfrak{s}. \tag{5.20}$$

Consequently $\alpha(f)$ is mapped into the zero element of $\mathfrak{B}(\mathfrak{H})$. Thus, all smeared observables with test functions satisfying (5.19) are equal regardless of the behavior of these test functions outside $\mathbf{S}^\alpha$. Therefore, $A_\alpha(f_\lambda)$ is a uniquely determined operator.

Due to the construction of the algebra $\mathfrak{A}$ and to the * isomorphism between $\mathfrak{A}$ and the C^* algebra $\mathfrak{A}_0$, we get that (5.18) is true and that

$$A_\alpha(f_\lambda^*) = A_\alpha(\bar{f}_\lambda). \tag{5.21}$$

By choosing f_λ to be a real function, i.e., $f_\lambda = \bar{f}_\lambda$, we can deduce that $A_\alpha = A_\alpha(f_\lambda)$ is Hermitian. As A_α is bounded it follows that it is self-adjoint.

According to the construction

$$||A_\alpha||^2 = ||X_\alpha(f_\lambda)||^2 = [\Phi(\xi_\alpha(f_\lambda))]^2 = \sup_{F \in \mathfrak{s}} F^{\alpha;\alpha}(\bar{f}_\lambda \times f_\lambda). \tag{5.22}$$

With the help of Axiom IC and Theorem 2, we get from (5.22)

$$||A_\alpha||^2 = \sup_{F \in \mathfrak{s}_0} F^\alpha(\bar{f}_\lambda f_\lambda). \tag{5.23}$$

We prove that $||A_\alpha|| = ||\alpha||$ by showing first that

$$||A_\alpha|| \leq ||\alpha|| \tag{5.24}$$

and afterwards that

$$||A_\alpha|| \geq ||\alpha||. \tag{5.25}$$

We can easily build an infinitely many times differentiable function $e_\epsilon(\lambda)$ such that

$$0 \leq e_\epsilon(\lambda) \leq 1 \tag{5.26}$$

and

$$e_\epsilon(\lambda) = \begin{cases} 0, & |\lambda| > ||\alpha|| + \epsilon, \\ 1, & |\lambda| \leq ||\alpha||. \end{cases} \tag{5.27}$$

We can choose such an f_λ that

$$|f_\lambda(\lambda)| \leq ||\alpha|| + \epsilon, \qquad |\lambda| \leq ||\alpha|| + \epsilon. \tag{5.28}$$

[24] For the sake of simplicity it is assumed that R_M^α (see footnote 14) does not contain any point belonging to the interval $-||\alpha|| \leq \lambda \leq +||\alpha||$. The theorem can be, however, formulated and proved also when that is not the case, only then the proof is technically somewhat more involved.

We obviously have that $\bar{f}_\lambda f_\lambda e_\epsilon \in \mathfrak{F}^\alpha$ and

$$|f_\lambda(\lambda)|^2 e_\epsilon(\lambda) \leq (||\alpha|| + \epsilon)^2 e_\epsilon(\lambda), \quad \lambda \in \mathbf{R}^\alpha. \tag{5.29}$$

Consequently, due to Axiom ID we can write

$$F^\alpha(\bar{f}_\lambda f_\lambda) = F^\alpha(\bar{f}_\lambda f_\lambda e_\epsilon) \leq (||\alpha|| + \epsilon)^2 F^\alpha(e_\epsilon). \tag{5.30}$$

As F^α is a probability functional we have

$$F^\alpha(e_\epsilon) = 1. \tag{5.31}$$

Thus we get from (5.23)

$$||A_\alpha||^2 \leq (||\alpha|| + \epsilon)^2. \tag{5.32}$$

The construction can be performed for arbitrarily small ϵ; hence (5.24) follows.

Assume now that $||\alpha||$ is a point of the spectrum $\mathbf{S}^\alpha$; in the case of the only other alternative, $-||\alpha|| \in \mathbf{S}^\alpha$, the argument will be completely analogous. For any $\epsilon > 0$ we can find two arbitrarily many-times differentiable functions $e_1(\lambda)$, $e_2(\lambda)$, $\lambda \in \mathbf{R}^\alpha$, having the following properties:

$$0 \leq e_i(\lambda) \leq 1, \qquad \lambda \in \mathbf{R}^\alpha, \qquad i = 1, 2; \tag{5.33}$$

$$\begin{aligned} e_1(\lambda) &= \begin{cases} 0, & \lambda < ||\alpha|| - \epsilon, \quad \lambda > ||\alpha|| + \epsilon, \\ 1, & ||\alpha|| - \tfrac{1}{2}\epsilon \leq \lambda \leq ||\alpha||, \end{cases} \\ e_2(\lambda) &= \begin{cases} 0, & \lambda < \epsilon, \quad \lambda > ||\alpha|| - \tfrac{1}{2}\epsilon, \\ 1, & 0 \leq \lambda \leq ||\alpha|| - \epsilon; \end{cases} \end{aligned} \tag{5.34}$$

$$e_1(\lambda) + e_2(\lambda) = 1, \qquad 0 \leq \lambda \leq +||\alpha||. \tag{5.35}$$

Obviously, we again have $\bar{f}_\lambda f_\lambda e_1$, $\bar{f}_\lambda f_\lambda e_2 \in \mathfrak{F}^\alpha$ and

$$|f_\lambda(\lambda)|^2 e_1(\lambda) \geq (||\alpha|| - \epsilon)^2 e_1(\lambda), \qquad \lambda \in \mathbf{R}^\alpha. \tag{5.36}$$

Consequently we can write for any F

$$\begin{aligned} F^\alpha(\bar{f}_\lambda f_\lambda) &= F^\alpha(\bar{f}_\lambda f_\lambda e_1) + F^\alpha(\bar{f}_\lambda f_\lambda e_2) \\ &\geq F^\alpha(\bar{f}_\lambda f_\lambda e_1) \geq (||\alpha|| - \epsilon)^2 F^\alpha(e_1). \end{aligned} \tag{5.37}$$

According to Axiom IIIB there always exists such an $F_1 \in \mathcal{S}$ that $F_0^\alpha(e_1) = 1$. Due to Theorem 2 there should also exist such an $F_0 \in \mathcal{S}_0$ that $F_0^\alpha(e_1) = 1$. Therefore, from (5.23) we get

$$||A_\alpha||^2 \geq (||\alpha|| - \epsilon)^2. \tag{5.38}$$

The whole procedure can be carried through with an arbitrarily small ϵ and, therefore, (5.25) follows. Q.E.D.

This theorem cannot be extended in general to s-dimensional observables ($s \geq 2$) with bounded spectrum, because the space $\mathfrak{F}^\alpha$ can be, under realistic conditions, restricted in many respects so that we cannot construct *a priori* generalizations of test functions fulfilling some kind of generalization of the conditions (5.19).

Theorem 4 cannot be extended to the general case of one-dimensional unbounded observables either if no additional assumptions are introduced. One such additional assumption would consist in requiring that the sequence

$$F^\alpha[\lambda^2 f_1(\lambda)], \qquad F^\alpha(\lambda^2 f_2(\lambda)], \cdots \tag{5.39}$$

is bounded for any $F \in \mathcal{S}_0$ and any identity-approximating sequence $f_1, f_2, \cdots \in \mathfrak{F}^\alpha$.

We see that in the most general case each observable α of the original formalism is not represented by a single self-adjoined operator but rather by an entire family of bounded operators $A_\alpha(f)$ associated with each smeared observable $\alpha(f) \in \mathfrak{B}_b$ which corresponds to α.

Ordering Theorems*

F. ROHRLICH AND J. G. WRAY†
Department of Physics, Syracuse University, Syracuse, New York
(Received 16 February 1966)

Various theorems related especially to time-ordered products are proven. Applications to quantum field theory include particularly simple derivations of the off-mass-shell extensions of the S-operator and of the interpolating field. The latter are a generalization of the results of Glaser, Lehmann, and Zimmermann.

1. ORDERING AND ORDERED PRODUCTS

THE idea of operator ordering in quantum field theory is not new[1] and has been used extensively since the work of Wick[2] made it generally known. However, due to the almost exclusive use of perturbation expansions, operator ordering has been carried out mainly with free fields.[3] While certain orderings, such as normal ordering, are only defined for such fields, other orderings are not. The theorems below will refer to orderings (especially time ordering) of arbitrary local operators.

Given a product of n operators, the ordered product of these operators under a given ordering must be defined. We define an operator O by

$$O(A_1B_2 \cdots Z_n) = (A_1B_2 \cdots Z_n)_{\text{ordered}}, \tag{1.1}$$

with two important properties. First, it is idempotent,

$$O^2 = O, \tag{1.2}$$

which is an obvious property. Secondly, it has a symmetry property. For spin-zero fields the left side of (1.1) is symmetric in the permutation of the operators, for spin-half fields it is antisymmetric. More precisely, if each factor in the product transforms according to some irreducible representation of the Lorentz group $D(\frac{1}{2}m, \frac{1}{2}n)$ then the left side of (1.1) changes sign under an interchange of two operators for each of which $m + n$ is odd (Fermi operators). Otherwise the left side remains unaffected.

If the n operators in the product are completely unrelated, the operator O is linear. But one is usually interested in operators for which commutation relations hold and possibly other equations. In that case linearity in general does not hold. As a particular consequence of this we note that an equation does not remain valid (in general) under the operator O.

The operator for normal ordering will be denoted by N, the normal ordered product by $: \cdots :$ in the notation of Wick. Its definition is well known.[2] Time ordering, however, requires some comments.

The operators T_+ and T_- indicate positive and negative time-ordering, i.e., nondecreasing time from right to left and from left to right, respectively. If $A_1, A_2, \cdots, A_n$ are n different time-dependent operators, then

$$T_+(A_1 \cdots A_n) = \delta_p A_{i_1} \cdots A_{i_n}, \quad t_{i_1} > t_{i_2} > \cdots > t_{i_n}, \tag{1.3}$$

$$T_-(A_1 \cdots A_n) = \delta'_p A_{i_1} \cdots A_{i_n}, \quad t_{i_1} < t_{i_2} < \cdots < t_{i_n}. \tag{1.4}$$

δ_p and δ'_p are the sign factors corresponding to the symmetry of the left sides (see above) under the permutations

$$\begin{pmatrix} 1 & \cdots & n \\ i_1 & \cdots & i_n \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & \cdots & n \\ j_1 & \cdots & j_n \end{pmatrix}$$

respectively.

When two or more operators occur at equal times their order is unspecified and seems to be essentially irrelevant, because it will be assumed throughout that two operators at equal times and different space points commute with each other unless they are both Fermi operators (in which case they anticommute). The sign arising in the latter case is included in δ_p and δ'_p. Not so, however, for equal-space points.

The correct expression of a time-ordered product[4] is then obtained by use of the step function $\theta(x) =$

* Work supported in part by a grant from the National Science Foundation.
† NASA Fellow.
[1] A. Houriet and A. Kind, Helv. Phys. Acta **22**, 319 (1949).
[2] G. C. Wick, Phys. Rev. **80**, 268 (1950).
[3] We shall use the word "operator" loosely here including operator valued distributions.
[4] The following definition is necessary only for time-ordered products of operators and is partly redundant for operator distributions.

1, $\frac{1}{2}$, 0 for $x^0 > 0$, $x^0 = 0$, $x^0 < 0$ and of the corresponding partition of unity,

$$1 = \theta_{x1} + \theta_{1x} = \theta_{x1} + \theta_{1x}(\theta_{x2} + \theta_{2x})$$
$$= \theta_{x1} + \theta_{1x2} + \theta_{1x}\theta_{2x}(\theta_{x3} + \theta_{3x}) = \cdots, \quad (1.5)$$

where we used the notation

$$\theta_{ij} \equiv \theta(x_i - x_j), \qquad \theta_{ijk} \equiv \theta_{ij}\theta_{jk}. \quad (1.6)$$

This partition is used as follows. For two operators $A_1 \equiv A(t_1)$ and $B_2 \equiv B(t_2)$ which we shall assume to be Bose operators for the sake of example,

$$T_+(A_1B_2) \equiv \theta_{12}A_1B_2 + \theta_{21}B_2A_1, \quad (1.7)$$

$$T_-(A_1B_2) \equiv \theta_{12}B_2A_1 + \theta_{21}A_1B_2. \quad (1.8)$$

Given a product such as A_1B_2 on the right of one of these equations, a third operator C_X is to be inserted in the place appropriate to the time ordering using (1.5). For $t_1 \geq t_2$, for example,

$$T'_+(A_1B_2C_X)$$
$$= \theta_{x1}C_XA_1B_2 + \theta_{1x2}A_1C_XB_2 + \theta_{1x}\theta_{2x}A_1B_2C_X.$$

Thus, for arbitrary times

$$T'_+(A_1B_2C_3) = \theta_{12}(\theta_{31}C_3A_1B_2$$
$$+ \theta_{132}A_1C_3B_2 + \theta_{13}\theta_{23}A_1B_2C_3) + (1 \rightleftharpoons 2).$$

The last symbol indicates repetition of the previous term with 1 and 2 interchanged. But this expression is not yet symmetrical in 1, 2, and 3. Symmetrization yields

$$T_+(A_1B_2C_3) = \frac{1}{3!}\sum_{\text{perm}} T'_+(A_1B_2C_3)$$
$$= \sum_{\text{perm}} \bar{\theta}_{123}A_1B_2C_3, \quad (1.9a)$$

where

$$\bar{\theta}_{123} \equiv \tfrac{1}{3}\theta_{123}(1 + 2\theta_{13}). \quad (1.9b)$$

We note that (1.9) is *not* the sum of all permutations of $\theta_{123}A_1B_2C_3$, but differs from it when the t_i are equal.

The partition of unity (1.5) permits one to insert an additional operator into a T_+ -product of $n - 1$ operators, yielding after symmetrization a T_+ -product of n operators. In this way time-ordered products of any number of operators can be constructed.

The T-product of n continuous operator functions is a discontinuous operator function because of the step functions. But even in this case the derivative of the T-product is to be regarded as a distribution. This suggests that the T-product and $\theta(x)$ should be regarded as distributions to begin with. While this is possible, it does not invalidate the approach adopted here. After all, a distribution is a generalization of the function concept. In particular, the definition of $\theta(0)$ in connection with the specification of $\theta(x)$ as a function does not prevent one to regard $\theta(x)$ as a distribution. It only makes the value of $\theta(0)$ irrelevant. Thus, we have for example

$$\theta_{13}\theta_{123}|_{\text{function}} \rightarrow \theta_{123}|_{\text{distribution}}. \quad (1.10)$$

Consequently, as a distribution, the T-product can be defined as, e.g.,

$$T_+(A_1B_2C_3) = \sum_{\text{perm}} \theta_{123}A_1B_2C_3. \quad (1.11)$$

Indeed, (1.9) becomes (1.11) because of (1.10).

We adopt our approach in terms of functions rather than distributions for the ordered products for two reasons. Firstly, many of the equations we derive below are of a purely algebraic nature and do not involve derivatives, and secondly, there is little point in regarding the θ-function as a distribution if its product with the operator distributions $A_1 \cdots Z_n$ remains undefined in a distribution-theoretic sense.

For the purpose of applications it will sometimes be necessary to consider the expansion of an operator in terms of time-ordered products of free fields. The latter are denoted by lower case letters and, unless otherwise noted, will always refer to the in-fields; their time-ordered products are denoted by a subscript:

$$T_\pm(a_1 \cdots z_n) \equiv (a_1 \cdots z_n)_\pm. \quad (1.12)$$

Thus, for a neutral scalar field $a(x)$, an operator functional of it would be

$$F = \sum_{n=0}^{\infty} \frac{1}{n!}\int f_n^{(\pm)}(x_1 \cdots x_n)(a_1 \cdots a_n)_\pm \, dx_1 \cdots dx_n, \quad (1.13)$$

where dx_i indicates integration over four-dimensional Minkowski space and $f_n^{(\pm)}(x_1 \cdots x_2)$ is a (generalized) function symmetric in its arguments.

The expansion (1.13) can be used to define the quantity $(a(x)F)_\pm$ by

$$(a(x)F)_\pm \equiv \sum_{n=0}^{\infty} \frac{1}{n!}\int f_n^{(\pm)}(x_1 \cdots x_n)$$
$$\times (a(x)a_1 \cdots a_n)_\pm \, dx_1 \cdots dx_n. \quad (1.14)$$

The quantities $f_n^{(\pm)}(x_1 \cdots x_n)$ in (1.13) can be related to the operator derivatives[5] of the operator

[5] F. Rohrlich, J. Math. Phys. 5, 324 (1964).

functional F as follows. We know that in the N-product expansion

$$F = \sum_{n=0}^{\infty} \frac{1}{n!} \int f_n(x_1 \cdots x_n) :a_1 \cdots a_n: dx_1 \cdots dx_n, \tag{1.15}$$

$$f_n(x_1 \cdots x_n) = \langle \delta^n F/(\delta x_1 \cdots \delta x_n)\rangle_0, \tag{1.16}$$

where we use the notation

$$\delta F/\delta x \equiv \delta F/\delta a(x). \tag{1.17}$$

The Wick theorem relating free T-products and N-products plus the symmetry of $f_n(x_1 \cdots x_n)$ gives

$$:a_1 \cdots a_n: = \sum_{m=\min}^{n} \frac{1}{m!} (a_1 \cdots a_m)_{\pm} \times C^{(\pm)}_{m,n}(x_{m+1} \cdots x_n) \tag{1.18}$$

with

$$C^{(\pm)}_{n,n} = n!, \tag{1.19a}$$

and[6]

$$C^{(\pm)}_{m,n}(x_{m+1} \cdots x_n) = \frac{n!}{(n-m)!} (\pm i)^{\frac{1}{2}(n-m)} \times \prod_{1A}^{*} \Delta_{1R}(x_{m+1} - x_{m+2}) \cdots \Delta_{1R}(x_{n-1} - x_n). \tag{1.19b}$$

Substitution into (1.15) and rearrangement of the double sum yields

$$f^{(\pm)}_m(x_1 \cdots x_m) = \sum_{l=0,2,\cdots}^{\infty} \frac{1}{(m+l)!} \int dx_{m+1} \cdots dx_{m+l} \times C^{(\pm)}_{m,m+l}(x_{m+1} \cdots x_{m+l}) f_{m+l}(x_1 \cdots x_{m+l}). \tag{1.20}$$

In Sec. 2 we derive the fundamental reduction formula for T-products and apply it to the derivation of the analog to the Matthews–Salam equations[7] as they arise in asymptotic quantum field theory. The exact relation of these equations with those obtained by Matthews and Salam will be discussed elsewhere.

In Sec. 3 the reduction formula for T-products is generalized to the factorization formula for T-products. As a by-product, the relation between retarded commutators and retarded products is obtained and applied to the derivation of the Glaser–Lehmann–Zimmermann result.

In Sec. 4 a "pseudo-mapping" in terms of T-products is derived for operators which are related in the way in which interpolating and asymptotic free fields are related,

$$A(x) = S^*[a(x)S]_+, \tag{1.21}$$

where S is unitary. This important theorem can be combined with the general expression for multiple operator derivatives in terms of T-products (Sec. 5), to yield a simple derivation of the off-mass shell behavior of any (multilocal) functional and in particular of the S-operator in quantum field theory.

Throughout this paper we shall restrict ourselves to the neutral scalar field whenever the transformation properties of the operators become relevant. Generalizations to charged fields, to higher spins, and to the massless case will be reported in a future publication.

2. REDUCTION OF T-PRODUCTS

Let $B_2 \cdots Z_n$ be a product of $n-1$ Bose operators, defined at points in time, $t_2, t_3 \cdots$ respectively, which is T_+ ordered:

$$T_+(B_2 \cdots Z_n) = B_2 \cdots Z_n \quad \text{for} \quad t_2 > t_3 > \cdots > t_n. \tag{2.1}$$

How is it related to a T_+ product of n factors? It is easy to see that

$$T_+(A_1 B_2 \cdots Z_n) = \theta_{12} A_1 B_2 \cdots Z_n + \theta_{213} B_2 A_1 C_3 \cdots Z_n + \theta_{314} B_2 C_3 A_1 D_4 \cdots Z_n + \cdots + \theta_{n1} B_2 \cdots Z_n A_1. \tag{2.2}$$

It is further noted that, subject to (2.1),

$$\theta_{213} = \theta_{21}(\theta_{13} + \theta_{31}) - \theta_{21}\theta_{31} = \theta_{21} - \theta_{21}\theta_{31} \underset{t_a > t_a}{\rightarrow} \theta_{21} - \theta_{31},$$

and in general,

$$\theta_{1,1,l+1} \underset{t_l > t_{l+1}}{\rightarrow} \theta_{l1} - \theta_{l+1,1}. \tag{2.3}$$

With (2.5) Eq. (2.2) becomes

$$\begin{aligned} T_+(A_1 \cdots Z_n) &= [(\theta_{12} + \theta_{21})A_1 B_2 \cdots Z_n - \theta_{21} A_1 B_2 \cdots Z_n] \\ &+ (\theta_{21} - \theta_{31}) B_2 A_1 C_3 \cdots Z_n \\ &+ (\theta_{31} - \theta_{41}) B_2 C_3 A_1 D_4 \cdots Z_n + \cdots \\ &+ (\theta_{n-1,1} - \theta_{n1}) B_2 C_3 \cdots Y_{n-1} A_1 Z_n \\ &+ \theta_{n1} B_2 \cdots Z_n A_1, \end{aligned} \tag{2.4}$$

[6] For notation see J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons* (Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1959, second printing with corrections).

[7] P. T. Matthews and A. Salam, Proc. Roy. Soc. (London) **A221**, 128 (1954).

which upon rearranging terms results in[8]

$$T_+(A_1 \cdots Z_n) = A_1 B_2 \cdots Z_n + [B_2, A_1]_R C_3 \cdots Z_n + B_2[C_3, A_1]_R D_4 \cdots Z_n + \cdots + B_2 \cdots Y_{n-1}[Z_n, A_1]_R$$

or

$$T_+(A_1 \cdots Z_n) = A_1 \cdots Z_n + [B_2 \cdots Z_n, A_1]_R. \quad (2.5)$$

We see now that (2.5) already contains the most general case, including the possibility of equal times, as can easily be checked. In particular, any of the explicit operators $A_1, B_2, \cdots, Z_n$ could be operator products at equal time.

Furthermore since any T_+ product can be reduced to an expression of the form (2.1), the result (2.5) also proves the equation

$$T_+(A_1 B_2 \cdots Z_n) = A_1 T_+(B_2 \cdots Z_n) + [T_+(B_2 \cdots Z_n), A_1]_R. \quad (2.6)$$

Variations on (2.6) are summarized below:

$$T_\pm(A_1 \cdots Z_n) = A_1 T_\pm(B_2 \cdots Z_n) \pm \underset{A}{[T_\pm(B_2 \cdots Z_n), A_1]_R}, \quad (2.7)$$

$$T_\pm(A_1 \cdots Z_n) = T_\pm(B_2 \cdots Z_n) A_1 \pm \underset{A}{[A_1, T_\pm(B_2 \cdots Z_n)]_R}, \quad (2.8)$$

where

$$\underset{A}{[T_\pm(B_2 \cdots Z_n), A_1]_R} = \underset{R}{[A_1, T_\pm(B_2 \cdots Z_n)]_A}. \quad (2.9)$$

Their proofs can easily be supplied by the reader.

Eqs. (2.7) and (2.8) imply the interesting equalities

$$\underset{A}{[[B_2, T_\pm(C_3 \cdots Z_n)]_R}, \underset{A}{A_1]_R} = [B_2, \underset{A\,A}{[T_\pm(C_3 \cdots Z_n), A_1]_R}]_R \pm \underset{A}{[[B_2, A_1]_R}, T_\pm(C_3 \cdots Z_n)]. \quad (2.10)$$

They can be proven as follows:

$$\begin{aligned} T_+(A_1 \cdots Z_n) &= A_1 T_+(B_2 \cdots Z_n) + [T_+(B_2 \cdots Z_n), A_1]_R \\ &= A_1\{T_+(C_3 \cdots Z_n)B_2 + [B_2, T_+(C_3 \cdots Z_n)]_R\} \\ &\quad + [\{T_+(C_3 \cdots Z_n)B_2 + [B_2, T_+(C_3 \cdots Z_n)]_R\}, A_1]_R \\ &= A_1 T_+(C_3 \cdots Z_n)B_2 + [B_2, A_1 T_+(C_3 \cdots Z_n)]_R \\ &\quad - [B_2, A_1]_R T_+(C_3 \cdots Z_n) \\ &\quad + [T_+(C_3 \cdots Z_n), A_1]_R B_2 + T_+(C_3 \cdots Z_n)[B_2, A_1]_R \\ &\quad + [[B_2, T_+(C_3 \cdots Z_n)]_R, A_1]_R. \end{aligned} \quad (2.11)$$

We also have

$$\begin{aligned} T_+(A_1 B_2 \cdots Z_n) &= T_+(A_1 C_3 \cdots Z_n)B_2 + [B_2, T_+(A_1 C_3 \cdots Z_n)]_R \\ &= T_+(A_1 C_3 \cdots Z_n)B_2 + [B_2, A_1 T_+(C_3 \cdots Z_n)]_R \\ &\quad + [B_2, [T_+(C_3 \cdots Z_n), A_1]_R]_R. \end{aligned} \quad (2.12)$$

Now comparison of terms in (2.11) and (2.12) yields Eq. (2.10) with the upper subscripts,

$$[B_2, [T_+(C_3 \cdots Z_n), A_1]_R]_R = [[B_2, T_+(C_3 \cdots Z_n)]_R, A_1]_R - [[B_2, A_1]_R, T_+(C_3 \cdots Z_n)]. \quad (2.13)$$

The proof for the T_--product follows in the same manner.

As a simple application of Eq. (2.7), consider the operator derivative of the T_+ -product of the interpolating neutral scalar field. That field satisfies[9,10]

$$i(\delta A_k/\delta x) = \mathfrak{N} K_x[A_k, A_x]_R, \quad (2.14)$$

where the operator[11] $\mathfrak{N}$ eliminates all terms with factors of the form $K_i\delta(x_i - x_j)$. Therefore,

$$\begin{aligned} (i\delta/\delta x)T_+(A_1 \cdots A_n) &= \sum_{k=1}^{n} T_+[A_1 \cdots i(\delta A_k/\delta x) \cdots A_n) \\ &= \mathfrak{N} K_x[T_+(A_1 \cdots A_n), A_x]_R \\ &= \mathfrak{N} K_x T_+(A_x A_1 \cdots A_n) - \mathfrak{N} J_x T_+(A_1 \cdots A_n). \end{aligned} \quad (2.15)$$

We used (2.7) and $K_x A_x = J_x$. The relation[12] (2.15) is the analog of the Matthews–Salam equations[7] in asymptotic Eq. field theory.

3. THE FACTORIZATION OF T-PRODUCTS

The generalization of (2.6) is

$$T_+(A_1 \cdots Z_n) = T_+(A_1 \cdots F_l)T_+(G_{l+1} \cdots Z_n) + [T_+(G_{l+1} \cdots Z_n), T_+(A_1 \cdots F_l)]_R, \quad (3.1)$$

[8] The retarded commutator $[A_1, B_2]_R \equiv \theta_{12}[A_1, B_2]$. It is a special case of the inhomogeneous commutator of Ref. 5.

[9] R. E. Pugh, J. Math. Phys. 6, 740 (1965).

[10] T. W. Chen, F. Rohrlich, and M. Wilner, J. Math. Phys. 7, 1365 (1966).

[11] R. E Pugh, J. Math. Phys. 7, 376 (1966).

[12] This relation was first obtained by T. W. Chen by a different method, Ph.D. thesis, Syracuse University (1966); see also Syracuse University preprint GP-4935, SU 66-01.

where[13]

$$[T_+(G_{l+1}\cdots Z_n), T_+(A_1\cdots F_l)]_R = \sum_{i=1}^{l} T_+(A_1\cdots \Lambda_i\cdots F_l)[T_+(G_{l+1}\cdots Z_n), C_i]_R$$
$$+ \sum_{i<j=1}^{l} T_+(A_1\cdots \Lambda_{ij}\cdots F_l)[[T_+(G_{l+1}\cdots Z_n), C_i]_R, D_j]_R$$
$$\vdots$$
$$+ [\cdots [T_+(G_{l+1}\cdots Z_n), A_1]_R, B_2]_R, \cdots, F_l]_R. \tag{3.2}$$

The consistency of Eq. (3.2) with our previous definitions is demonstrated in Appendix A. The proof of (3.1) follows by induction. For $l = 1$, (3.2) becomes an identity and (3.1) reduces to (2.6) and is therefore satisfied. Assume it now for l and prove it for $l + 1$.

$$T_+(A_1\cdots Z_n) = A_1T_+(B_2\cdots Z_n) + [T_+(B_2\cdots Z_n), A_1]_R$$
$$= A_1\{T_+(B_2\cdots G_{l+1})T_+(H_{l+2}\cdots Z_n) + [T_+(H_{l+2}\cdots Z_n), T_+(B_2\cdots G_{l+1})]_R\}$$
$$+ [\{T_+(B_2\cdots G_{l+1})T_+(H_{l+2}\cdots Z_n) + [T_+(H_{l+2}\cdots Z_n), T_+(B_2\cdots G_{l+1})]_R\}, A_1]_R, \tag{3.3}$$

$$T_+(A_1\cdots Z_n) = A_1T_+(B_2\cdots G_{l+1})T_+(H_{l+2}\cdots Z_n) + A_1[T_+(H_{l+2}\cdots Z_n), T_+(B_2\cdots G_{l+1})]_R$$
$$+ [T_+(B_2\cdots G_{l+1}), A_1]_RT_+(H_{l+2}\cdots Z_n) + T_+(B_2\cdots G_{l+1})[T_+(H_{l+2}\cdots Z_n), A_1]_R$$
$$+ [[T_+(H_{l+2}\cdots Z_n), T_+(B_2\cdots G_{l+1})]_R, A_1]_R \tag{3.4}$$
$$= T_+(A_1\cdots G_{l+1})T_+(H_{l+2}\cdots Z_n) + [T_+(H_{l+2}\cdots Z_n), T_+(A_1\cdots G_{l+1})]_R. \tag{3.5}$$

The last relation is a consequence of the following

$$[[T_+(H_{l+2}\cdots Z_n), T_+(B_2\cdots G_{l+1})]_R, A_1]_R$$
$$= \sum_{i=2}^{l+1} [T_+(B_2\cdots \Lambda_i\cdots G_{l+1})[T_+(H_{l+2}\cdots Z_n), C_i]_R, A_1]_R$$
$$+ \sum_{i<j=2}^{l+1} [T_+(B_2\cdots \Lambda_{ij}\cdots G_{l+1})[[T_+(H_{l+2}\cdots Z_n), C_i]_R, D_j]_R, A_1]_R$$
$$\vdots$$
$$+ [\cdots [T_+(H_{l+2}\cdots Z_n), B_2]_R, C_3]_R, \cdots G_{l+1}]_R, A_1]_R$$
$$= \sum_{i=2}^{l+1} \{[T_+(B_2\cdots \Lambda_i\cdots G_{l+1}), A_1]_R[T_+(H_{l+2}\cdots Z_n), C_i]_R$$
$$+ T_+(B_2\cdots \Lambda_i\cdots G_{l+1})[[T_+(H_{l+2}\cdots Z_n), C_i]_R, A_1]_R\} + \cdots$$
$$= \sum_{i=2}^{l+1} \{[T_+(A_1B_2\cdots \Lambda_i\cdots G_{l+1}) - A_1T_+(B_2\cdots \Lambda_i\cdots G_{l+1})][T_+(H_{l+2}\cdots Z_n), C_i]_R$$
$$+ T_+(B_2\cdots \Lambda_i\cdots G_{l+1})[[T_+(H_{l+2}\cdots Z_n), C_i]_R, A_1]_R\} + \cdots$$
$$= [T_+(H_{l+2}\cdots Z_n), T_+(A_1\cdots G_{l+1})]_R - T_+(B_2\cdots G_{l+1})[T_+(H_{l+2}\cdots Z_n), A_1]_R$$
$$- A_1[T_+(H_{l+2}\cdots Z_n), T_+(B_2\cdots G_{l+1})]_R. \tag{3.6}$$

When combined with (3.4) this yields (3.5). This completes the proof of (3.1).

In the same manner factorization can be carried out for the T_--product,

$$T_-(A_1\cdots Z_n) = T_-(A_1\cdots F_l)T_-(G_{l+1}\cdots Z_n) - [T_-(G_{l+1}\cdots Z_n), T_-(A_1\cdots F_l)]_A, \tag{3.7}$$

[13] The letter Λ_{ijk} indicates missing operators at the positions $ijk\cdots$.

where

$$[T_-(G_{l+1}\cdots Z_n), T_-(A_1\cdots F_l)]_A = +\sum_{i=1}^{l} T_-(A_1\cdots \Lambda_i\cdots F_l)[T_-(G_{l+1}\cdots Z_n), C_i]_A$$
$$+(-)\sum_{i<j=1}^{l} T_-(A_1\cdots \Lambda_{ij}\cdots F_l)[[T_-(G_{l+1}\cdots Z_n), C_i]_A, D_i]_A$$
$$\vdots$$
$$+(-)^{l-1}[\cdots[T_-(G_{l+1}\cdots Z_n), A_1]_A, B_2]_A, \cdots F_l]_A. \tag{3.8}$$

Alternatives to (3.2) and (3.8) are given by the following:

$$[T_\pm(G_{l+1}\cdots Z_n), T_\pm(A_1\cdots F_l)]_{\substack{R\\A}} = +\sum_{i=l+1}^{n} [C_i, T_\pm(A_1\cdots F_l)]_{\substack{R\\A}} T_\pm(G_{l+1}\cdots \Lambda_i\cdots Z_n)$$
$$+(\pm)\sum_{i<j=l+1}^{n} [C_i, [D_i, T_\pm(A_1\cdots F_l)]_{\substack{R\\A}}]_{\substack{R\\A}} T_\pm(G_{l+1}\cdots \Lambda_{ij}\cdots Z_n)$$
$$\vdots$$
$$+(\pm)^{n-l-1}[G_{l+1}, [H_{l+2}, \cdots [Z_n, T_\pm(A_1\cdots F_l)]_{\substack{R\\A}}\cdots]_{\substack{R\\A}}. \tag{3.9}$$

An equivalent set of equations to (3.2), (3.8), and (3.9) is related by the symmetries

$$[T_\pm(G_{l+1}\cdots Z_n), T_\pm(A_1\cdots F_l)]_{\substack{R\\A}} = [T_\pm(A_1\cdots F_l), T_\pm(G_{l+1}\cdots Z_n)]_{\substack{A\\R}} \tag{3.10}$$

and

$$[\cdots[T_\pm(G_{l+1}\cdots Z_n), A_1]_{\substack{R\\A}}, B_2]_{\substack{R\\A}}, \cdots, F_l]_{\substack{R\\A}}$$
$$\text{is symmetric under permutation of the operators } A_1, B_2, \cdots F_l. \tag{3.11}$$

These equations are

$$[T_\pm(A_1\cdots F_l), T_\pm(G_{l+1}\cdots Z_n)]_{\substack{A\\R}} = +\sum_{i=1}^{l} T_\pm(A_1\cdots \Lambda_i\cdots F_l)[C_i, T_\pm(G_{l+1}\cdots Z_n)]_{\substack{A\\R}}$$
$$+(\pm)\sum_{i<j=1}^{l} T_\pm(A_1\cdots \Lambda_{ij}\cdots F_l)[C_i, [D_i, T_\pm(G_{l+1}\cdots Z_n)]_{\substack{A\\R}}]_{\substack{A\\R}}$$
$$\vdots$$
$$+(\pm)^{l-1}[A_1, [B_2, \cdots [F_l, T_\pm(G_{l+1}\cdots Z_n)]_{\substack{A\\R}}\cdots]_{\substack{A\\R}} \tag{3.12}$$

and

$$[T_\pm(A_{1\pm}\cdots F_l), T_\pm(G_{l+1}\cdots Z_n)]_{\substack{A\\R}} = +\sum_{i=l+1}^{n} [T_\pm(A_1\cdots F_l), C_i]_{\substack{A\\R}} T_\pm(G_{l+1}\cdots \Lambda_i\cdots Z_n)$$
$$+(\pm)\sum_{i<j=l+1}^{n} [[T_\pm(A_1\cdots F_l), C_i]_{\substack{A\\R}}, D_j]_{\substack{A\\R}} T_\pm(G_{l+1}\cdots \Lambda_{ij}\cdots Z_n)$$
$$\vdots$$
$$+(\pm)^{n-l-1}[\cdots[T_\pm(A_1\cdots F_l), G_{l+1}]_{\substack{A\\R}}, H_{l+2}]_{\substack{A\\R}}, \cdots Z_n]_{\substack{A\\R}}. \tag{3.13}$$

As a check of this result consider the neutral scalar field $A(x)$ and its asymptotic free field $a(x)$, symbolically $A(x) \to a(x)$ for $t \to -\infty$. A well known consequence of microcausality is that the interpolating field can be written as[14]

[14] V. Glaser, H. Lehmann, and N. Zimmermann, Nuovo Cimento 6, 1122 (1957).

$$A(x) = a(x) + \sum_{n=2}^{\infty} \frac{1}{n!} \int K_1 \cdots K_n \langle R(x; x_1 \cdots x_n) \rangle_0 \, :a_1 \cdots a_n: \, dx_1 \cdots dx_n. \tag{3.14}$$

Consequently, the off-mass-shell extension of $A(x)$ is given by

$$\langle \delta^n A(x)/(\delta x_1 \cdots \delta x_n) \rangle_0 = \mathfrak{N} K_1 \cdots K_n \langle R(x; x_1 \cdots x_n) \rangle_0. \tag{3.15}$$

The corresponding operator equation

$$\delta^n A(x)/(\delta x_1 \cdots \delta x_n) = \mathfrak{N} K_1 \cdots K_n R(x, x_1 \cdots x_n), \tag{3.16}$$

which provides the off-mass-shell extension of the results of Ref. 14 also holds. This relation was first obtained by T. W. Chen[12] who did not use the $\mathfrak{N}$ operator in the operator derivative. While this operator is not necessary, it simplifies the argument considerably as can be seen by the following very short proof of (3.16).

First we define

$$i^n R'(x; x_1 \cdots x_n) \equiv [\cdots [A_x, A_1]_R, A_2]_R, \cdots A_n]_R. \tag{3.17}$$

Then it follows by induction that

$$\delta^n A_x/(\delta x_1 \cdots \delta x_n) = \mathfrak{N} K_1 \cdots K_n R'(x; x_1 \cdots x_n). \tag{3.18}$$

For $n = 1$ this is the "dynamical axiom" of Pugh[9] which was derived from the fundamental axioms in a recent paper.[10] Assume it now for n. Then

$$\begin{aligned} \delta^{n+1} A_x/(\delta x_1 \cdots \delta x_{n+1}) &= (\delta/\delta x_{n+1}) \mathfrak{N} K_1 \cdots K_n R'(x; x_1 \cdots x_n) \\ &= \mathfrak{N} K_1 \cdots K_n (\delta/\delta x_{n+1}) R'(x; x_1 \cdots x_n) \\ &= \mathfrak{N} K_1 \cdots K_n K_{n+1} (-i)[R'(x; x_1 \cdots x_n), A_{n+1}]_R. \end{aligned} \tag{3.19}$$

The last step follows by noting that $R'(x; x_1 \cdots x_n)$ is a linear combination of polynomials in the A-field, that Euler's theorem applies, and that $i\delta A_1/\delta x_2 = \mathfrak{N} K_2 [A_1, A_2]_R$. But by definition

$$-i[R'(x; x_1 \cdots x_n), A_{n+1}]_R = R'(x; x_1 \cdots x_n x_{n+1}), \tag{3.20}$$

proving the contention (3.18). In Appendix B it is further demonstrated that $R' = R$, concluding the proof of (3.16).

Thus, using (3.16), (3.20), and (3.2) for a real scalar field, we obtain the interesting form:

$$\begin{aligned} \mathfrak{N} K_x K_1 \cdots K_n [A_x, T_+(A_1 \cdots A_n)]_R = \mathfrak{N} \Big\{ & \sum_{i=1}^{n} K_x K_1 \cdots \Lambda_i \cdots K_n T_+(A_1 \cdots \Lambda_i \cdots A_n) \frac{i \, \delta A_x}{\delta x_i} \\ & + \sum_{i<j=1}^{n} K_x K_1 \cdots \Lambda_{ij} \cdots K_n T_+(A_1 \cdots \Lambda_{ij} \cdots A_n) \frac{i^2 \, \delta^2 A_x}{\delta x_i \, \delta x_j} \\ & \vdots \\ & + K_x (i^n \, \delta^n A_x/\delta x_1 \cdots \delta x_n) \Big\}. \end{aligned} \tag{3.21}$$

Using (2.8) the left of (3.21) becomes,

$$\begin{aligned} \mathfrak{N} K_x K_1 \cdots K_n [A_x, T_+(A_1 \cdots A_n)]_R &= \mathfrak{N} K_x K_1 \cdots K_n \{ T_+(A_x A_1 \cdots A_n) \\ &\quad - T_+(A_1 \cdots A_n) A_x \}. \end{aligned} \tag{3.22}$$

The off-mass-shell extension of S is given by (see Sec. 5)

$$i^n \delta^n S/(\delta x_1 \cdots \delta x_n) = S \mathfrak{N} K_1 \cdots K_n T_+(A_1 \cdots A_n). \tag{3.23}$$

Consequently, with (3.22) and (3.23), (3.21) can be written as

$$\begin{aligned} \frac{i^{n+1} \, \delta^{n+1} S}{\delta x \, \delta x_1 \cdots \delta x_n} &= \mathfrak{N} K_x \frac{i^n \, \delta^n (S A_x)}{\delta x_1 \cdots \delta x_n} \\ &= \mathfrak{N} \frac{i^n \, \delta^n (S J_x)}{\delta x_1 \cdots \delta x_n} \\ &= \mathfrak{N} \frac{i^n \, \delta^n}{\delta x_1 \cdots \delta x_n} \left(\frac{i \, \delta S}{\delta x} \right) \\ &= \frac{i^{n+1} \, \delta^{n+1} S}{\delta x \, \delta x_1 \cdots \delta x_n}, \end{aligned} \tag{3.24}$$

thus demonstrating the consistency of the T factorization with the functional derivative notation.

4. $T_+(A_1 \cdots A_n) = S^*(a_1 \cdots a_n S)_+$

Let a_i be a free Hermitian scalar field at the space-time point x_i. Then the corresponding interpolating field $A_i \to a_i$ for $t_i \to -\infty$ is given by[15]

$$A_i = S^*(a_i S)_+. \tag{4.1}$$

A_i is also Hermitian:

$$\begin{aligned} A_i^* &= (a_i S^*)_- S = \{S^* a_i - [a_i, S^*]_a\} S \\ &= S^*\{(a_i S)_+ - [a_i, S]_a\} - [a_i, S^*]_a S \\ &= S^*(a_i S)_+ - [a_i, S^* S]_a. \end{aligned} \tag{4.2}$$

The last term vanishes because of unitarity, so that

$$A_i^* = A_i.$$

To prove the contention of this section we proceed by induction. For $n = 1$ we regain (4.1). Let us therefore assume it for n and prove it for $n + 1$. Starting with (2.10),

$$\begin{aligned} T_+(A_1 \cdots A_{n+1}) &= A_{n+1} T_+(A_1 \cdots A_n) \\ &\quad + [A_{n+1}, T_+(A_1 \cdots A_n)]_a. \end{aligned} \tag{4.3}$$

By Eq. (4.2) and assumption,

$$\begin{aligned} &A_{n+1} T_+(A_1 \cdots A_n) \\ &= (a_{n+1} S^*)_-(a_1 \cdots a_n S)_+ \\ &= \{S^* a_{n+1} - [a_{n+1}, S^*]_a\}(a_1 \cdots a_n S)_+ \\ &= S^*\{(a_1 \cdots a_{n+1} S)_+ - [a_{n+1}, (a_1 \cdots a_n S)_+]_a\} \\ &\quad - [a_{n+1}, S^*]_a (a_1 \cdots a_n S)_+ \\ &= S^*(a_1 \cdots a_{n+1} S)_+ - [a_{n+1}, S^*(a_1 \cdots a_n S)_+]_a. \end{aligned} \tag{4.4}$$

However, by assumption

$$\begin{aligned} &[a_{n+1}, S^*(a_1 \cdots a_n S)_+]_a \\ &\quad = [a_{n+1}, T_+(A_1 \cdots A_n)]_a \\ &\quad = -\int d\zeta\, \Delta_A(x_{n+1} - \zeta) \\ &\qquad \times \sum_{j=1}^{n} T_+(A_1 \cdots i\, \delta A_j/\delta\zeta \cdots A_n) \\ &\quad = -\int d\zeta\, \Delta_A(x_{n+1} - \zeta) K_\zeta \\ &\qquad \times \sum_{j=1}^{n} T_+(A_1 \cdots [A_j, A_\zeta]_R \cdots A_n), \end{aligned} \tag{4.5}$$

where Bogoliubov causality[16] has been used in obtaining (4.5). Integration by parts reduces (4.5) to

$$\begin{aligned} &[a_{n+1}, S^*(a_1 \cdots a_n S)_+]_a \\ &\quad = [T_+(A_1 \cdots A_n), A_{n+1}]_R \\ &\quad = [A_{n+1}, T_+(A_1 \cdots A_n)]_A, \end{aligned} \tag{4.6}$$

such that with (4.3) and (4.4) we have the desired result

$$T_+(A_1 \cdots A_{n+1}) = S^*(a_1 \cdots a_{n+1} S)_+. \tag{4.7}$$

Some time after conclusion of the work reported here we discovered a paper by Medvedev[17] who, by a different argument, has also arrived at the conclusion, (4.7); he has also correlated his result with the off-mass-shell extension of the S_{op} as we do in Sec. 5.

5. MULTIPLE OPERATOR DERIVATIVES AS T-PRODUCTS

We return to the result (2.8) and derive from it the strong operator equation

$$\begin{aligned} i^n\, \delta^n F/(\delta a_1 \cdots \delta a_n) \\ = \mathfrak{N} K_1 \cdots K_n (a_1 \cdots a_n F)_+. \end{aligned} \tag{5.1}$$

The proof proceeds by induction. For $n = 1$ we have

$$\begin{aligned} \mathfrak{N} K_1 (a_1 F)_+ &= \mathfrak{N} K_1 (F a_1 + [a_1, F]_r) \\ &= \mathfrak{N} K_1 [a_1, F]_r = i\, \frac{\delta F}{\delta a_1}. \end{aligned} \tag{5.2}$$

Assume now that (5.1) is valid for n, then

$$\begin{aligned} &\mathfrak{N} K_1 \cdots K_{n+1} (a_1 \cdots a_{n+1} F)_+ \\ &\quad = \mathfrak{N} K_1 \cdots K_n \cdot K_{n+1} (a_1 \cdots a_{n+1} F)_+. \end{aligned} \tag{5.3}$$

The last factor becomes

$$\begin{aligned} &K_{n+1}(a_1 \cdots a_{n+1} F)_+ \\ &\quad = K_{n+1}\{(a_1 \cdots a_n F)_+ a_{n+1} + [a_{n+1}, (a_1 \cdots a_n F)_+]_r\} \\ &\quad = (i\, \delta/\delta a_{n+1})(a_1 \cdots a_n F)_+. \end{aligned}$$

When substituted into (5.3) this gives

---

[15] The subscripts $(\pm)$ have been defined in Sec. 1. The subscripts a, r are similarly defined as advanced and retarded commutators with respect to the implicit free fields. E.g.,

$$\begin{aligned} (a_1 \cdots a_m S)_\pm &\equiv \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int d\zeta_1 \cdots d\zeta_n \\ &\quad \times w_n^{(\pm)}(\zeta_1 \cdots \zeta_n) T_\pm(a_1 \cdots a_m a_{\zeta_1} \cdots a_{\zeta_n}), \\ [a_x, S]_{\substack{a \\ r}} &\equiv \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int d\zeta_1 \cdots d\zeta_n \\ &\quad \times w_n^{(+)}(\zeta_1 \cdots \zeta_n)[a_x, T_+(a_{\zeta_1} \cdots a_{\zeta_n})]_{\substack{A \\ R}}. \end{aligned}$$

[16] See Ref. 10; it is shown there that Pugh's "dynamical axiom" (Ref. 9)

$$\frac{i\, \delta A_1}{\delta x_2} = K_2[A_1, A_2]_R$$

is a consequence of Bogoliubov causality.

[17] B. V. Medvedev, Zh. Eksperim. i Teor. Fiz. **48**, 1479 (1965) [English transl.: Soviet Phys.—JETP **21**, 989 (1965)].

$$\mathfrak{N}K_1 \cdots K_{n+1}(a_1 \cdots a_{n+1}F)_+$$
$$= \mathfrak{N}K_1 \cdots K_n(i\delta/\delta a_{n+1})(a_1 \cdots a_nF)_+$$
$$= i(\delta/\delta a_{n+1})\cdot i^n[\delta^nF/(\delta a_1 \cdots \delta a_n)],$$

and establishes (5.1) for $n + 1$. This completes the induction proof of (5.1).

As an application of the result (5.1), we derive the off-mass-shell form of the S-operator. Let $F = S$, then

$$i^nS^*[\delta^nS/(\delta a_1 \cdots \delta a_n)]$$
$$= NK_1 \cdots K_nS^*(a_1 \cdots a_nS)_+$$
$$= NK_1 \cdots K_nT_+(A_1 \cdots A_n). \quad (5.4)$$

In the last step we made use of (4.7). Now, the φ -product differs from the T-product only in terms containing one or more factors of $\Delta_c(x_i - x_j)$. Because of the factors K_iK_j in (5.4), these become $-K_j\delta(x_i - x_j)$ and are to be omitted due to the appearance of the operator $\mathfrak{N}$. Thus,

$$\mathfrak{N}K_1 \cdots K_nT_+(A_1 \cdots A_n)$$
$$= \mathfrak{N}K_1 \cdots K_n\varphi(A_1 \cdots A_n) \quad (5.5)$$

and (5.4) can also be written as

$$i^nS^*[\delta^nS/(\delta a_1 \cdots \delta a_n)]$$
$$= \mathfrak{N}K_1 \cdots K_n\varphi(A_1 \cdots A_n). \quad (5.6)$$

That the operator $\mathfrak{N}$ can be omitted in (5.6) [but not in (5.4)] will not be shown here.[18]

Another application of (5.1) refers to the expansion of an operator in terms of normal ordered products,

$$F = \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int f_n(x_1 \cdots x_n) \times :a_1 \cdots a_n: \, dx_1 \cdots dx_n. \quad (5.7)$$

The definition of the derivative specifies

$$f_n(x_1 \cdots x_n) = \langle \mathfrak{N}i^n[\delta^nF/(\delta a_1 \cdots \delta a_n)]\rangle_0. \quad (5.8)$$

Therefore,

$$f_n(x_1 \cdots x_n) = \langle \mathfrak{N}K_1 \cdots K_n(a_1 \cdots a_nF)_+\rangle_0. \quad (5.9)$$

The off-mass-shell behavior of the $f_n(x_1 \cdots x_n)$ is therefore determined by this expression for any F of the form (5.7). The vacuum expectation value of (5.6) is a special case of this result.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge helpful discussions with our colleagues, especially Professor M. Wellner, Dr. V. Gorgé, and Dr. Tuan-Wu Chen.

[18] An indirect proof is contained in Ref. 9 where (5.6) is derived without the benefit of the $\mathfrak{N}$ operator.

Note Added in Proof: As mentioned following (3.16), the use of the operator $\mathfrak{N}$ can be avoided. To this end one can admit the free-field equation $K_xa(x) = 0$ only as a weak equation, i.e., it cannot remain valid under operator differentiation. The operator derivative can then be defined to commute with differentiation and integration. The interpolating field, defined in (1.21), will still satisfy the usual integral equation, but $K_xA(x) = J(x)$ will now be valid only weakly [this is not the choice of $A(x)$ adopted in Ref. 12].

With these definitions all equations of Secs. 1 through 4 and the Appendices remain valid as strong equations without the operator $\mathfrak{N}$, except (3.23) and (3.24). These equations as well as Sec. 5 from which they arise are modified. However, the important result (5.6) remains valid without $\mathfrak{N}$ as a weak equation.

APPENDIX A

The compatibility of Eqs. (3.2) and (2.5) is easily shown. The two statements are

$$[F_x, T_+(A_1 \cdots Z_n)]_R$$
$$= \sum_{i=1}^{n} T_+(A_1 \cdots \Lambda_i \cdots Z_n)[F_x, C_i]_R$$
$$+ \sum_{i<j=1}^{n} T_+(A_1 \cdots \Lambda_{ij} \cdots Z_n)[[F_x, C_i]_R, D_j]_R$$
$$\vdots$$
$$+ [\cdots [F_x, A_1]_R, B_2]_R, \cdots Z_n]_R \quad (A1)$$

and

$$[F_x, T_+(A_1 \cdots Z_n)]_R$$
$$\equiv \sum_{i=1}^{n} T_+(A_1 \cdots [E_x, C_i]_R \cdots Z_n). \quad (A2)$$

From Appendix B we have that

$$[\cdots [F_x, A_1]_R, \cdots, D_m]_R = i^nR(F_x; A_1 \cdots D_m)$$
$$= \sum_{\substack{\text{perm} \\ (1\cdots m)}} \theta_{z_1\cdots m}[\cdots [F_x, A_1], B_2], \cdots D_m]. \quad (A3)$$

Further, for convenience, we assume the time ordering

$$T_+(A_1 \cdots Z_n) = A_1 \cdots Z_n, \quad \text{i.e.,}$$
$$t_1 > t_2 > \cdots > t_n. \quad (A4)$$

Then using (A3), (A1) becomes

$$[F_x, A_1 \cdots Z_n]_R = \sum_{i=1}^{n} A_1 \cdots \Lambda_i \cdots Z_n[F_x, C_i]_R$$
$$+ \sum_{i<j=1}^{n} A_1 \cdots \Lambda_{ij} \cdots Z_n[[F_x, C_i]_R, D_j]$$
$$+ \cdots + [\cdots [F_x, A_1]_R, B_2], \cdots Z_n], \quad (A5)$$

and (A.2) becomes

$$[F_x, A_1 \cdots Z_n]_R = \sum_{i=1}^{n} A_1 \cdots [F_x, C_i]_R \cdots Z_n \quad \text{(A6)}$$

$$= [F_x, A_1 \cdots Y_{n-1}]_R Z_n + A_1 \cdots Y_{n-1}[F_x, Z_n]_R. \quad \text{(A7)}$$

We shall now prove that this expression is the same as (A5) by using induction. Eqs. (A7) and (A5) are clearly identical for $n = 1$. Assume now that they are both valid for $n - 1$. Then

$$[F_x, A_1 \cdots Z_n]_R = A_1 \cdots Y_{n-1}[F_x, Z_n]_R + \Big\{ \sum_{i=1}^{n-1} A_1 \cdots \Lambda_i \cdots Y_{n-1}[F_x, C_i]_R + \sum_{i<j=1}^{n-1} A_1 \cdots \Lambda_{ij} \cdots Y_{n-1}[[F_x, C_i]_R, D_j] \vdots + [\cdots [F_x, A_1]_R, B_2], \cdots, Y_{n-1}] \Big\} Z_n. \quad \text{(A8)}$$

The *first two* terms on the right of (A8) give

$$\sum_{i=1}^{n} A_1 \cdots \Lambda_i \cdots Z_n[F_x, C_i]_R + \sum_{i=1}^{n-1} A_1 \cdots \Lambda_i \cdots Y_{n-1}[[F_x, C_i]_R, Z_n].$$

The first term agrees exactly with the first term in (A5) while the second term together with the *third* term of (A8) give

$$\sum_{i<j=1}^{n} A_1 \cdots \Lambda_{ij} \cdots Z_n[[F_x, C_i]_R, D_j] + \sum_{i<j=1}^{n-1} A_1 \cdots \Lambda_{ij} \cdots Y_{n-1}[[[F_x, C_i]_R, D_j], Z_n].$$

Again, the first term agrees with the second term of (A5) while the second term combined with the next (fourth) term of (A8) yields again two terms, etc. In this way the expression (A8) is completely rewritten in the form (A5).

This completes the induction proof that (A1) and (A2) are equivalent.

APPENDIX B

The R product, defined by

$$i^n R(A_x; B_1 C_2 \cdots Z_n) \equiv \sum_{\substack{\text{perm} \\ 1 \cdots n}} \theta_{x1\ldots n}[\cdots [A_x, B_1], C_2], \cdots, Z_n], \quad \text{(B1)}$$

can also be expressed in terms of the R commutator, R', defined by

$$i^n R'(A_x; B_1 C_2 \cdots Z_n) \equiv [\cdots [A_x, B_1]_R, C_2]_R, \cdots Z_n]_R. \quad \text{(B2)}$$

In fact, we have

$$R = R' \quad \text{(B3)}$$

or

$$i^{n+1} R(A_x; B_1 \cdots Z_{n+1}) = i^n [R(A_x; B_1 \cdots Y_n), Z_{n+1}]_R, \quad \text{(B4)}$$

which implies (B3) by iteration.

To prove (B4) assume the particular time ordering (we set $t_x \equiv t_0$ for convenience)

$$t_x \equiv t_0 > t_1 > t_2 > \cdots > t_n \quad \text{(B5)}$$

such that

$$i^n R(A_0; B_1 \cdots Y_n) = [\cdots [A_0, B_1], C_2], \cdots Y_n]. \quad \text{(B6)}$$

Then

$$i^n [R(A_0; B_1 \cdots Y_n), Z_{n+1}]_R = [\cdots [A_0, B_1], C_2], \cdots Y_n], Z_{n+1}]_R = \sum_{i=0}^{n} [\cdots [A_0, B_1], \cdots [D_i, Z_{n+1}]_R], \cdots, Y_n], \quad \text{(B7)}$$

$$i^n [R(A_0; B_1 \cdots Y_n), Z_{n+1}]_R = \theta_{0n+1}[\cdots [A_0, Z_{n+1}], B_1], \cdots Y_n] + \sum_{i=1}^{n} \theta_{in+1}\{[\cdots [A_0, B_1], \cdots, D_i], Z_{n+1}], \cdots Y_n] - [\cdots [A_0, B_1], \cdots Z_{n+1}], D_i], \cdots Y_n]$$

$$= \sum_{i=1}^{n} (\theta_{i-1,n+1} - \theta_{in+1}) \times [\cdots [A_0, B_1], \cdots D_{i-1}], Z_{n+1}], E_i], \cdots Y_n] + \theta_{nn+1}[\cdots [A_0, B_1], C_2], \cdots Y_n], Z_{n+1}]$$

$$= \sum_{i=1}^{n} \theta_{i-1,n+1,i}[\cdots [A_0, B_1], \times \cdots D_{i-1}], Z_{n+1}], E_i], \cdots Y_n] + \theta_{nn+1}[\cdots [A_0, B_1], \cdots, Y_n], Z_{n+1}], \quad \text{(B8)}$$

using

$$\theta_{i-1,n+1} - \theta_{in+1} \to \theta_{i-1,n+1,i} \quad \text{for} \quad t_{i-1} > t_i. \quad \text{(B9)}$$

But, Eq. (B8) is just $i^{n+1} R(A_0; B_1 \cdots Y_n Z_{n+1})$ subject to the condition (B5). Thus, since a proof for one particular time ordering is a proof for arbitrary time ordering, Eq. (B.3) is proved.

Algebraic Aspects of Trilinear Commutation Relations

H. SCHARFSTEIN
5500 Fieldston Road, Bronx, New York 10471
(Received 11 October 1965)

An attempt is made to derive from the formalism of Schwinger's action principle, in a more convincing manner than previously described, a set of trilinear equal-time commutation relations which contains the commutation relations first discussed by H. S. Green as special cases. Matrix representations of field operators satisfying the trilinear commutation relations are considered. Two representations are explicitly discussed: a four-dimensional and an eight-dimensional representation. The representations considered and the bilinear equal-time commutation relations between the associated "component fields" obeying ordinary statistics are specified by irreducible representations of an algebra which is suggested by the trilinear commutation relations. The component fields associated with the same representation and of the same spin differ from each other in their bilinear equal-time commutation relations with other fields. This difference is reflected in the interactions into which the various fields can enter.

INTRODUCTION

IN the process of deriving commutation relations from the formalism of Schwinger's action principle,[1] it is customary to subject the field variations to a variety of restrictions. The commutation relations between kinematically related and unrelated field variables are then obtained from the generator equations and the restrictions imposed on the variations. Different assumptions about the operator properties of the field variations naturally lead to different commutation relations.[1-6] It is known, for example, that the trilinear commutation relations first considered by Green[7] are special cases of a set of trilinear commutation relations which can be obtained from the action principle if particularly simple restrictions are imposed on the variations.[6] The procedure of selecting for consideration classes of variations having certain operator properties of course also has implications bearing on the interactions which can be considered in the Lagrangian context. It is therefore of interest to inquire to what extent it is necessary to impose restrictions on the variations other than those implied by the definition of the generators of the infinitesimal transformations, and by the equations these generators satisfy. These equations [Eqs. (3) and (4)] amount to trilinear commutation relations between kinematically related and unrelated fields. The commutation relations (9) and (16) are derived with the definite assumption that a field variation and the corresponding field have identical equal-time commutation behavior with respect to all fields, except that the contraction of a variation with any field operator vanishes identically.

The algebra defined by Eqs. (20) and (21) may be of interest even outside the context in which it is discussed, i.e., outside the context of trilinear equal-time commutation relations.

THE GENERATOR EQUATIONS AND COMMUTATION RELATIONS

In the framework of Schwinger's action principle, it is natural that two different types of fields have to be considered. These are fields whose infinitesimal transformations are generated by antisymmetrized generators, and fields whose infinitesimal transformations are generated by symmetrized generators. The former will be denoted by ψ and will be referred to as fields of type one, while the latter will be denoted by ϕ and will be referred to as fields of type two.

The generators of the infinitesimal transformations, which, in the context of Schwinger's action principle, generate the infinitesimal transformations of the two types of fields and the infinitesimal transformations of the respective adjoints or canonical conjugates $\bar{\psi}$ and π, are respectively defined by

$$G(\psi) = i \int [\bar{\psi}(\mathbf{x}), \gamma_4 \, \delta\psi(\mathbf{x})]_- \, d\mathbf{x}, \tag{1a}$$

$$G(\bar{\psi}) = i \int [\delta\bar{\psi}(\mathbf{x}), \gamma_4\psi(\mathbf{x})]_- \, d\mathbf{x}, \tag{1b}$$

[1] J. Schwinger, Phys. Rev. **82**, 914 (1951); *ibid.* **91**, 713 (1953); Proc. Natl. Acad. Sci. U.S. **44**, 223, 617 (1958); *1959 Brandeis University Summer Institute in Theoretical Physics Lecture Notes* (notes transcribed by A. M. Kaufman), p. 81.
[2] T. W. B. Kibble and J. C. Polkinghorne, Proc. Roy. Soc. (London) **A243**, 252 (1957).
[3] D. V. Volkov, Zh. Eksperim. i Teor. Fiz. **36**, 1560 (1959) [English transl.: Soviet Phys.—JETP **9**, 1107 (1959)].
[4] R. Arnowitt and S. Deser, J. Math. Phys. **3**, 637 (1962).
[5] H. Scharfstein, Thesis, New York University (1962) (unpublished), and Clarification to the Thesis (1962) (unpublished).
[6] H. Scharfstein, Nuovo Cimento **30**, 740 (1963).
[7] H. S. Green, Phys. Rev. **90**, 270 (1953).

and

$$\tilde{G}(\phi) = \int [\pi(\mathbf{x}),\ \delta\phi(\mathbf{x})]_+\, d\mathbf{x}, \tag{2a}$$

$$\tilde{G}(\pi) = \int [\delta\pi(\mathbf{x}),\ \phi(\mathbf{x})]_+\, d\mathbf{x}. \tag{2b}$$

The factor γ_4 in Eqs. (1a) and (1b) appears to bias antisymmetrized generators in favor of spinors. Actually, this factor has been introduced into the definition of the antisymmetrized generators only for convenience.

The generators defined in Eqs. (1) satisfy the following relations:

$$[\psi(\mathbf{x}),\ G(\psi)]_- = i\ \delta\psi(\mathbf{x}), \tag{3a}$$

$$[\bar{\psi}(\mathbf{x}),\ G(\bar{\psi})]_- = -i\ \delta\bar{\psi}(\mathbf{x}), \tag{3b}$$

$$[\psi(\mathbf{x}),\ G(\bar{\psi})]_- = [\bar{\psi}(\mathbf{x}),\ G(\psi)]_- = 0, \tag{3c}$$

and

$$[\Theta(\mathbf{x}),\ G(\psi)]_- = [\Theta(\mathbf{x}),\ G(\bar{\psi})]_- = 0, \tag{3d}$$

where in (3d) the operator $\Theta(\mathbf{x})$ refers to a field variable of either type one or two, which is kinematically unrelated to the type one field ψ.

The relations satisfied by the symmetrized generators are

$$[\phi(\mathbf{x}),\ \tilde{G}(\phi)]_- = i\ \delta\phi(\mathbf{x}), \tag{4a}$$

$$[\pi(\mathbf{x}),\ \tilde{G}(\pi)]_- = -i\ \delta\pi(\mathbf{x}), \tag{4b}$$

$$[\phi(\mathbf{x}),\ \tilde{G}(\pi)]_- = [\pi(\mathbf{x}),\ \tilde{G}(\phi)]_- = 0, \tag{4c}$$

and

$$[\Theta(\mathbf{x}),\ \tilde{G}(\phi)]_- = [\Theta(\mathbf{x}),\ \tilde{G}(\pi)]_- = 0, \tag{4d}$$

where in (4d) the operator $\Theta(\mathbf{x})$ refers to a field variable of either type one or two, which is kinematically unrelated to the type-two field ϕ.

With repeated application of suitable trilinear algebraic identities, the following expressions are obtained from Eqs. (1)–(4):

$$i\int [\delta\psi_\mu(\mathbf{x})(\gamma_4)_{\mu\alpha},\ (\{\bar{\psi}_\alpha(\mathbf{x}),\ [\chi_\beta(\mathbf{x}'),\ \chi'_\gamma(\mathbf{x}'')]_-\}_- - \delta(\mathbf{x}-\mathbf{x}')(\gamma'_4)_{\alpha\beta}\chi'_\gamma(\mathbf{x}'') + \delta(\mathbf{x}-\mathbf{x}'')(\gamma'_4)_{\alpha\gamma}\chi_\beta(\mathbf{x}'))]_-\, d\mathbf{x}$$
$$- i\int [\bar{\psi}_\alpha(\mathbf{x})(\gamma_4)_{\mu\alpha},\ \{\delta\psi_\mu(\mathbf{x}),\ [\chi_\beta(\mathbf{x}'),\ \chi'_\gamma(\mathbf{x}'')]_-\}_-]_-\, d\mathbf{x} = 0, \tag{5a}$$

$$i\int [\delta\bar{\psi}_\mu(\mathbf{x})(\gamma_4)_{\mu\alpha},\ (\{\psi_\alpha(\mathbf{x}),\ [\chi_\beta(\mathbf{x}'),\ \chi'_\gamma(\mathbf{x}'')]_-\}_- - \delta(\mathbf{x}-\mathbf{x}')(\gamma'_4)_{\alpha\beta}\chi'_\gamma(\mathbf{x}'') + \delta(\mathbf{x}-\mathbf{x}'')(\gamma'_4)_{\alpha\gamma}\chi_\beta(\mathbf{x}'))]_-\, d\mathbf{x}$$
$$- i\int [\psi_\alpha(\mathbf{x})(\gamma_4)_{\mu\alpha},\ \{\delta\bar{\psi}_\mu(\mathbf{x}),\ [\chi_\beta(\mathbf{x}'),\ \chi'_\gamma(\mathbf{x}'')]_-\}_-]_-\, d\mathbf{x} = 0, \tag{5b}$$

$$\int [\delta\phi_\mu(\mathbf{x}),\ (\{\pi_\mu(\mathbf{x}),\ [\chi_\nu(\mathbf{x}'),\ \chi'_\rho(\mathbf{x}'')]_+\}_- - i\ \bar{\delta}_{\mu\nu}\ \delta(\mathbf{x}-\mathbf{x}')\chi'_\rho(\mathbf{x}'') - i\ \bar{\delta}_{\mu\rho}\ \delta(\mathbf{x}-\mathbf{x}'')\chi_\nu(\mathbf{x}'))]_+\, d\mathbf{x}$$
$$+ \int [\pi_\mu(\mathbf{x}),\ \{\delta\phi_\mu(\mathbf{x}),\ [\chi_\nu(\mathbf{x}'),\ \chi'_\rho(\mathbf{x}'')]_+\}_-]_+\, d\mathbf{x} = 0, \tag{6a}$$

and

$$\int [\delta\pi_\mu(\mathbf{x}),\ (\{\phi_\mu(\mathbf{x}),\ [\chi_\nu(\mathbf{x}'),\ \chi'_\rho(\mathbf{x}'')]_+\}_- - i\ \bar{\delta}_{\mu\nu}\ \delta(\mathbf{x}-\mathbf{x}')\chi'_\rho(\mathbf{x}'') - i\ \bar{\delta}_{\mu\rho}\ \delta(\mathbf{x}-\mathbf{x}'')\chi_\nu(\mathbf{x}'))]_+\, d\mathbf{x}$$
$$+ \int [\phi_\mu(\mathbf{x}),\ \{\delta\pi_\mu(\mathbf{x}),\ [\chi_\nu(\mathbf{x}'),\ \chi'_\rho(\mathbf{x}'')]_+\}_-]_+\, d\mathbf{x} = 0. \tag{6b}$$

The modified gamma-four matrix, γ'_4, entering Eqs. (5) and the ordered Kronecker delta, $\bar{\delta}_{\mu\nu}$, occurring in Eqs. (6), are respectively defined as follows:

$$(\gamma'_4)_{\alpha\beta} = (\gamma'_4)_{\beta\alpha} \equiv \begin{cases} (\gamma_4)_{\alpha\beta} & \text{if } \alpha \text{ and } \beta \text{ refer to canonically conjugate (type-one) field components,} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and

$$\bar{\delta}_{\mu\nu} = -\bar{\delta}_{\nu\mu} \equiv \begin{cases} \delta_{\mu\nu} & \text{if } \mu \text{ refers to a field component and } \nu \text{ to the canonically conjugate component,} \\ -\delta_{\mu\nu} & \text{if } \nu \text{ refers to a field component and } \mu \text{ to the canonically conjugate component,} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In the derivation of Eqs. (5) and (6), no restrictions have been imposed on the variations other than those implied by Eqs. (1)–(4). Were it not for the second integral on the left-hand side of each one of Eqs. (5) and (6), commutation relations could be obtained from these equations, because in the first integral in each case the variations stand to the right or left, symmetrically disposed, with the right and left coefficients differing at most by a phase factor. It seems reasonable to require that the commutation behavior of a field variation be parallel to that of the corresponding field, except that the contraction of a variation with any field operator vanishes identically. This requirement is consistent with and indicated by Eqs. (1)–(4). Moreover, this requirement also ensures that two field operators, which are obtained from each other by infinitesimal transformations of the type considered in Schwinger's action principle, have the same commutation properties. The commutation relations, which can be obtained from Eq. (5a) when the second integral on the left-hand side of the equation is ignored and when identical right and left coefficients of the variations in the first integral are equated to zero simultaneously, imply that the second integral on the left-hand side of Eq. (5b) vanishes. The commutation relations which can then be obtained from the first integral on the left-hand side of Eq. (5b) in turn guarantee the vanishing of the second integral on the left-hand side of Eq. (5a). Commutation relations can be obtained from Eqs. (6) in a similar manner. Since the variations involved in the process, $\delta\psi$, $\delta\bar{\psi}$, $\delta\phi$, and $\delta\pi$, can presumably be performed independently, the above procedure of deriving commutation relations from Eqs. (3) and (4) appears to be self-consistent. The trilinear equal-time commutation relations thus obtained are

$$[\psi_\alpha(\mathbf{x}), [\chi_\beta(\mathbf{x}'), \chi'_\gamma(\mathbf{x}'')]_-]_- = \delta(\mathbf{x}-\mathbf{x}')(\gamma'_4)_{\alpha\beta}\chi'_\gamma(\mathbf{x}'') - \delta(\mathbf{x}-\mathbf{x}'')(\gamma'_4)_{\alpha\gamma}\chi_\beta(\mathbf{x}'), \quad (9a)$$

$$[\phi_\mu(\mathbf{x}), [\chi_\nu(\mathbf{x}'), \chi'_\rho(\mathbf{x}'')]_+]_- = i\,\delta_{\mu\nu}\,\delta(\mathbf{x}-\mathbf{x}')\chi'_\rho(\mathbf{x}'') + i\,\delta_{\mu\rho}\,\delta(\mathbf{x}-\mathbf{x}'')\chi_\nu(\mathbf{x}'). \quad (9b)$$

While the symmetrization of the generators of the infinitesimal tranformations of the fields ψ and ϕ is used in the derivation of the above commutation relations and is reflected in the trilinear algebraic identities employed in the derivation, the symmetrization of the generators associated with the fields χ and χ' does not play an explicit role. The fields χ and χ' can therefore, independently of each other and independently of the fields ψ and ϕ, be either of type one or of type two. Moreover, any two operators occurring in the same trilinear commutation relation may be kinematically related or unrelated. Furthermore, any field quantity appearing in Eqs. (9) can denote, independently of the other operators, either a field operator or its canonical conjugate.

The commutation relations derived from the generator equations should in turn imply these equations. Since Eqs. (3) and (4) are of third order in field quantities, commutation relations of order higher than third, involving different fields, need hardly be considered.

MATRIX REPRESENTATIONS

If Eq. (9a) is assumed to be valid only when all the operators occurring in it are kinematically related, then the type-one fields satisfying the resulting commutation relation are known to be representable as linear superpositions of arbitrary numbers of fermion fields, the various fermions commuting with each other at equal times.[5–13] Similarly, if Eq. (9b) is assumed to be valid only when all the operators occurring in it are kinematically related, then the type-two fields satisfying the resulting commutation relation are known to be representable as linear superpositions of arbitrary numbers of anticommuting boson fields.[5–13] In view of these representations, it appears to be natural to try to represent the generalized fields, i.e., those field operators which satisfy the commutation relations (9) or, more correctly, the commutation relations (16) (cf. below), with no restriction of their validity to the kinematically related case, in terms of fermion and boson fields. The following representations of type-one and type-two fields will therefore be considered:

$$\psi(\mathbf{x}) = 2^{-\frac{1}{2}} \sum_{i=1}^{m} A_i \psi_i(\mathbf{x}), \quad (10a)$$

[8] K. Johnson, Math. Rev. 21, 7745 (1960).
[9] G. F. Dell'Antonio, O. W. Greenberg, and E. C. G. Sudarshan, University of Rochester preprint NYO-10241 (1962), and *Group Theoretical Concepts and Methods in Elementary Particle Physics*, F. Gürsey, Ed. (Gordon and Breach Science Publishers, Inc., New York, 1964), p. 403.
[10] L. O'Raifeartaigh and C. Ryan, Proc. Roy. Irish Acad. 62, 93 (1963).
[11] C. Ryan and E. C. G. Sudarshan, Nucl. Phys. 47, 207 (1963).
[12] T. F. Jordan, N. Mukunda, and S. V. Pepper, J. Math. Phys. 4, 1089 (1963).
[13] D. G. Boulware and S. Deser, Nuovo Cimento 30, 230 (1963).

$$\bar{\psi}(\mathbf{x}) = 2^{-\frac{1}{2}} \sum_{i=1}^{m} A_i^{\dagger} \bar{\psi}_i(\mathbf{x}), \tag{10b}$$

$$\phi(\mathbf{x}) = 2^{-\frac{1}{2}} \sum_{i=1}^{n} B_i \phi_i(\mathbf{x}), \tag{10c}$$

$$\pi(\mathbf{x}) = 2^{-\frac{1}{2}} \sum_{i=1}^{n} B_i^{\dagger} \pi_i(\mathbf{x}). \tag{10d}$$

The field operators on the right-hand side of Eqs. (10a) and (10b), respectively, are fermion field operators and their Pauli adjoints, the field operators on the right-hand side of Eqs. (10c) and (10d), respectively, are boson field operators and their adjoints, and the A's and B's are numerical matrices. The relationship between a matrix A_i and its 'conjugate' $A_i^{\dagger}$, as well as the relationship between B_i and $B_i^{\dagger}$, are not specified at present, except that these matrices must be related in such a manner that Eqs. (12) are satisfied. The integers m and n in Eqs. (10) remain to be determined.

While the associative law of multiplication is valid not only for numerical matrices but also for matrices whose nonvanishing elements are field operators obeying commutation relations, some properties of matrices valid in the former case are not in general valid in the latter. In order to bring the methods of matrix algebra to bear on the problem at hand, it is therefore expedient to decouple the numerical matrix coefficients occurring on the right-hand side of Eqs. (10) from the field operators. This can readily be accomplished if it is assumed that for equal times the commutation relations between the various component fields are bilinear, i.e., distinct component fields obeying ordinary statistics either commute or anticommute for equal times. The question whether, in the present context, a component field and its adjoint have identical equal-time commutation behavior with respect to other component fields is left open; i.e., in the derivation of Eqs. (12) for the sake of generality a component field and its adjoint are treated as independent entities as far as their bilinear equal-time commutation relations with other component fields are concerned.

In any theory in which fields of both types are to be considered, the trilinear equal-time commutation relations implied by Eqs. (9), which have to be satisfied by the generalized fields, are[6]

$$[\psi_\alpha(\mathbf{x}), [\psi'_\beta(\mathbf{x}'), \psi''^{\dagger}_\gamma(\mathbf{x}'')]_-]_- = \delta(\mathbf{x}-\mathbf{x}')(\gamma'_4)_{\alpha\beta}\psi''^{\dagger}_\gamma(\mathbf{x}'') - \delta(\mathbf{x}-\mathbf{x}'')(\gamma'_4)_{\alpha\gamma}\psi'_\beta(\mathbf{x}'), \tag{11a}$$

$$[\psi_\alpha(\mathbf{x}), [\psi'_\beta(\mathbf{x}'), \phi(\mathbf{x}'')]_-]_- = \delta(\mathbf{x}-\mathbf{x}')(\gamma'_4)_{\alpha\beta}\phi(\mathbf{x}''), \tag{11b}$$

$$[\psi(\mathbf{x}), \phi(\mathbf{x}')\phi'(\mathbf{x}'')]_- = 0, \tag{11c}$$

$$[\phi_\mu(\mathbf{x}), [\phi'_\nu(\mathbf{x}'), \phi''_\rho(\mathbf{x}'')]_+]_- = i\,\tilde{\delta}_{\mu\nu}\,\delta(\mathbf{x}-\mathbf{x}')\phi''_\rho(\mathbf{x}'') + i\,\tilde{\delta}_{\mu\rho}\,\delta(\mathbf{x}-\mathbf{x}'')\phi'_\nu(\mathbf{x}'), \tag{11d}$$

$$[\phi_\mu(\mathbf{x}), [\phi'_\nu(\mathbf{x}'), \psi(\mathbf{x}'')]_+]_- = i\,\tilde{\delta}_{\mu\nu}\,\delta(\mathbf{x}-\mathbf{x}')\psi(\mathbf{x}''), \tag{11e}$$

$$[\phi(\mathbf{x}), \psi(\mathbf{x}')\psi'(\mathbf{x}'')]_- = 0. \tag{11f}$$

Trying to satisfy the commutation relations (11) with two generalized fields, one of each type, having representations of the kind indicated in Eqs. (10), and assuming that for equal-times the component fermion and boson fields satisfy bilinear commutation relations, as explained above, without specifying what the bilinear commutation relations between any two component fields are, the following numerical matrix equations are obtained:

$$[A_i A_i,\ A_i^{\dagger}]_- = [A_i A_i,\ A'_i]_- = 0, \tag{12a}$$

$$[B_i B_i,\ B_i^{\dagger}]_- = [B_i B_i,\ B'_i]_- = 0, \tag{12b}$$

$$[A_i,\ [A_i^{\dagger},\ A'_j]_{\pm}]_{\mp} = 0, \qquad i \neq j, \tag{12c}$$

$$[B_i,\ [B_i^{\dagger},\ B'_j]_{\pm}]_{\mp} = 0, \qquad i \neq j, \tag{12d}$$

$$\tfrac{1}{4}\sum_{i\neq j} [A_i,\ [A_i^{\dagger},\ A'_j]_{\pm}]_{\pm} + \tfrac{1}{4}[A_j,\ [A_j^{\dagger},\ A'_j]_{+}]_{+} = A'_j, \tag{12e}$$

$$\tfrac{1}{4}\sum_{i\neq j} [B_i,\ [B_i^{\dagger},\ B'_j]_{\pm}]_{\pm} + \tfrac{1}{4}[B_j,\ [B_j^{\dagger},\ B'_j]_{+}]_{+} = B'_j, \tag{12f}$$

$$[A_i,\ [A'_j,\ A''_k]_{\pm}]_{\mp} = 0, \qquad i \neq j,\ i \neq k,\ j \neq k, \tag{12g}$$

$$[B_i,\ [B'_j,\ B''_k)_{\pm}]_{\mp} = 0, \qquad i \neq j,\ i \neq k,\ j \neq k, \tag{12h}$$

$$[A_i,\ [A'_j,\ B_k]_{\mp}]_{\pm} = 0, \qquad i \neq j, \tag{12i}$$

$$[B_i,\ [B'_j,\ A_k]_{\pm}]_{\mp} = 0, \qquad i \neq j, \tag{12j}$$

$$[A_i,\ [A_i^{\dagger},\ B_j]_{\mp}]_{\pm} = 0, \tag{12k}$$

$$[B_i,\ [B_i^{\dagger},\ A_j]_{\pm}]_{\mp} = 0, \tag{12l}$$

$$\tfrac{1}{4}\sum_{i} [A_i,\ [A_i^{\dagger},\ B_j]_{\mp}]_{\mp} = B_j, \tag{12m}$$

$$\tfrac{1}{4}\sum_{i} [B_i,\ [B_i^{\dagger},\ A_j]_{\pm}]_{\pm} = A_j, \tag{12n}$$

$$[A_i, B_jB_j]_- = 0, \tag{12o}$$

$$[B_i, A_jA_j]_- = 0, \tag{12p}$$

$$[A_i, B_jB_j^\dagger]_\mp = [A_i, B_jB_k']_\mp = 0, \qquad j \neq k, \tag{12q}$$

$$[B_i, A_jA_j^\dagger]_\mp = [B_i, A_jA_k']_\mp = 0, \qquad j \neq k. \tag{12r}$$

In Eqs. (12c)–(12n) each of the four indicated possibilities have, *a priori*, to be considered for each set of values of the subscripts i and j, or, i, j, and k, with the subscripts being unequal where so restricted. A similar statement applies with respect to the two possibilities which can in principle occur in each one of Eqs. (12q) and (12r). The various alternatives can *a priori* be realized depending on whether the component fields concerned commute or anticommute for equal times. If for two distinct component fields obeying ordinary statistics a bilinear equal-time commutation assignment is assumed or determined, it will, of course, be reflected in those of Eqs. (12), where the matrices associated with these two fields enter, and where several possibilities have in principle to be considered, depending on the bilinear equal-time commutation behavior of the two component fields under consideration. Furthermore, the matrices associated with any component fields occur in several of the above matrix equations. The various alternatives indicated in different equations are, of course, not independent when they apply to the same matrices. If the locality between the component fields is assumed to be "normal", the upper signs are obtained in those equations where several possibilities can be realized.

In Eqs. (12) no summation is to be performed over repeated indices unless expressly indicated. The subscripts of the A matrices in Eqs. (12) can assume the values $1, \cdots, m$, while the subscripts of the B matrices run from 1 to n, where m and n are integers which are determined by explicit matrix representations satisfying Eqs. (12). Furthermore, the symbols A_i can, independently of the A's with a different subscript and independently of the B's, stand either for the matrix A_i or for its "conjugate" $A_i^\dagger$. A similar statement applies to the symbols A_j', B_i, B_j' etc., and it is understood that for all subscripts

$$(A_i^\dagger)^\dagger = A_i, \tag{13a}$$

and

$$(B_i^\dagger)^\dagger = B_i. \tag{13b}$$

Eqs. (12a), (12b), (12o), and (12p) imply that, irrespective of the bilinear equal-time commutation relations between any two distinct component fields, the square of each one of the matrices under consideration commutes with all the other matrices.[14]

A GENERALIZATION OF THE TRILINEAR COMMUTATION RELATIONS

The trilinear commutation relations (9) do not properly take into account the matrix structure of the generalized fields. It is expedient to regard the trilinear commutation relations as transformations of the fields χ and χ'. Eqs. (9) can symbolically be expressed in the form

$$T_\psi[\chi, \chi']_- = (\psi, \chi)\chi' - (\psi, \chi')\chi, \tag{14a}$$

$$T_\phi[\chi, \chi']_+ = (\phi, \chi)\chi' + (\phi, \chi')\chi. \tag{14b}$$

In particular,

$$T_\psi[\bar\psi, \phi]_- = (\psi\bar\psi + \bar\psi\psi)\phi - \psi\phi\bar\psi - \bar\psi\phi\psi = (\psi, \bar\psi)\phi. \tag{15}$$

Equation (15) indicates that contributions to the contraction factor, symbolically denoted by $(\psi, \bar\psi)$, occurring on the right-hand side of Eq. (15) come from terms in which the field ϕ stands between the fields ψ and $\bar\psi$, and from terms in which ϕ stands to the right (or left) of the fields ψ and $\bar\psi$.

The labels attached to the component fields specify the matrices with which the components respectively are associated. The matrices describe the interdependence of the component fields and, as will be shown below, also completely determine the bilinear equal-time commutation relations of the components. Since the generalized fields do not necessarily obey bilinear commutation relations, it is unrealistic to expect that, under the operations T_ψ and T_ϕ, the components of the generalized fields will not be shuffled. Indeed, Eq. (15) suggests that the commutation relations (9) should properly be generalized to

$$\begin{aligned}[\psi_\alpha(\mathbf{x}), [\chi_\beta(\mathbf{x}'), \chi_\gamma'(\mathbf{x}'')]_-]_- \\ = \tfrac{1}{2}\,\delta(\mathbf{x} - \mathbf{x}')(\gamma_4')_{\alpha\beta}(\chi_\gamma'(\mathbf{x}'') \pm M\chi_\gamma'(\mathbf{x}'')M^{-1}) \\ - \tfrac{1}{2}\,\delta(\mathbf{x} - \mathbf{x}'')(\gamma_4')_{\alpha\gamma}(\chi_\beta(\mathbf{x}') \pm M\chi_\beta(\mathbf{x}')M^{-1})\end{aligned} \tag{16a}$$

[14] If instead of Eqs. (10) type-one fields were to be represented as linear superpositions of boson fields with numerical matrix coefficients and type-two fields as linear superpositions of fermion fields with matrix coefficients, or if boson fields with matrix coefficients were to be added to the right-hand side of Eqs. (10a) and (10b) and fermion fields with matrix coefficients were to be added to the right-hand side of Eqs. (10c) and (10d), the numerical matrix equations obtained in analogy to Eqs. (12) no longer imply that the square of each matrix coefficient commutes with all the other matrices, unless all the matrices vanish.

and

$$[\phi_\mu(\mathbf{x}), [\chi_\nu(\mathbf{x}'), \chi'_\rho(\mathbf{x}'')]_+]_- = \tfrac{1}{2}i\,\tilde{\delta}_{\mu\nu}\,\delta(\mathbf{x}-\mathbf{x}')(\chi'_\rho(\mathbf{x}'') \pm N\chi'_\rho(\mathbf{x}'')N^{-1}) + \tfrac{1}{2}i\,\tilde{\delta}_{\mu\rho}\,\delta(\mathbf{x}-\mathbf{x}'')(\chi_\nu(\mathbf{x}') \pm N\chi_\nu(\mathbf{x}')N^{-1}). \quad (16b)$$

In order that Heisenberg's equations of motion for free fields be obtainable from Eqs. (16), it is necessary that the trilinear commutation (11a) and (11d), first discussed by H. S. Green,[7,15] be contained in Eqs. (16) as special cases. This will be the case if

$$\tfrac{1}{2}(\psi \pm M\psi M^{-1}) = \psi, \quad (17a)$$

$$\tfrac{1}{2}(\phi \pm N\phi N^{-1}) = \phi, \quad (17b)$$

i.e., fields of type one either commute or anticommute with the matrix M, and fields of type two either commute or anticommute with N. Furthermore, regardless of the numerical transformation matrices M and N, Eqs. (16) also imply the following commutation relations considered by Kibble and Polkinghorne[2]:

$$[\psi(\mathbf{x}), [\phi(\mathbf{x}'), \phi'(\mathbf{x}'')]_+]_- = 0, \quad (18a)$$

$$[\phi(\mathbf{x}), [\psi(\mathbf{x}'), \psi'(\mathbf{x}'')]_-]_- = 0. \quad (18b)$$

As far as commutation relations are concerned, only Eqs. (11a), (11d), (18a), and (18b) are necessary for the construction of Hilbert spaces for free fields.

In those trilinear equal-time commutation relations in which the generalized field quantities are either all kinematically related or all kinematically unrelated, the transformation matrices M and N do not explicitly occur. Hence, in Eqs. (11), only Eqs. (11b) and (11e) have to be modified. Similarly, in Eqs. (12) only Eqs. (12m) and (12n) must be changed in an obvious manner.

It turns out that, at least for the representations to be discussed, the matrix structure of the generalized fields and the bilinear equal-time commutation relations between the component fields are derived from those of Eqs. (12) which are homogeneous in the matrices, i.e., from those matrix equations in which the matrices M and N do not explicitly appear. Equations (16) therefore have to be regarded as equations for M and N with the generalized fields given. This procedure is somewhat analogous to the procedure referred to in the classical Lagrangian formalism as the "method of undetermined multipliers."[16]

[15] E. P. Wigner, Phys. Rev. 77, 711 (1950).

[16] Cf., for example, C. Lanczos, *The Variational Principles of Mechanics* (The University of Toronto Press, Toronto, 1949), p. 43 *et seq.*

In fact, the commutation relations (16) can be readily obtained from the action principle. The numerical matrices M and N, in addition to satisfying Eqs. (16) and (17), must also satisfy the following relations, which are obtained with the aid of Eqs. (3a), (3b), (4a), and (4b):

$$[[\psi(\mathbf{x}), G(\psi)]_-, \chi(\mathbf{x}')]_- = i[\delta\psi(\mathbf{x}), \chi(\mathbf{x}')]_- = i[\delta\psi(\mathbf{x}), \tfrac{1}{2}(\chi(\mathbf{x}') \pm M\chi(\mathbf{x}')M^{-1})]_-, \quad (19a)$$

$$[[\bar{\psi}(\mathbf{x}), G(\bar{\psi})]_-, \chi(\mathbf{x}')]_- = -i[\delta\bar{\psi}(\mathbf{x}), \chi(\mathbf{x}')]_- = -i[\delta\bar{\psi}(\mathbf{x}), \tfrac{1}{2}(\chi(\mathbf{x}') \pm M\chi(\mathbf{x}')M^{-1})]_-, \quad (19b)$$

$$[[\phi(\mathbf{x}), \tilde{G}(\phi)]_-, \chi(\mathbf{x}')]_+ = i[\delta\phi(\mathbf{x}), \chi(\mathbf{x}')]_+ = i[\delta\phi(\mathbf{x}), \tfrac{1}{2}(\chi(\mathbf{x}') \pm N\chi(\mathbf{x}')N^{-1})]_+, \quad (19c)$$

and

$$[[\pi(\mathbf{x}), \tilde{G}(\pi)]_-, \chi(\mathbf{x}')]_+ = -i[\delta\pi(\mathbf{x}), \chi(\mathbf{x}')]_+ = -i[\delta\pi(\mathbf{x}), \tfrac{1}{2}(\chi(\mathbf{x}') \pm N\chi(\mathbf{x}')N^{-1})]_+. \quad (19d)$$

If in the process of deriving commutation relations from the action principle, i.e., if in the derivation of Eqs. (5) and (6) use is made of Eqs. (19), the trilinear commutation relations (16) are obtained instead of Eqs. (9).

In general, the matrices M and N are expected to depend on the generalized fields in whose trilinear commutation relations M or N appear.

ALGEBRAIC CONSIDERATIONS

In view of Eqs. (12), the algebra generated by a set of matrices C_i will be studied under the assumption that the generators satisfy the equations

$$[C_i^2, C_j]_- = 0, \quad \text{for all } i \text{ and } j, \quad (20)$$

and for $i \neq j$, $i \neq k$, $j \neq k$

$$[C_i, [C_j, C_k]_\pm]_\pm = 0. \quad (21)$$

In Eq. (21) all the four indicated possibilities compatible with trilinear algebraic identities are considered for any set of values of the subscripts i, j and k.

Eqs. (20) and (21) imply that for all values of i, j, and k

$$[[(C_iC_jC_iC_j)^n + (C_jC_iC_jC_i)^n], C_k]_- = 0, \quad (22)$$

where n is an arbitrary (nonnegative) integer. Equation (22) is valid regardless of which of the alternatives indicated in Eq. (21) are realized by any triplet of matrices.

Eqs. (20) and (21) imply that, for any irreducible matrix representation of operators satisfying Eqs. (20) and (21),

$$C_i^2 = d_i I \tag{23}$$

and

$$C_iC_jC_iC_j + C_jC_iC_jC_i = d_{ij}I = d_{ji}I, \tag{24}$$

where d_i and d_{ij} are numerical factors and I is the unit matrix.

Eqs. (21) and (23) suggest that the generators of an irreducible representation can be normalized in such a manner that their square, unless it vanishes, is equal to unity, i.e., for any irreducible representation,

$$C_i^2 = I \text{ or } 0. \tag{25}$$

The normalization (25) implies that for the algebra in general the generators can be chosen in such a manner that for all i

$$C_i^4 = C_i^2. \tag{26}$$

Any algebraic expression of order higher than m^2 in the generators of an irreducible representation of the algebra generated by m generators can, with the aid of Eqs. (21), (23), and (24), be reduced to a linear superposition of terms, none of which is of order higher than m^2 in the C_i's. Any irreducible matrix representation of the algebra generated by a finite number of operators satisfying Eqs. (20) and (21) is therefore of finite dimensionality.

In order to satisfy Eqs. (16), the trilinear commutation relations satisfied by the generating matrices must be chosen and the numerical factors occurring on the right-hand side of Eqs. (23) and (24) must be assigned in such a manner that, not for all the associated irreducible matrix representations, a bilinear commutation behavior of each pair of generators is also implied. Moreover, two distinct component fields obeying ordinary statistics are supposed to either commute or anticommute for equal times but not both. Furthermore, it should be possible to define a relationship between A_i and $A_i^\dagger$ and between B_i and $B_i^\dagger$ in such a manner that Eqs. (16) are satisfied. These requirements single out the possibilities which have to be examined.

Two special cases will presently be considered: (A) The case of three generators. (B) The case of five generators. In each case any triplet of generators is supposed to satisfy trilinear commutation relations of the type indicated in Eq. (21). Moreover, in both cases the generators are required to be divisible into two subsets in such a manner that each generator belonging to either subset commutes with the product, taken in any order, of any two generators belonging to the other subset. In case (A) C_1 and C_2 belong to one subset and C_3 to the other, whereas in case (B) C_1, C_2, and C_3 belong to one subset and C_4 and C_5 to the other. Specifically, the generators are defined by the following equations:

$$C_i^2 = I, \begin{cases} i = 1, 2, 3 & \text{for case (A)}, \\ i = 1, \cdots, 5 & \text{for case (B)}, \end{cases} \tag{27}$$

and

$$C_iC_jC_iC_j + C_jC_iC_jC_i = -2I, \tag{28a}$$

if C_i and C_j ($i \neq j$) belong to the same subset, while

$$C_iC_jC_iC_j + C_jC_iC_jC_i = 0, \tag{28b}$$

if C_i and C_j do not belong to the same subset. Furthermore, (if no misunderstanding can arise the generating matrices will simply be indicated by their subscripts only):

$$\text{For case (A): } [1, [2, 3]_-]_- = [2, [3, 1]_-]_- = [3, 12]_- = [3, 21]_- = 0. \tag{29a}$$

$$\text{For case (B): } [1, [2, 3]_-]_- = [2, [3, 1]_-]_- = [3, [1, 2]_-]_- = 0; \tag{29b}$$

and nine equations of the form

$$[i, [j, k]_-]_- = [j, [k, i]_-]_- = [k, ij]_- = [k, ji]_- = 0, \tag{29c}$$

where C_k does not belong to the subset containing C_i and C_j.

From Eqs. (27), (28), and (29), the following bilinear commutation relations are obtained:

$$\text{For case (A): } [1, 2]_+ = 0. \tag{30a}$$

$$\text{For case (B): } [1, 2]_+ = [1, 3]_+ = [2, 3]_+ = [4, 5]_+ = 0; \tag{30b}$$

i.e., any two generators contained in the same subset anticommute. However, Eqs. (27), (28), and (29) also imply that any two generators C_i and C_j belonging to different subsets cannot possibly satisfy a bilinear commutation relation of the form

$$C_iC_j \pm C_jC_i = kI, \tag{31}$$

where k is a numerical factor equal to or different from zero. The above equations also imply that

$$C_iC_jC_k = -C_kC_jC_i \quad \text{for } i \neq j, \ i \neq k, \ j \neq k. \tag{32}$$

The linearly independent elements of the two irreducible representations of the algebra under consideration can be chosen in such a way that the square of each is a multiple of the identity:

Case (A) 16 linearly independent matrices

I

$1, 2, 3$

$12 - 21, 13 \pm 31, 23 \pm 32$

$123, 132, 131, 313, 323$

$1313, 1323.$

Case (B): 64 linearly independent matrices

I

$1, 2, 3, 4, 5$

$12 - 21, 13 - 31, 23 - 32, 45 - 54$

$14 \pm 41, 15 \pm 51, 24 \pm 42, 25 \pm 52, 34 \pm 43, 35 \pm 53$

123

$(12 - 21)4, (12 - 21)5, (13 - 31)4, (13 - 31)5, (23 - 32)4, (23 - 32)5$

$(45 - 54)1, (45 - 54)2, (45 - 54)3$

$142, 143, 243, 152, 153, 253, 415, 425, 435$

$141, 151, 414, 424, 434$

$(12 - 21)(45 - 54), (13 - 31)(45 - 54), (23 - 32)(45 - 54)$

1414

$1425, 1435, 2435$

$12(34 \pm 43), 12(35 \pm 53)$

$1415, 4142, 4143, 4243$

$41234, 41235, 12345.$

The generators can be chosen to be Hermitian. For example,

Case (A):

$$C_1 = \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix}; \quad C_2 = \begin{bmatrix} \sigma_y & 0 \\ 0 & \sigma_x \end{bmatrix}; \quad C_3 = \begin{bmatrix} 0 & \sigma_x \\ \sigma_x & 0 \end{bmatrix}, \tag{33a}$$

where σ_x and σ_y are Pauli matrices.

Case (B):

$$C_1 = \begin{bmatrix} 0 & \gamma_1 \\ -\gamma_1 & 0 \end{bmatrix}; \quad C_2 = \begin{bmatrix} 0 & \gamma_2 \\ -\gamma_2 & 0 \end{bmatrix}; \quad C_3 = \begin{bmatrix} 0 & \gamma_3 \\ -\gamma_3 & 0 \end{bmatrix}, \tag{33b}$$

$$C_4 = \begin{bmatrix} \gamma_4 & 0 \\ 0 & \gamma_5 \end{bmatrix}, \quad C_5 = \begin{bmatrix} \gamma_5 & 0 \\ 0 & -\gamma_4 \end{bmatrix}, \tag{33c}$$

where the 4×4 gamma matrices are conventionally defined by

$$\bar{\gamma}_\mu \bar{\gamma}_\nu + \gamma_\nu \gamma_\mu = 2 g_{\mu\nu} I, \tag{34a}$$

$$-g_{11} = -g_{22} = -g_{33} = g_{44} = 1, \tag{34b}$$

and

$$\bar{\gamma}_5 \equiv i \gamma_1 \gamma_2 \gamma_3 \gamma_4. \tag{34c}$$

It is expedient to select for consideration all those of the above listed elements of the two representations which are of odd order in the generators and to make the following definitions:

Case (A):

$$A_1 = A_1^\dagger = k1; \qquad B_1 = B_1^\dagger = k3,$$
$$A_2 = A_2^\dagger = k2; \qquad B_2 = B_2^\dagger = ik132,$$
$$A_3 = A_3^\dagger = k313; \qquad B_3 = B_3^\dagger = k131,$$
$$A_4 = A_4^\dagger = k323; \qquad B_4 = B_4^\dagger = ik123. \tag{35a}$$

Case (B):

$$A_{1,1} = A_{1,1}^\dagger = k1; \qquad B_{1,1} = B_{1,1}^\dagger = k4,$$
$$A_{1,2} = A_{1,2}^\dagger = ik415; \qquad B_{1,2} = B_{1,2}^\dagger = k5,$$
$$A_{1,3} = A_{1,3}^\dagger = k414; \qquad B_{1,3} = B_{1,3}^\dagger = k141,$$
$$A_{1,4} = A_{1,4}^\dagger = ik145; \qquad B_{1,4} = B_{1,4}^\dagger = k151,$$
$$A_{2,1} = A_{2,1}^\dagger = k2; \qquad B_{2,1} = B_{2,1}^\dagger = ik124,$$
$$A_{2,2} = A_{2,2}^\dagger = ik425; \qquad B_{2,2} = B_{2,2}^\dagger = ik125,$$
$$A_{2,3} = A_{2,3}^\dagger = k424; \qquad B_{2,3} = B_{2,3}^\dagger = ik142,$$
$$A_{2,4} = A_{2,4}^\dagger = ik245; \qquad B_{2,4} = B_{2,4}^\dagger = ik152,$$
$$A_{3,1} = A_{3,1}^\dagger = k3; \qquad B_{3,1} = B_{3,1}^\dagger = ik234,$$
$$A_{3,2} = A_{3,2}^\dagger = ik435; \qquad B_{3,2} = B_{3,2}^\dagger = ik235,$$
$$A_{3,3} = A_{3,3}^\dagger = k434; \qquad B_{3,3} = B_{3,3}^\dagger = ik243,$$
$$A_{3,4} = A_{3,4}^\dagger = ik345; \qquad B_{3,4} = B_{3,4}^\dagger = ik253,$$
$$A_{4,1} = A_{4,1}^\dagger = ik123; \qquad B_{4,1} = B_{4,1}^\dagger = ik134,$$
$$A_{4,2} = A_{4,2}^\dagger = k41235; \qquad B_{4,2} = B_{4,2}^\dagger = ik135,$$
$$A_{4,3} = A_{4,3}^\dagger = ik41234; \qquad B_{4,3} = B_{4,3}^\dagger = ik143,$$
$$A_{4,4} = A_{4,4}^\dagger = k12345; \qquad B_{4,4} = B_{4,4}^\dagger = ik153. \tag{35b}$$

The matrices listed above have the following properties: The square of each is equal to $k^2 I$, where k is a normalization factor. Any two A's, belonging to the same representation, either commute or anticommute. The same is true with respect to each pair of B's. No A satisfies a bilinear commutation relation with a B. Any triplet of matrices belonging to the same representation satisfies trilinear commutation relations of the type exhibited in Eq. (21). Moreover, each A commutes or anticommutes with the product, taken in any order, of each pair of B's. The same is true for each B with respect to the product of any two A's.

The A's will be associated with fermions and the B's with bosons. If the opposite procedure is adopted, which can be done, no new results are obtained. Tables I–IV, which summarize the bilinear equal-time commutation relations between the component fields for the case that the A's are associated with fermions and the B's with bosons, are identical to or can be transformed (by suitable permutations of rows and columns) into the tables obtained when the A's are associated with bosons and the B's with fermions.

Only those of Eqs. (12) which are homogeneous in the matrices have so far been used explicitly. These homogeneous equations also determine the bilinear equal-time commutation relations between the component fields. Equations (16) or (9) imply the following commutation rules: If the matrices respectively associated with two fermion fields commute, the fermions anticommute for equal-times. If the matrices anticommute, the fermions commute. If the matrices respectively associated with two boson fields commute, the bosons commute for equal-times. If the matrices anticommute, the bosons anticommute. For bilinear equal-time commutation relations involving a fermion and a boson field, there are two possibilities in each case. For each representation a fermion and a boson field selected at random can be assumed to commute or to anticommute for equal times. The bilinear equal-time commutation relations between all fermion and boson fields associated with the same representation are then determined. The following alternatives can, for example, be considered:

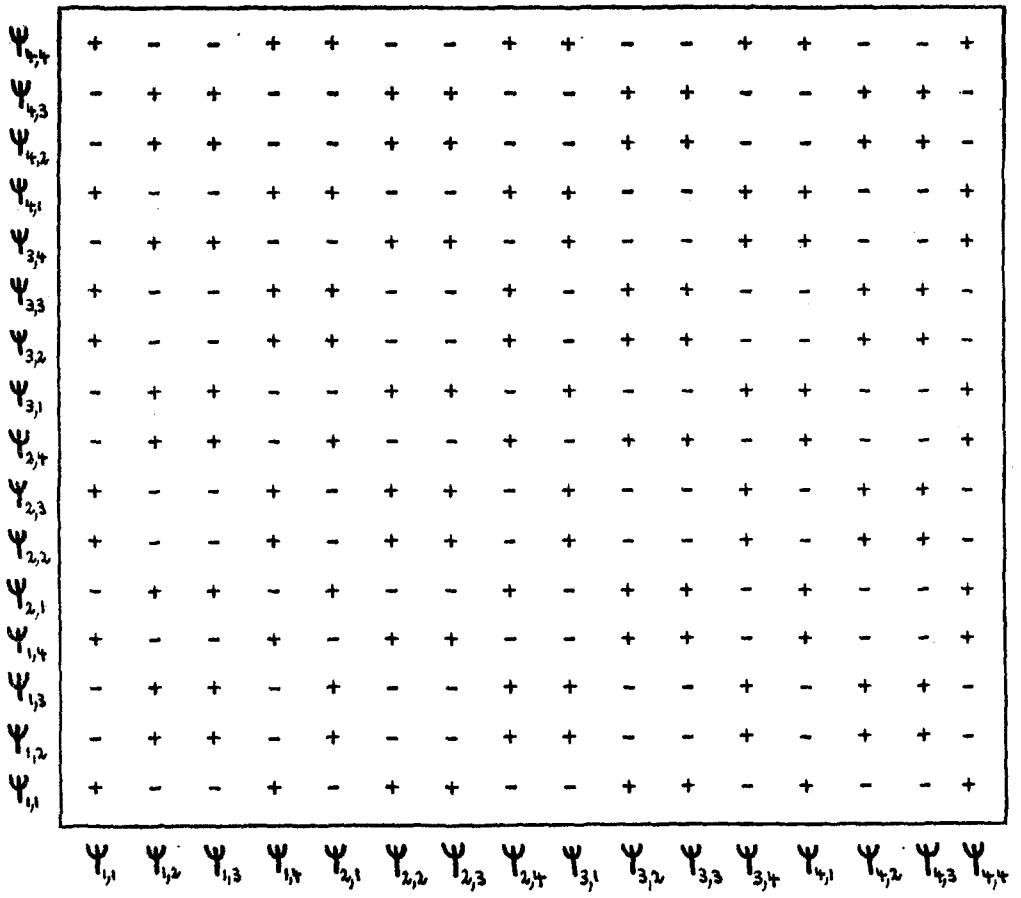
$$[\psi_1(\mathbf{x}), \phi_1(\mathbf{x}')]_- = 0, \tag{36a}$$

$$[\psi_1(\mathbf{x}), \phi_1(\mathbf{x}')]_+ = 0, \tag{36b}$$

TABLE I. Bilinear equal-time commutation relations between the component fields for case (A).

	ψ_1	ψ_2	ψ_3	ψ_4	ϕ_4	ϕ_3	ϕ_2	ϕ_1
ψ_1	+	−	−	+	+	+	−	−
ψ_2	−	+	+	−	+	+	−	−
ψ_3	−	+	+	−	−	−	+	+
ψ_4	+	−	−	+	−	−	+	+
ϕ_4	+	+	−	−	−	+	+	−
ϕ_3	+	+	−	−	+	−	−	+
ϕ_2	−	−	+	+	+	−	−	+
ϕ_1	−	−	+	+	−	+	+	−

TABLE II. Bilinear equal-time commutation relations between the fermions [case (B)].



$$[\psi_{1,1}(\mathbf{x}), \phi_{1,1}(\mathbf{x}')]_- = 0, \tag{36c}$$

$$[\psi_{1,1}(\mathbf{x}), \phi_{1,1}(\mathbf{x}')]_+ = 0, \tag{36d}$$

where ψ_1, ϕ_1, $\psi_{1,1}$, and $\phi_{1,1}$ are the component fields associated with the matrices A_1, B_1, $A_{1,1}$, and $B_{1,1}$, respectively.

The bilinear equal-time commutation relations between the component fields for the case of three generating matrices are summarized in Table I, and for the case of five generators in Tables II–IV. Table I is based on alternative (36a) and Table IV on alternative (36c). If alternatives (36b) and (36d) were respectively adopted instead, all the signs in the lower left-hand quadrant of Table I and in Table IV have to be reversed while the other tables are not affected (a + sign indicates that the two-component fields concerned anticommute for equal-times). Actually, it makes no difference which one of the alternatives (36a) and (36b) is adopted for case (A), because the tables associated with one alternative can be transformed into the tables associated with the other alternative by suitably permuting rows or columns. A similar statement applies to alternatives (36c) and (36d) of case (B).

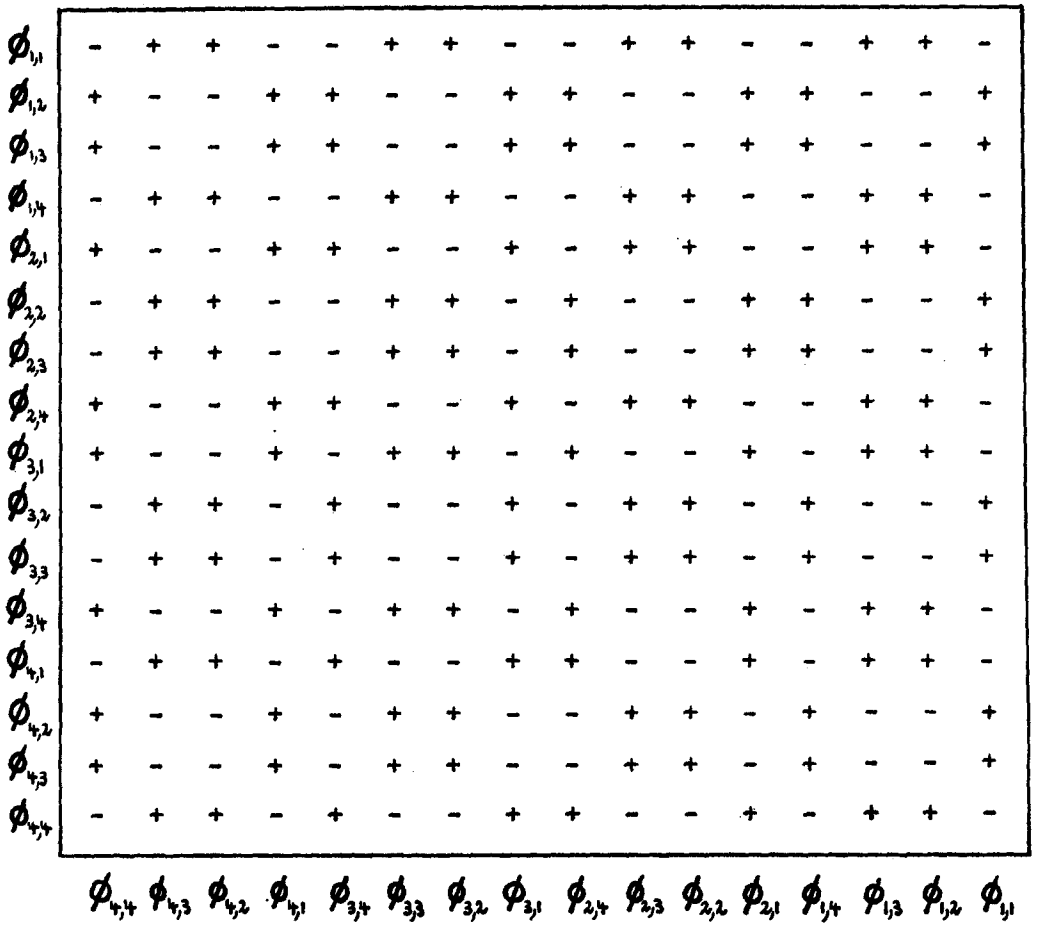
Tables I–IV show that, if the commutation relations of a component field with itself are taken into account, each component field commutes for equal times with as many component fields as it anticommutes with. Commutation and anticommutation relations thus naturally occur in a symmetric way, each row and each column in the tables containing + and − signs in equal numbers.

The simplest assumption consistent with Eqs. (16) concerning A_i and $A_i^\dagger$ (and B_i and $B_i^\dagger$) has been made in Eqs. (35), namely that they are equal. Other possibilities can be investigated. Because of the assumption made, a component field and its adjoint have identical equal-time commutation behavior with respect to all the other component fields.[17]

When each one of the matrices of Eqs. (35) is multiplied by its component field, two sets of gen-

[17] In this connection see G. F. Dell'Antonio, Ann. Phys. (N.Y.) 16, 153 (1961).

TABLE III. Bilinear equal-time commutation relations between the bosons [case (B)].



eralized fields of different dimensionality are obtained, provided the component fields satisfy the bilinear equal-time commutation relations given in Table I in the case of the four-dimensional representation, and in Tables II–IV in the case of the eight-dimensional representation. It is a simple matter to verify that, for each triplet of the resulting generalized field quantities associated with the same representation, it is possible to construct nonsingular matrices M and N in such a manner that Eqs. (16), (17), and (19) are satisfied. Each generalized field thus contains only one component. This means that, as a free field, each component can be considered independently of the other component fields. However, for the four-dimensional representation, four different component fields are readily available for each spin.[18] For the eight-dimensional representation sixteen component fields differing from each other in their bilinear equal-time commutation relations are similarly available for each spin.[18] For the eight-dimensional representation, several component fields can conveniently be grouped together into one generalized field, provided the normalization factor k in Eqs. (35b) is suitably adjusted. For example, applying Eqs. (10a) and (10b) to those A matrices multiplied by their associated fermion field components which, in Eqs. (35b) and Tables II and IV, have the same second subscript, one obtains four generalized type-one fields, each having four components. Similarly, all the B's with the same second subscript and their associated bosons can be grouped into four generalized type-two fields. For each triplet of generalized fields so obtained, it is possible to construct matrices M and N in such a manner that Eqs. (16), (17), and (19) are satisfied. For example, for the two generalized fields

$$\psi = 2^{-\frac{1}{2}} \sum_{i=1} A_{i,1}\psi_{i,1} \tag{37a}$$

[18] The matrices (35) associated with the fermion and boson fields are linearly independent. If the requirement of linear independence is relaxed, a greater number of component fields can be obtained, and the tables summarizing the bilinear equal-time commutation relations would consequently have to be enlarged.

and

$$\phi = 2^{-\frac{1}{2}} \sum_{i=1}^{4} B_{i,1}\phi_{i,1}, \tag{37b}$$

all relevant commutation relations are satisfied provided $k = \frac{1}{2}$,

$$M = M^{-1} = C_1C_2C_3C_4C_5, \tag{38a}$$

$$N = N^{-1} = C_4, \tag{38b}$$

and the + signs are chosen on the right-hand sides of Eqs. (16) and (19).

Higher dimensional representations of operators satisfying the commutation relations (16) can be generated from the four-dimensional representation by forming the direct products of the matrices given in Eqs. (35a) with the Pauli matrices or with the gamma matrices.

Table I indicates that, any two distinct fermions associated with the four-dimensional representation, have either identical or opposite bilinear equal-time commutation behavior with all boson components. In the eight-dimensional representation, there is a wider range of possibilities, as Table IV shows. This difference between the two representations is reflected in the interactions of the component fields respectively associated with the two representations.

INTERACTIONS

The requirement that all relevant variational expressions possess right-left symmetry with respect to the coefficients and positions of the variations necessitates the introduction of a numerical matrix into trilinear interactions of generalized fields.[6] For example, if the numerical matrix K in the local trilinear interaction Lagrangian density

$$L_{\text{int.}} = [\bar{\psi}_a(\mathbf{x}), \Gamma\psi_b(\mathbf{x})]_-[K, \phi(\mathbf{x})]_+ \quad \text{Hermitian conjugate} \tag{39}$$

has the property that

$$[\bar{\psi}_a(\mathbf{x}), [K, \phi(\mathbf{x})]_+]_- = [\psi_b(\mathbf{x}), [K, \phi(\mathbf{x})]_+]_- = 0, \tag{40}$$

TABLE IV. Bilinear equal-time commutation relations between fermion and boson fields [case (B)].

	$\Psi_{1,1}$	$\Psi_{1,2}$	$\Psi_{1,3}$	$\Psi_{1,4}$	$\Psi_{2,1}$	$\Psi_{2,2}$	$\Psi_{2,3}$	$\Psi_{2,4}$	$\Psi_{3,1}$	$\Psi_{3,2}$	$\Psi_{3,3}$	$\Psi_{3,4}$	$\Psi_{4,1}$	$\Psi_{4,2}$	$\Psi_{4,3}$	$\Psi_{4,4}$
$\phi_{1,1}$	−	−	+	+	−	−	+	+	−	−	+	+	−	−	+	+
$\phi_{1,2}$	−	−	+	+	−	−	+	+	−	−	+	+	−	−	+	+
$\phi_{1,3}$	+	+	−	−	+	+	−	−	+	+	−	−	+	+	−	−
$\phi_{1,4}$	+	+	−	−	+	+	−	−	+	+	−	−	+	+	−	−
$\phi_{2,1}$	+	+	−	−	+	+	−	−	−	−	+	+	−	−	+	+
$\phi_{2,2}$	+	+	−	−	+	+	−	−	−	−	+	+	−	−	+	+
$\phi_{2,3}$	−	−	+	+	−	−	+	+	+	+	−	−	+	+	−	−
$\phi_{2,4}$	−	−	+	+	−	−	+	+	+	+	−	−	+	+	−	−
$\phi_{3,1}$	−	−	+	+	+	+	−	−	+	+	−	−	−	−	+	+
$\phi_{3,2}$	−	−	+	+	+	+	−	−	+	+	−	−	−	−	+	+
$\phi_{3,3}$	+	+	−	−	−	−	+	+	−	−	+	+	+	+	−	−
$\phi_{3,4}$	+	+	−	−	−	−	+	+	−	−	+	+	+	+	−	−
$\phi_{4,1}$	+	+	−	−	−	−	+	+	+	+	−	−	−	−	+	+
$\phi_{4,2}$	+	+	−	−	−	−	+	+	+	+	−	−	−	−	+	+
$\phi_{4,3}$	−	−	+	+	+	+	−	−	−	−	+	+	+	+	−	−
$\phi_{4,4}$	−	−	+	+	+	+	−	−	−	−	+	+	+	+	−	−

the variations, which are obtained in the process of deriving the equations of motion from Eq. (39), can be moved to the right or left in such a manner that for type-one field variations the resulting coefficients are equal except for sign, while for type-two field variations the coefficients are equal including sign, which corresponds to the symmetrization of the free-field Lagrangians of the two types of fields. Furthermore, the total Lagrangian density must have a self-consistent matrix structure, so that, as the net result, a compatible set of scalar equations of motion for the component fields is obtained. As a consequence, the two fermions respectively associated with the generalized type-one fields in the interaction (39) must have identical equal-time bilinear commutation relations with the boson associated with the generalized type-two field entering the interaction: if the fermions commute for equal times they must anticommute with the boson; if the fermions anticommute they must commute with the boson field. If the two fermion fields in Eq. (39) have different first subscripts, Table IV indicates that, for the eight-dimensional representation, the boson entering the interaction can, for any suitable integral spin, be realized *a priori* in four different ways. Furthermore, if the subscripts a and b in Eq. (39) are set equal, there are, for each half-integral spin value in the eight-dimensional representation, eight fermions which can, and eight fermions which cannot, enter into the resulting minimal interactions with any specific type-one field of suitable integral spin.

The matrix

$$C_1C_3C_1C_3 = C_2C_3C_2C_3 \tag{41a}$$

anticommutes with all the matrices (35a). Similarly, the matrix

$$\begin{aligned} C_1C_4C_1C_4 &= C_1C_5C_1C_5 = C_2C_4C_2C_4 \\ &= C_2C_5C_2C_5 = C_3C_4C_3C_4 = C_3C_5C_3C_5 \end{aligned} \tag{41b}$$

anticommutes with all the matrices (35b). As a consequence, the fermions respectively associated with the matrices A_i and $C_3A_iC_3$ (and, similarly, the bosons respectively associated with B_i and $C_1B_iC_1$), where A_i is any one of the A's given in Eqs. (35a), have opposite bilinear equal-time commutation behavior with respect to all the other component fields associated with the same representation [cf., Table I]. A similar statement applies to the matrices A_{ij} and $C_4A_{ij}C_4$ in the case of the eight-dimensional representation [Eqs. (35b) and Tables II–IV]. As a consequence, in the case of the eight-dimensional representation, the fermions respectively associated with A_{ij} and $C_4A_{ij}C_4$ cannot "decay" into each other via a trilinear interaction.

The selection rules mentioned are, of course, valid regardless of which of the alternatives (36) are selected for consideration in each case.

CONCLUSIONS

The irreducible representations of generalized field operators considered not only specify the bilinear equal-time commutation relations between the component fields in an essentially unique manner, but also contain information concerning the number of different ordinary fields which can be expected in each case.[18] Since the trilinear equal-time commutation relations are specified, no Klein-transformations to the "normal case" can be effected for the fermion and boson fields associated with the matrices (35), and in this respect the component fields differ from conventional fields.[19–27] For the representations discussed

$$[\psi_\alpha(\mathbf{x}),\ \psi'^\dagger_\beta(\mathbf{x}')]_+ = \delta(\mathbf{x}-\mathbf{x}')(\gamma'_4)_{\alpha\beta}I \tag{42a}$$

and

$$[\phi_\mu(\mathbf{x}),\ \phi'_\nu(\mathbf{x}')]_- = i\,\delta(\mathbf{x}-\mathbf{x}')\,\delta_{\mu\nu}I, \tag{42b}$$

i.e., kinematically related and unrelated generalized type-one field quantities obey equal-time anti-commutation relations and kinematically related and unrelated generalized type-two field quantities obey equal-time commutation relations. Although generalized field quantities of different types neither commute nor anticommute for equal times, all physical causality requirements are satisfied, because, for the representations considered, any two observables commute for spacelike separations of the arguments. Because of the matrix structure of the generalized fields, the bilinear equal-time commutation relations between the component fields appear in a new context.

The possible physical relevance of any representations of field operators satisfying the trilinear commutation relations (16) and the interactions of these fields are currently being investigated.

[19] O. Klein, Zh. Techn. Fiz. (USSR) **9**, 1 (1938).
[20] S. Oneda and H. Umezawa, Progr. Theoret. Phys. (Kyoto) **9**, 685 (1953).
[21] T. Kinoshita, Phys. Rev. **96**, 199 (1954).
[22] H. Umezawa, *Quantum Field Theory* (North-Holland Publishing Company, Amsterdam, 1956), p. 197.
[23] R. Spitzer, Phys. Rev. **105**, 1919 (1957).
[24] G. Lüders, Z. Naturforsch. **13a**, 254 (1958).
[25] T. Kinoshita, Phys. Rev. **110**, 978 (1958).
[26] H. Araki, J. Math. Phys. **2**, 267 (1961).
[27] For an illuminating comment concerning the normal case compare the lecture notes of A. S. Wightman in *Theoretical Physics* (International Atomic Energy Agency, Vienna, 1963), footnote on p. 29.

On the Existence of Field Theory. I. The Analytic Approach

J. G. TAYLOR
Department of Physics, Rutgers, The State University, New Brunswick, New Jersey
(Received 16 June 1965)

The problem of existence of solutions to local field equations is studied. We set up the field equations so that the solutions correspond to fixed points of a mapping of the space of Green's functions into themselves. We attempt to use analytic methods to determine these fixed points, in particular, the contraction mapping principle. To do this we perform a rotation to Euclidean space from Lorentz space; in Euclidean space we prove the existence of solutions to a large class of approximating equations to the field equations, obtained by requiring the Green's functions to be zero if they have more than a certain number of external particles. By this method we prove that there is only the tirvial zero solution to certain types of bootstrap equations. The contraction mapping theorem does not appear powerful enough to discuss the complete field equations.

1. INTRODUCTION

SINCE the field equations were introduced thirty-five years ago[1] as a natural way to describe processes involving the creation or annihilation of particles, there has been a lively discussion as to whether or not these equations are self-consistent. The high-energy divergences which initially arise in their perturbation expansion were thought to lead to their inconsistency. In other words, it was claimed that it was not possible to reconcile the usual theory of quantum mechanics and the theory of special relativity with a formalism involving local interactions between particles at well-defined points of space–time. This was especially evident, since all physical quantities were infinite when calculated to high enough orders of perturbation theory.

The renormalization procedure for quantum electrodynamics, developed after 1947,[2] enabled finite results to be obtained, with remarkable experimental agreement, for all observable quantities. This process still left a nasty taste in many mouths, nor did it apply to strong, weak, or gravitational interactions.

For both these reasons, two new approaches to field theory were attempted, these being the axiomatic field theory[3] and the S-matrix theory.[4] These approaches tried to use the essential properties contained in local field equations. It was hoped that these properties would be strong enough to give dynamical predictions, but weak enough to avoid the high-energy divergences. However, these approaches are still not able to give a complete dynamical theory of elementary particles.

We would like to suggest here that, of the many possible ways of attempting to set up a useful and complete dynamical theory, the one most likely to lead to the sure ground of experimental facts is that which tries to understand the present success of quantum electrodynamics by taking a closer look at the methods used to achieve this success. This means giving a careful discussion of the equations of conventional field theory.

A first step in this discussion has been given earlier,[5] where the equations of conventional field theory were maneuvered into a form—the canonical form, which most likely avoids high-energy divergences—and discussed further. Great difficulty was met in actually solving the equations. In such a situation, it may be useful to consider the existence of solutions to the equations. This is what we attempt in this and succeeding papers. In the process of this investigation, it is hoped that we will gain insight into the nature of the solutions, if they exist, and into the approximate methods which may be set up to obtain their numerical value.

We found in Ref. 5 that the Green's functions equations arising from a particular conventional field theory are an infinite set of coupled nonlinear integro–differential equations. They are coupled in such a way that we cannot obtain an exact set of

[1] W. Heisenberg and W. Pauli, Z. Physik 56, 1 (1929).
[2] See, for example, the collected papers in *Quantum Electrodynamics*, J. Schwinger, Ed. (Dover Publications, Inc., New York, 1958).
[3] See, for example, A. S. Wightman and R. Streater, "*PCT, Spin and Statistics and All That* (W. A. Benjamin Company, Inc., New York, 1961).
[4] See, for example, G. Chew, *S-Matrix Theory of Strong Interactions* (W. A. Benjamin Company, Inc., New York, 1961).
[5] J. G. Taylor, Nuovo Cimento Suppl. 1, 857 (1963). The six papers contained there will be referred to as I–VI. See also K. Symanzik, *Lectures in High-Energy Physics*, B. Jakšić Ed., (Federal Nuclear Energy Commission of Yugoslavia, Zagreb, 1961), pp. 485–517.

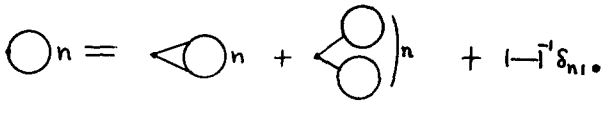
equations for any finite subset of the functions. Evidently, the methods we use have to be as powerful as modern mathematical techniques allow; we find that we even have to improve some of these techniques in order to apply them to our problems. Our main tool is the fixed-point theorems of both the analytical and topological type.

Our basic approach is constructive; we take a particular set of Green's functions equations and attempt to find suitable mathematical conditions to be imposed so that a solution of the equations exists. This approach is the opposite to that of the axiomatic field theorists. In our approach, we may start by talking nonsense, but we have to make good sense out of the nonsense in the end; in axiomatic theory, nonsense is ruled out from the beginning. In our making sense out of nonsense, we use any analyticity properties which seem consistent with our equations and useful for solving them; such analyticity is only proved sensible in the final steps of our discussion.

That is the purpose of this and following papers. Our discussion is, of necessity, highly mathematical, depending on techniques of topology, functional analysis, and functions of many complex variables. We cannot avoid this heavy use of mathematical techniques; we feel that it is necessary to actually prove results about this very complex set of equations, since almost anything can be conjectured about them (as has been the case in the past).

2. THE EQUATIONS

We will consider in this paper only field theories with no high-energy divergences in perturbation theory. In particular, let us consider a single neutral scalar particle of mass m in a space–time of one time and one space dimension. The field theory we take will be given by the interaction Lagrangian $\mathcal{L}_I = g\phi^3$ for the field ϕ of the particle. Upon canonical quantization, the resulting quantized field equations give rise to a set of coupled nonlinear integral equations connecting the various Green's functions which can be defined in terms of ϕ. These equations were developed and discussed in Ref. 5. We will use the same notation as that of Ref. 5, so that these Green's functions equations (GFE's) become [see Eq. (59) of I]



(1)

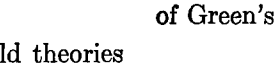
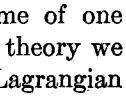
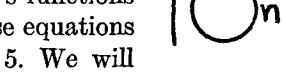
In (1) the bubble  multiplied by $\delta^4(\sum p_i)$ denotes the connected part of the Fourier transform

$$\int \prod dx_j \, e^{-i\sum p_j x_j} K_{x_j} \langle 0 | \, T(\phi(x_1) \cdots \phi(x_n)) \, | 0 \rangle,$$

where $K_x = (\Box_x^2 - m^2)$, and so is the off-mass-shell continuation of the S-matrix elements given by the field ϕ. The internal lines denote $-i\tilde{D}_F(p) = i[-p_0^2 + \mathbf{p}^2 + m^2 - i\varepsilon]^{-1}$. There is integration over all momenta carried by internal lines, and there is energy momentum conservation at each vertex, which denotes $-g$.

We note that $\bigcirc^n$ may be regarded as the (formal) sum of all Feynman diagrams for a process with exactly n internal particles, but with the external bare propagators for these particles removed.

We do not have to consider here the canonical or renormalized form of field equations discussed in Ref. 5, since our Eq. (1) for a convergent theory will be equivalent to this form. When we turn to theories with high-energy divergences, it will be necessary to discuss the canonical form only, since (1) will very likely contain the high-energy divergences. Also, we will consider only the connected part of the Green's functions, since the disconnected parts will contain δ-functions which would make our discussion of existence more difficult.

How do we go about proving the existence of a solution to (1)? We are interested in the infinite set of Green's functions  , and may regard each of these Green's functions as belonging to a vector space of functions. The space of functions  will be denoted by E_n; we will not specify it immediately, but assume it has a topology τ_n in which E_n is a locally convex topological vector space.[6] Thus, we are interested in a point $x =$  in the product space $E = \prod_{n\geq 2} E_n$.

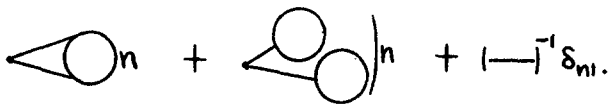
It is natural to take on E the product topology $\tau_\pi = \prod_{n\geq 2} \tau_n$, though we will discuss this and other topologies on E later. We may write Eq. (1) in the form

$$x = T(x), \qquad x_n = T_n(x), \tag{2,2'}$$

[6] We use the definitions in N. Bourbaki, *Elements de Mathematique, Livre V, Espaces Vectoriels Topologiques* (Hermann & Cie., Paris, 1954), Chaps. I–V.

where

$T_n(x) =$



So we see that a solution to the field equation (1) is a fixed point of the mapping T, and vice-versa. T is a mapping of E into itself, so our problem is to find a suitable space E and a suitable topology on E so that T has a fixed point in E. In the mathematical discussion of fixed-point theory,[7,8] there are two distinct avenues of approach. One is analytical, via the contraction mapping theorem, the other is topological, via the definition of the topological degree of a mapping.

The contraction mapping theorem justified the iteration approach to Eq. (2), and, besides giving an existence theorem, it proves uniqueness and provides an approximation scheme. The properties required for the theorem to apply are, however, much stronger than can be satisfied by many mappings, so that we will have to be prepared to relinquish the analytical approach if we still wish to talk about solving the actual equations.

The topological approach only proves existence in general, so that both uniqueness and how to approximate the solution are not going to come out of the analysis immediately. However, at least as a first step, we would like to prove existence. To do that our mapping still has to have suitable properties, or rather we have to be able to choose E and τ so that it has these properties, the main one of which is the complete continuity[9] of T. This is the property which we later find not to be satisfied by T for any (E, τ), and we therefore have to extend the topological approach to accommodate our case. But this will not be done in this paper.

For either of the approaches mentioned, it is necessary to have a topology on E so that T is at least continuous. In considering the type of topology best suited for our needs, we would naturally attempt to make estimates of the size of $T(x)$, or of $T(x) - T(y)$ as x approaches y. This immediately brings us to the difficulty that it is not possible to estimate the sizes of terms on the right-hand side of (1) or (2) in a simple fashion because (a) the principal-value singularities arising in the internal propagator lines; (b) the Lorentz group is locally compact but not compact, so that power counting does not apply. Thus, an integral like

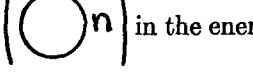
$$\int d^4 k[(k^2)^2 + 1]^{-137}$$

is divergent if k^2 is the Lorentz-invariant $k_0^2 - \mathbf{k}^2$.

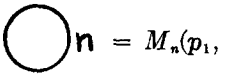
Both of these difficulties are inherent in the Lorentz metric of momentum space. We would avoid these difficulties in a Euclidean space, so we would like to continue (1) and (2) to a Euclidean metric. This will be done, in the next section, by a generalized Wick rotation[10] through 90° in the complex energy plane, $k_0 \to ik_0$, in *all* energy variables, both internal and external in (1).

3. ANALYTIC CONTINUATION

The GFE (1) is an equation relating functions evaluated at real momenta only. (1) may only have solutions which possess no analyticity in these momenta. In order to achieve the generalized Wick rotations mentioned above, we have to search for solutions which possess some analyticity in energy. If there are none, then other methods will have to be developed. But until we fail, we will search for analytic solutions. What analyticity may we expect for solutions of Eq. (1)? To answer that we need to determine the region of analyticity which is preserved by T in (2). We will naturally start by looking for the largest such region of analyticity, which is given by

Theorem 1: The largest region of analyticity of the set of Green's functions  in the energy variables and consistent with T is a product of cut planes, the cuts starting from the normal thresholds.

In detail, if



then M_n is analytic in the product of complex planes for p_{i0} $(i = 1, \cdots, n)$ with cuts given by

$$p_{I0} = \sum_{i \in I} p_{i0} \text{ real}, \quad |p_{I0}| \geq (\mathbf{p}_I^2 + 4m^2)^{\frac{1}{2}},$$

where I is any sub-interval of the interval $[1, n]$ in the set of integers, together with possible poles at $p_I^2 = m^2$.

[7] J. Cronin, "Fixed Points and Topological Degree in Nonlinear Analysis," Amer. Math. Soc. Math. Surveys, No. 11 (1964).

[8] K. Krasnoselskü, *Topological Methods in the Theory of Nonlinear Integral Equations* (Pergamon Press, London, 1964).

[9] Ref. 7, p. 131.

[10] G. C. Wick, Phys. Rev. **96**, 1124 (1954).

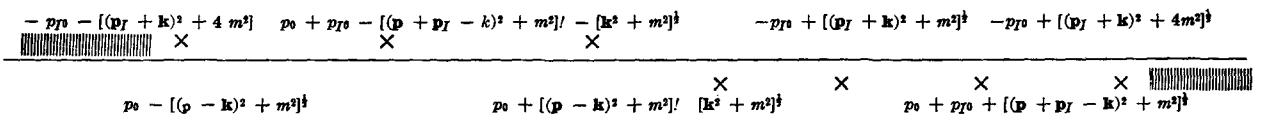


FIG. 1. The position of singularities in the k_0 plane arising in the integrand of Eq. (3).

As may be seen by inspection, this analyticity is exactly that satisfied by every term in perturbation theory and contains no anomalous thresholds. The latter only arise when continuation is also made in the space-parts $\mathbf{p}_i$ of the momenta.

We prove the theorem as follows. We assume that x is composed of components  , each of which has the analyticity stated in the theorem. This is then substituted into the right-hand side of (1), and the analyticity of the resulting terms investigated. We wish to show that the same cut-plane analyticity results.

Since this is immediate in the second and third terms on the right of (1), we just investigate the first term. We do not discuss the problem of convergence of the integral, but leave this to be made more precise when we consider the continuation problem in more detail in a later paper. We use the technique of distortion of the integration contour[11] to avoid pinches or end-point singularities as the external energy variables are moved about in their complex planes. The initial singularities in the k_0-plane for the quantity

$$= n = g \int dk[k^2 - m^2]^{-1}$$

$$\times[(p - k)^2 - m^2]^{-1} M_{n+2}(k, p - k, p_1, \cdots, p_n) \quad (3)$$

are given in Fig. 1. We remark here that, in order to distort the contour of integration away from moving poles, we must assume that the amplitude on the right-hand side of (3) is evaluated from above the cuts for positive energy and below the cuts for negative energy, exactly as if m^2 is replaced by $m^2 - i\epsilon$ everywhere, as in the Feynman convention.

We now allow p_0 and any number of the p_{i0} ($i = 1, \cdots, n$) to wander around in their complex planes. The singularities of the integral can arise from pinches of the contour of integration. For the pole singularities, these can be seen to arise when one or more of the following situations arises:

(i) $[\mathbf{k}^2 + m^2)^{\frac{1}{2}} = p_0 - [(\mathbf{p} - \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,
$-p_{I0} - [(\mathbf{p}_I + \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,
or $p_0 + p_{I0} - [(\mathbf{p} - \mathbf{k} + \mathbf{p}_I)^2 + m^2]^{\frac{1}{2}}$,

(ii) $p_0 + [(\mathbf{p} - \mathbf{k})^2 + m^2]^{\frac{1}{2}} = -[\mathbf{k}^2 + m^2]^{\frac{1}{2}}$,
$-p_{I0} - [(\mathbf{p}_I + \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,
or $p_0 + p_{I0} - [(\mathbf{p} - \mathbf{k} + \mathbf{p}_I)^2 + m^2]^{\frac{1}{2}}$,

(iii) $-p_{I0} + [(\mathbf{p}_I + \mathbf{k})^2 + m^2]^{\frac{1}{2}} = -[\mathbf{k}^2 + m^2]^{\frac{1}{2}}$,
$p_0 - [(\mathbf{p} - \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,
$p_0 + p_{J0} - [(\mathbf{p} - \mathbf{k} + \mathbf{p}_J)^2 + m^2]^{\frac{1}{2}}$,
or $-p_{J0} - [(\mathbf{p}_J + \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,

(iv) $p_0 + p_{I0} + [(\mathbf{p} + \mathbf{p}_I - \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,
$= -[\mathbf{k}^2 + m^2]^{\frac{1}{2}}$,
$p_0 - [(\mathbf{p} - \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,
$-p_{J0} - [(\mathbf{p}_J + \mathbf{k})^2 + m^2]^{\frac{1}{2}}$,
or $p_0 + p_{J0} - [(\mathbf{p} + \mathbf{p}_J - \mathbf{k})^2 + m^2]^{\frac{1}{2}}$.

One of these will happen provided that a subset I of the interval $(0, 1, \cdots, n)$ has $p_{I0} = \sum_{i \in I} p_{i0}$ with p_{I0} real, $|p_{I0}| \geq [\mathbf{p}_I^2 + 4m^2)^{\frac{1}{2}}$. Exactly the same set of conditions arises if pinches occur due to the wandering of the branch points. Further singularities arise, independently of the integration, from poles and branch points in the fixed external variables. When all these singularities are put together, we obtain exactly the set of singularities specified by the theorem.

We may now perform the analytic continutation of our mapping T. As we continue in our external variables $p_{i0} \to ip_{i0}$ in an anticlockwise direction, we may rotate the internal k_0 integration contour at the same time and in the same direction without crossing any of the singularities. After a 90° rotation the singularities in the k_0-plane are now displaced, and in fact lie outside a band around the imaginary k_0 axis of width $2m$, as shown in Fig. 2. We have thus removed the principal-value singularities, and achieved our continuation to the Euclidean region.

We define the set of points x in E whose components have the analyticity in energy described in Theorem 1 to be energy-analytic. Then our

[11] R. J. Eden in *Brandeis Summer Institute in Theoretical Physics, 1962* (W. A. Benjamin Company, Inc., New York, 1962).

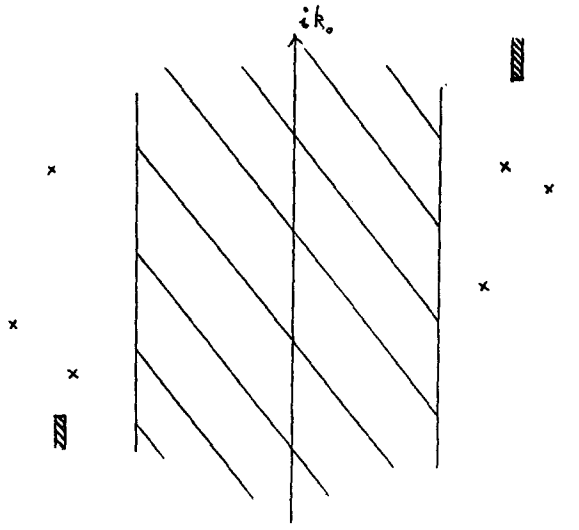


FIG. 2. The position of singularities in the k_0 plane arising in the integrand of Eq. (3) after rotation of the external energy variables through 90° in their complex planes.

previous discussion has proved, to within contributions from infinity, the following theorem.

Theorem 2: A necessary condition that the mapping T defined by (2) has a fixed point in the set of energy-analytic elements is that the corresponding Euclidean mapping $\bar{T}$ has a fixed point.

The theorem extends immediately to the case of an interaction $\mathcal{L}_I = \sum g_n\phi^n$, following the remarks after the proof of Theorem 1.

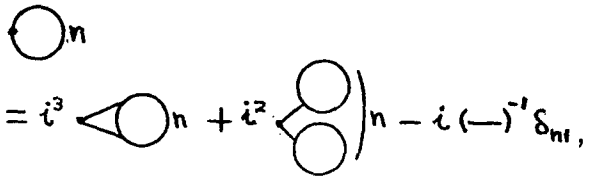
It would be nice if the existence of a fixed point for $\bar{T}$ was also a sufficient condition for the existence of a fixed point of T; this is a very difficult question which we are not yet in a position to discuss. One of the problems in discussing it is in the contributions from infinity which may arise during the analytic continuation of T to $\bar{T}$. We cannot attempt to discuss these contributions without going into a full discussion of the reverse continuation from the Euclidean to the Lorentz metric. We will do this in a later paper.

We note that the position of the threshold is not important for achieving the Wick rotation (provided it is not zero). Thus the Wick rotation can be done for either the unrenormalized equations (1) or for the renormalized equations that we will write down later [Eqs. (12)]. There will in fact be a mass shift in the unrenormalized equations so that the thresholds do not involve m (the bare mass) but the renormalized mass m_r. The energy analytic functions entering Theorems 1 and 2 for the renormalized field equations will naturally have their thresholds involving m_r. Since here we will not be using the particular space of analytic functions on which T is defined, except to enable the Wick rotation to be done, we need not concern ourselves further with the question of thresholds.

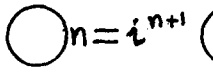
We assume in the remainder of this paper that we are working in the Euclidean region, and that we consider the mapping $\bar{T}$. The main reason for doing this is to ascertain which methods will apply to prove the existence of a fixed point for $\bar{T}$; the problem for T is harder, but our understanding gained in our discussion of $\bar{T}$ should help us when we finally face up to T.

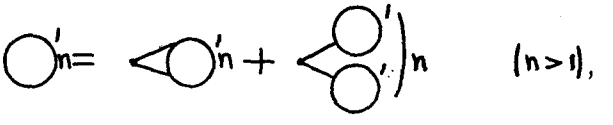
4. THE CONTRACTION MAPPING

We now wish to apply the contraction mapping theorem[12] to $\bar{T}$. To do this we need to introduce a norm on E. We will attempt to do this in a number of stages; we will start by trying to gain an idea of the type of norm useful in considering an approximation to $\bar{T}$, and then attempting to remove the approximation. Before we do this we will discuss the reality properties of $\bar{T}$. On performing the analytic continuation, we see that each internal line gives $i[p_0^2 + \mathbf{p}^2 + m^2]^{-1}$ and each internal integration gives an extra i. Thus in the Euclidean region (1) becomes



(4)

where now each line denotes $[p_0^2 + \mathbf{p}^2 + m^2]^{-1}$. If we introduce  then

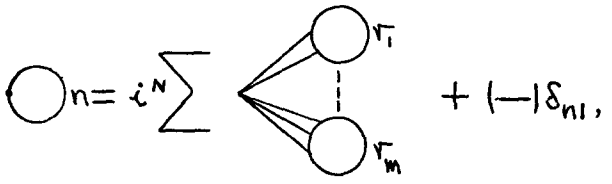


(4′)

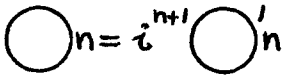
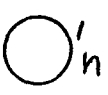
Thus (4′) becomes a real mapping, so we may search for solutions of (4′) when the function spaces E_n are composed of real functions only.

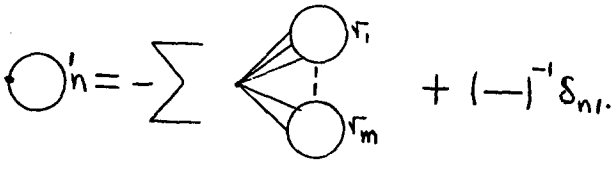
[12] Ref. 7, p. 141.

The same result follows for an interaction Lagrangian $\mathfrak{L}_I = g\phi^N$ for $N = 4, 5$, etc. The Green's functions equations in the general case are



(5)

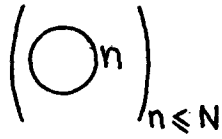
where the summation in (5) is over all sets of integers $(t_1, \cdots, t_m)$, $(r_1, \cdots, r_m)$ so that $\sum t_i = N$, $\sum r_i = n$, and then over all possible m. The vertex denotes g. In Euclidean space the general term in (5) has an extra factor $i^{N+\sum(t_i-1)}$, and if we define as before  , then the general term in the mapping for  has a factor i^{4N+n}, and we obtain the real equation

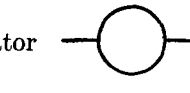


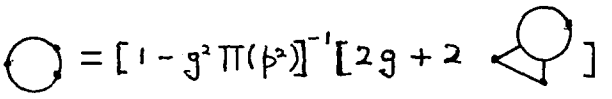
(5′)

Thus, we restrict ourselves to considering the mapping $\bar{T}$ on a product of real function spaces.

We now turn to approximating our mappings. The approximation we will use is to cut off the equations in (1) for large enough n. That is we will take  if $m > N$ on the right-hand side of (1) for $n > N$, and rearrange the resulting equations suitably to relate only the set of functions



Let us consider as a first approximation, the case $N = 3$, but even further approximate by taking the complete propagator  to be the free propagator (−). Then (1) becomes, in the Euclidean metric,



(6)

where $\prod(p^2) = \int d^2k[k^2 - m^2]^{-1}[(p-k)^2 + m^2]^{-1}$. $\prod(p^2)$ vanishes as $p^2 \to \infty$ and we choose $g^2 < [\max_p \prod(p^2)]^{-1} = g_0^2$. Equation (6) is a linear integral equation, but we are interested in applying the contraction mapping theorem. Evidently, if  is uniformly bounded in its external variables, it will preserve this property on the left-hand side of (6). If we denote this left-hand side by  , then for $\sup|M_3(p_1p_2p_3)| < \Lambda$, where $M_3 =$  , then

$$\sup|S(M_3)| < (1 - g^2/g_0^2)^{-1}(2g + 2g^2\alpha\Lambda), \quad (7)$$

where $\alpha = \sup|\Delta|$, Δ being the triangle function

$$\Delta = \Delta(p_1, p_2, p_3) = \int d^2k[k^2 + m^2]^{-1} \times [(p_1 - k)^2 + m]^{-1}[(p_2 + k)^2 + m^2]^{-1}.$$

In 2-dimensional space–time, α is finite. Further,

$$\sup|S(M_3') - S(M_3)| < 2(1 - g^2/g_0^2)^{-1}\alpha \sup|M_3' - M_3|. \quad (8)$$

If we now take E_3 to be the real Banach space of uniformly bounded real functions of three 2-vector momenta p_1, p_2, p_3 (satisfying $p_1 + p_2 + p_3 = 0$), with norm

$$||M_3|| = \sup|M_3(p_1p_2p_3)|$$

the supremum taken over all p_1, p_2, p_3 with $p_1 + p_2 + p_3 = 0$, then (7) and (8) show that S is a contraction mapping on E_3 which maps the closed ball $B_\Lambda = [M_3 : ||M_3|| \leq \Lambda]$ into itself provided $2\alpha g^2 < 1 - g^2/g_0^2$ and $2g + 2\alpha g^2\Lambda \leq \Lambda(1 - g^2/g_0^2)$. Thus if $g^2 < (2\alpha + g_0^{-2})^{-1} = g_c^2$ and $\Lambda \geq 2g(1 - g^2/g_c^2)^{-1}$ there will exist a unique solution in the ball B_Λ.

We have thus proved that Eq. (6) has a unique solution in the Banach space E_3, provided $g < g_c$. This solution is that obtained by iteration from any initial trial function on E_2, and, in particular, it is the convergent perturbation expansion of (6). In fact, we have, in the contraction mapping

theorem, a very effective tool for proving convergence of the perturbation expansion of our field equations.

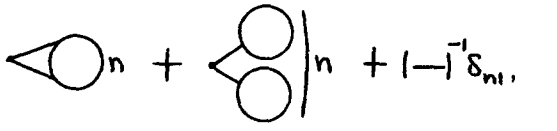
Let us now consider the higher approximations to (1). We make the Nth approximation as follows. We define the mapping $\bar{T}^{(N)}$ to be obtained by applying $\bar{T}$ to the vector $x^{(N)}$ where $x_i^{(N)} = 0 (i > N)$. We may thus regard $x^{(N)}$ as a vector in the subspace $E^{(N)} = \prod_{2 \leq n \leq N} E_n \times \{0\}$ of E, whose elements have their components beyond the Nth all zero. We then consider the first N components of $\bar{T}^{(N)}(x^{(N)})$ to be the components of $\bar{T}^{(N)}$. Thus if $p^{(N)}$ is the projection from E onto $E^{(N)}$, $\bar{T}^{(N)} = p^{(N)} \bar{T} p^{(N)}$. We wish to prove the existence of a fixed point of $\bar{T}^{(N)}$ in $E^{(N)}$, for each N.

Again, we let E_n be composed of bounded functions and given the uniformly bounded topology for $n > 2$, while for E_2 we take functions $M_2(p)$ for which $(p^2 + m^2)^{-1} M_2(p)$ is bounded with corresponding norm $||M_2|| = \sup (p^2 + m^2)^{-1} m^2 M_2(p)$. We impose on $E^{(N)}$ the product τ_π of the normed topologies on each $E_n (2 \leq n \leq N)$, so τ_π is a normed topology. The norm on $E^{(N)}$ may be taken as

$$||X^{(N)}|| = \sup_{1 \leq n \leq N} ||X_n^{(N)}||.$$

We now determine if $\bar{T}^{(N)}$ is a contraction mapping. We have

$\bar{T}_{n+1}^{(N)}(X^{(N)}) =$



(1′)

so

$$||\bar{T}_{n+1}^{(N)}(X^{(N)}) - \bar{T}_{n+1}^{(N)}(Y^{(N)})|| \leq g g_0^{-2} ||X_{n+2}^{(N)} - Y_{n+2}^{(N)}|| + g m^{-4} \sum_{r=1}^{n-1} {}^nC_r [||X_{r+1}^{(N)}|| \, ||X_{n-r+1}^{(N)} - Y_{n-r+1}^{(N)}|| + ||Y_{n-r+1}^{(N)}|| \, ||X_{r+1}^{(N)} - Y_{r+1}^{(N)}||]. \quad (9)$$

In (9) the factor ${}^nC_r$ has arisen from the symmetrization of the external momenta in ${}^nC_r$ ways for each division of n into r and $n - r$ momenta. If we now consider only elements $x^{(N)}$, $y^{(N)}$ in the ball $B_\Lambda = [x^{(N)}; ||x^{(N)}|| \leq \Lambda]$ we have

$$||\bar{T}^{(N)}(X^{(N)}) - \bar{T}^{(N)}(Y^{(N)})|| \leq \left(g g_0^{-2} + 2 g m^{-4} \Lambda \sup_{r \leq N} \sum_{r=1}^{n-1} {}^nC_r \right) ||X^{(N)} - Y^{(N)}|| \leq (g_0^{-2} + 2^N \Lambda m^{-4}) g \, ||X^{(N)} - Y^{(N)}||. \quad (10)$$

Further $\bar{T}^{(N)}(B_\Lambda) \subset B_\Lambda$ if

$$\left. \begin{aligned} g g_0^{-2} \Lambda + 2^N g m^{-4} \Lambda^2 &\leq \Lambda, \\ g g_0^{-2} \Lambda + m^{-2} &\leq \Lambda, \end{aligned} \right\}, \quad (11)$$

where the second equation in (11) comes from $n = 1$ in (1′). We thus have the conditions of the contraction mapping theorem satisfied if (11) is satisfied. The curves of (11) in the (g, Λ) plane are drawn in Fig. 3; we see that for any N there is a region in which both equations (11) are satisfied, for suitably small g. We notice that the uniqueness and existence for each g below the critical value g_c only holds for a suitable small B_Λ, and does not extend to the whole of $E^{(N)}$; this is characteristic of nonlinear mappings, when tackled in this manner. We have thus proved the following theorem.

Theorem 3: For each value of N, the approximate Euclidean field equation

$$\bar{T}^{(N)}(X) = X \quad (2'')$$

has a unique solution in some bounded region, in $E^{(N)}$, for an interval of values of the coupling constant g about the origin.

We note that our discussion of Eqs. (10) and (11) assumed g to be positive. If it is negative our previous discussion can go through with g replaced by $|g|$.

We may extend Theorem 3 to any nonlinear interaction in 2-dimensional space–time with no difficulty. The method is so similar that we do not give it here, but simply state the following.

Theorem 3′: For each value of N, the approximate Euclidean field equation

$$\bar{T}^{(N)}(X) = X$$

derived from the interaction Lagrangian $\mathfrak{L}_I = \sum_2^M g_n \phi^n$ in 2-dimensional space–time has a unique solution in some bounded region in $E^{(N)}$ for values of the coupling constants g_n contained in a set of nonzero intervals, each containing the origin.

We now are faced with two problems:

(a) Can we extend the existence theorem to larger values of g, for a given N?

(b) Does the interval for g converge to zero as $N \to \infty$?

These problems are evidently related; the answer to the second is that it does, from our conditions (11); the maximum value of g goes to zero when $N \to \infty$ as $4m^4/({}^N g_0^2)$. We may try to alleviate this difficulty by numerous methods: change the topology on $E^{(N)}$, on E_n, change B_Λ, change $\bar{T}^{(N)}$, and so on.

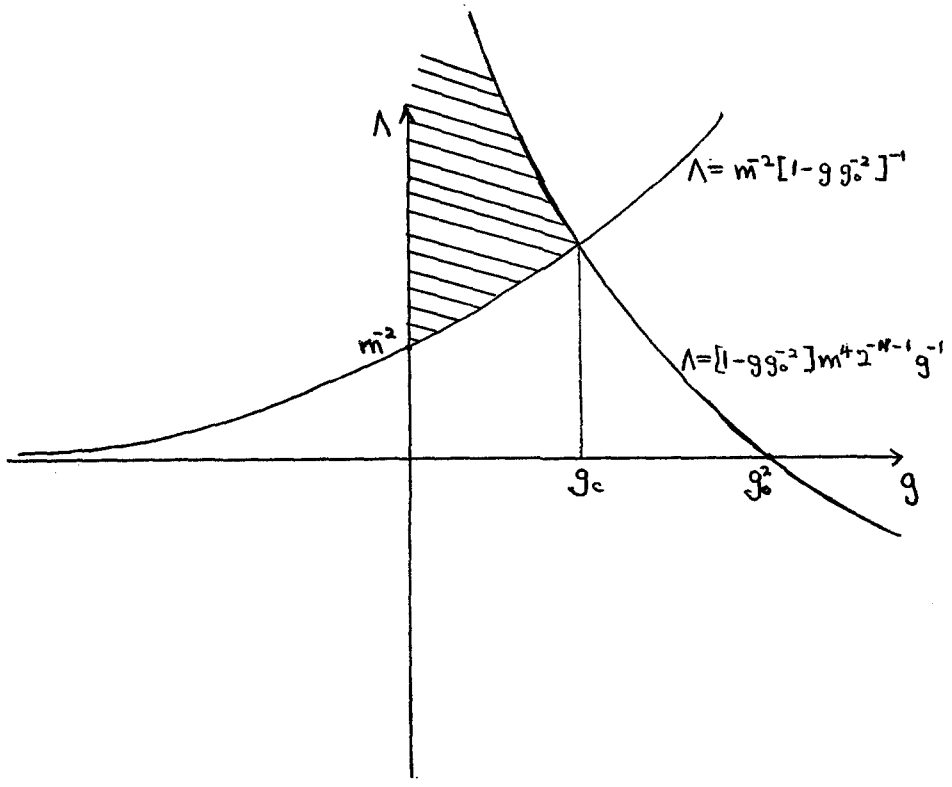


FIG. 3. The shaded region is the set of values of g and Λ satisfying Eqs. (10′) and (11′).

Let us first keep $\bar{T}$ and $\bar{T}^{(N)}$ and change the topologies. It seems that changing topologies on each E_n will not remove the troublesome factor 2^{N+1} in (10) and (11). What we need to do is to change the topology on $E^{(N)}$. This can be done by choosing the norm on $E^{(N)}$ to be $||x^{(N)}|| = \sup_{1 \le n \le N} ||\lambda_n^{-1} x_n^{(N)}||$, where λ_n are a suitable set of constants, depending on n in some way so that the contraction mapping principle will apply. With the new norm, conditions (10) and (11) for the contraction mapping theorem to apply are now

$$\left[g g_0^{-2} \lambda_{n+2} + 2 g m^{-4} \Lambda \sum_{r=1}^{n-1} \lambda_{r+1} \lambda_{n-r+1} {}^nC_r \right] < \lambda_{n+1}, \tag{10′}$$

$$g g_0^{-2} \lambda_3 \Lambda + m^{-2} \le \lambda_2 \Lambda, \tag{11′}$$

for $n \le N$. To answer question (b) above in the negative, we have to be able to satisfy (10′) and (11′) for all n and some nonzero g by a suitable choice of the sequence $\{\lambda_r\}$. But such a choice is not possible, as we see from the fact that, from (10′), $\lambda_{n+1} > a\lambda_{n+2} > \cdots > a^r \lambda_{n+1+r}$, where $a = g g_0^{-2}$. Thus, we have the summation term in (10′), (11′),

$$\sum_{r=1}^{n-1} \lambda_{r+1} \lambda_{n-r+1} {}^nC_r > a^n \lambda_{n+1}^2 \sum_{r=1}^{n-1} {}^nC_r > a^n 2^{n-2} \lambda_{r+1}^2. \tag{11″}$$

Again, from (10′) and (11′), we need $\lambda_{n+1} < c(2a)^{-n}$ for $c = 2m^4/(g\Lambda)$. But we also need from (10′), $c_1 \lambda_n (n - 1) < \lambda_{n+1}$, $c_1 = 2 g m^{-4} \Lambda \lambda_2$. These conditions $\lambda_{n+1} < O[(2a)^{-n}]$ and $\lambda_{n+1} > O[(n - 1)!]$ are evidently incompatible for large n, unless $g = 0$. Thus, no possible choice of product topology will enable us to prove the existence of a solution to (2″) for an arbitrarily large N.

We find this same problem however we change the topology on each E_n, at least for any of the L_p norms ($1 \le p \le \infty$). It is always the summation term on the right-hand side of (10′) which causes trouble. This is the same thing as saying that there are too many new graphs at each step of iteration in perturbation theory. That we cannot hope to prove the conditions of the contraction mapping principle are satisfied for arbitrarily large N follows from Jaffe,[13] who states that the perturbation expansion of our equations (1) is divergent, and the radius of convergence in g of the perturbation series is zero. For if we had been able to apply the contraction mapping principle, by obtaining a solution to (10′) and (11′), we would have been able to obtain a solution to (1) which would have been the perturbation expansion. To see how such a solution would have arisen, we remark that, if we

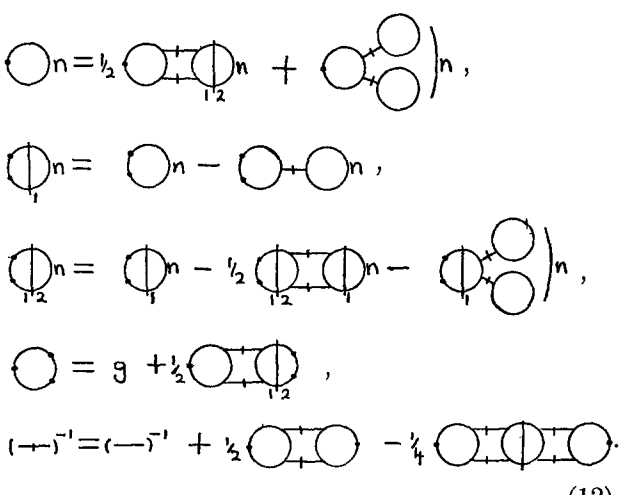
[13] A. Jaffe, Commun. Math. Phys. 1, 127 (1965).

take for E_n the space C_n of continuous bounded functions in the topology of uniform convergence (with suitable modifications for E_2), then it would be possible, by the contraction mapping theorem, to prove the existence of a unique solution of $T^{(N)}(x^{(N)}) = x^{(N)}$ in a compact $K^{(N)} = \prod_{1\le n\le N} K_n$. Here, each K_n is is composed of uniformly bounded (by $\lambda_n\Lambda$), equicontinuous functions vanishing uniformly at infinity (except for K_3, which would be the convex hull of such a compact in C_3 and the constant function g). But then $K = \lim_{N\to\infty} K^{(N)}$ is compact in $C = \prod C_n$ in the product topology, so the solutions have a convergent subsequence, converging to a point x_0 of K; such a point would be a solution of the complete field equation (2). x_0 would also be obtained as the limit of the perturbation series $x^{(N)}$, so proving the convergence of perturbation theory.

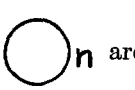
Jaffe's discussion was given for unrenormalized perturbation theory. We still have a chance of proving the convergence of perturbation theory for the renormalized equations. We turn to this in the next section.

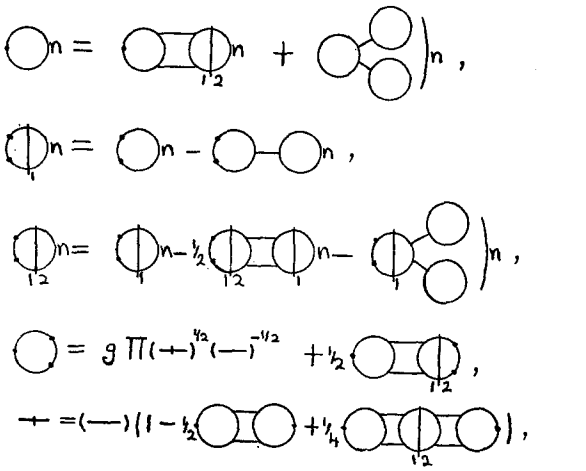
5. THE RENORMALIZED EQUATIONS

We are still working in 2-dimensional space–time, so that the renormalization procedure is just a reshuffling of the mapping T. As discussed in Ref. 5, the renormalized equations are



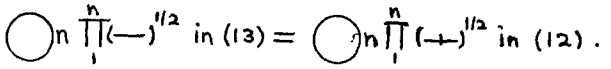
(12)

In (12) the functions  are completely amputated with respect to their external complete propagators. We can deal with (12) most conveniently if we replace the square root of the complete propagators in each diagram so that (12) now reads, in terms of partially amputated functions,

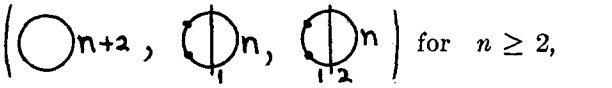


(13)

where the functions in (13) are related to those in (12) by



In the Euclidean metric, (12) is a real mapping; (13) need not be, due to the indefiniteness of $(+)^{\frac{1}{2}}$. However, this need not disturb us in our application of the contraction mapping theorem. Equation (13) is a mapping R on the space $E = \prod_n E_n$, where E_{n+2} is a space of triples of functions

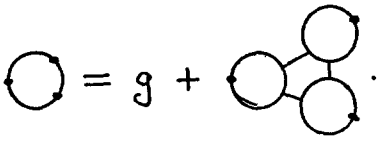


$$E_3 = \left|\bigcirc 3\right| \quad \text{and} \quad E_2 = \{+\}.$$

Then the solution of (13) satisfied

$$R(X) = X. \tag{13'}$$

We consider the corresponding Euclidean mapping $\bar{R}$, and the approximating mapping $\bar{R}^{(N)} = P^{(N)}\bar{R}P^{(N)}$. Let us consider the case $N = 2$, with $(-) = (+)$. The equation is



(14)

When $g = 0$ this equation may be regarded as the basis of off-the-mass-shell bootstrap equations (where our discussion will even go through for this equation in four dimensions due to the absence of divergences). If we apply the same considerations

to (14) as we did to (6), then (14) satisfies the contraction mapping theorem provided

$$\alpha\Lambda^3 + g \leq \Lambda, \qquad 3\Lambda^2\alpha < 1, \tag{15}$$

where α is the same constant as introduced in (7). Equation (15) will be satisfied provided $g < 2/3(3\alpha)^{\frac{1}{2}}$, and a suitable $\Lambda < (3\alpha)^{-\frac{1}{2}}$ may then be chosen to satisfy (15). We have thus proved that, for suitably small g, Eq. (14) has a unique solution equal to the convergent perturbation expansion; if $g = 0$ the only solution of the equation in the ball $B_\Lambda = [M_3 : ||M_3|| < (3\alpha)^{-\frac{1}{2}}]$ is the trivial vanishing solution. While this does not prove this to be the only solution, it does give an indication to this effect.

A similar theorem to Theorem 3 may be proved by exactly the same method for the mapping $\bar{R}^{(N)}$. We do not go into the details here, since they are so similar; we see immediately that such a result will have the same difficulties as was met in the previous section when we attempt to let $N \to \infty$. This arises from the last term on the right-hand side of (13), which has the same increasing number of terms when n increases, as does the second term on the right-hand side of (1). As in the last section, we cannot avoid this problem by suitably changing topologies.

Our existence and uniqueness theorem for $\bar{R}^{(N)}$ show that no bootstrap system is possible, at least in a certain bounded region of $E^{(N)}$.[14]

By bootstrap we mean the set of equations obtained by requiring there to be no local interaction for the bootstrapped particle to itself or other particles, so $g = 0$ for such a system. One can take a less stringent bootstrap with $g \neq 0$, but the wave-function renormalization constant Z of the particle being strapped set equal to zero. We do not have anything to say yet on the resulting system of equations.

[14] A similar result for R^4 has been proved by a very different approach by M. Broido (private communication).

6. CONCLUSIONS

We have seen in the preceeding sections that we may prove the existence and uniqueness of solutions to approximated field equations, for any order of approximation, by analytic methods. These methods are not strong enough to prove even the existence of solutions to the complete set of field equations. This result is rather disturbing, since the particular equations we have been discussing [except possibly for (14) in four dimensions] are far from the realistic cases, such as quantum electrodynamics. We have to develop methods for solving the existence problem which work, not only for our complete set of field equations (1) in two dimensions in the Euclidean metric, but also work when we analytically continue back to the Lorentz metric, and also include high-energy divergences. This we will start to do in the next paper of this series, where we also try to extend Theorem 3 so that a solution exists for all values of the coupling constant.

Evidently, as far as our existence program is concerned, the results of this paper are negative. It shows that the direct analytic approach is a blind alley. However, our existence and uniqueness results for Eq. (14) in four dimensions will be of value in off-the-mass-shell bootstrap and resonance approximation discussion.[15] It is exactly this equation (14) that arises in the resonance approximation, and it is encouraging to learn that this equation has a unique solution for at least some values of g.

[15] J. G. Taylor, "Composite, Elementary, and Resonance Particles" in *Proceedings of the International Conference on Elementary Particles, Sienna, September, 1963* (Italian Physical Society Publishers, Bologna, 1964).

The Structure of 4-Spinors

S. TEITLER
U. S. Naval Research Laboratory, Washington, D. C.
(Received 3 December 1965)

The possible underlying spinor spaces used in the description of Dirac 4-spinors are enumerated and classified within the framework of vector Clifford algebra. The relation with the usual matrix formulation is reviewed. Proper Lorentz transformation, space and time inversions are described in the Clifford algebra formalism. A representation-independent definition of charge sign conjugation is given.

I. INTRODUCTION

RECENTLY,[1,2] in an effort to describe fermion and boson one-particle fields from a unified viewpoint, we have reviewed, and in certain ways, extended Riesz's approach[3-5] to the description of fields by means of aggregates belonging to the Clifford algebra C_{16} generated with the basis vectors of Lorentz space–time L_4. This algebra has sixteen basis elements e^A; namely, a scalar unit, 1; four vector units e^λ; six bivector units $e^\lambda e^\mu \equiv e^{\lambda\mu}$, $(\lambda < \mu)$; four trivector units $e^{\lambda\mu\nu} \equiv e^\lambda e^\mu e^\nu$, $(\lambda < \mu < \nu)$; and the pseudoscalar unit $e^{0123} \equiv e^0e^1e^2e^3$. Here $\lambda, \mu, \nu = 0, 1, 2, 3$; the ascending order of indices has been chosen for specificity, and the algebra is defined by the anticommutation relations

$$e^\mu e^\lambda + e^\nu e^\mu = 2[g^{\mu\nu}]; \qquad (\mu, \nu = 0, 1, 2, 3). \tag{1.1}$$

Here, $[g^{\mu\nu}]$ is the Lorentz metric which we choose with diagonal elements $(1, -1, -1, -1)$ and vanishing nondiagonal elements. The products and commutation rules of the e^A derived using Eq. (1.1) are given in Table I.

As a means of unification in I, we considered various solutions to the Dirac-like equation

$$\left(\kappa - \sum_{\mu=0}^{3} ie^\mu \partial_\mu\right)\psi = 0, \tag{1.2}$$

where $\kappa = mc/\hbar$, and ψ is a Clifford aggregate or linear combination of Clifford basis elements. For example,

$$\kappa\Phi_s = \kappa\Phi + i\sum_{\mu=0}^{3} \varphi_\mu e^\mu \tag{1.3}$$

corresponds to a scalar field, and the Proca field is made up of a linear combination of vector and bivector components. Further, following Riesz,[3] Sommerfeld,[6] and Sauter[7] we generalized the usual Dirac column 4-spinor by identifying it with an element of a left-minimal ideal.

As may be recalled, a left ideal $\mathfrak{s}_L$ in an algebra C is a set of elements such that if $a \in \mathfrak{s}_L$, $b \in \mathfrak{s}_L$, then $(a + b) \in \mathfrak{s}_L$, and if $a \in \mathfrak{s}_L$ and $c \in C$, then $ca \in \mathfrak{s}_L$. A right ideal is defined similarly except that $\mathfrak{s}_R \ni ac$. A left-(right-) minimal ideal is a left (right) ideal which contains no other ideal but itself and the null ideal.

As has been discussed before[1,3] and as is shown more fully below, minimal ideals may be generated by using a primitive Clifford aggregate π such that, for any Clifford aggregate C, $\pi C\pi = t\pi$, where t is a scalar. For C_{16}, a certain class of four-component idempotents are primitive aggregates and may be used to generate other primitive aggregates. The relationship among primitive idempotents, other primitive aggregates, and minimal ideals of C_{16} are discussed in detail in Sec. II. There, the possible minimal ideals which correspond to 4-spinor spaces are enumerated and classified.

In Sec. III, we discuss the matrix representation for left- and right-minimal ideals and make some observations concerning the types of minimal-ideal basis. In Sec. IV we consider homogeneous Lorentz transformations, space and time inversions, and charge sign conjugation of the minimal-ideal basis.

II. IDEMPOTENTS, PRIMITIVE AGGREGATES, AND MINIMAL IDEALS

We consider idempotents made up of elements of equal weight, i.e., all of whose elements have square ± 1. Clearly there is only one one-component idempotent and that is the scalar unit, 1. In contrast

[1] S. Teitler, Nuovo Cimento Suppl. 3, 1 (1965), referred to as I.
[2] S. Teitler, Nuovo Cimento Suppl. 3, 15 (1965).
[3] M. Riesz in *Comptes Réndus du Dixième Congrès des Mathematiques des Pays Scandanaves* (Copenhagen, 1946), pp. 123 ff.
[4] M. Riesz in *Comptes Réndus du Dousième Congrès des Mathematiques des Pays Scandanaves* (Lund, 1953), pp. 241 ff.
[5] M. Riesz in *Lecture Series No. 38* (University of Maryland, 1958), Chaps. I–IV.
[6] A. Sommerfeld, *Atombau und Spektrallinien* (Braunschweig, 1939), Vol. II, pp. 217–268.
[7] F. Sauter, Z. Physik 63, 803 (1930); *ibid.* 64, 295 (1930).

we may form two-component idempotents by an appropriate linear combination of the scalar unit and any other basis element of the Clifford algebra. Thus, we have $(\frac{1}{2})(1 \pm e^0)$, $(\frac{1}{2})(1 \pm ie^1)$, $\cdots$, $(\frac{1}{2})(1 \pm e^{123})$, $(\frac{1}{2})(1 \pm ie^{0123})$. Clearly the two-component idempotents are not primitive aggregates since, as is readily seen from the multiplication table, every basis element of the Clifford algebra commutes with six other elements in addition to itself and the scalar unit.

We turn next to four-component idempotents of the form

$$E_4 = N^{\frac{1}{2}}(1 + \varepsilon^{A_1} + \varepsilon^{A_2} + \varepsilon^{A_3}). \tag{2.1}$$

Here,

$$\varepsilon^{Ai} = \eta_i e^{Ai}, \qquad \eta_i = \pm 1, \pm i, \tag{2.2}$$

where the e^{Ai} may be any of the Clifford basis elements [the η_i are chosen so that $(\varepsilon^{Ai})^2 = \pm 1$] and $N^{\frac{1}{2}}$ is a normalization scalar constant which is

TABLE I. The products and commutation rules for the basis elements of the vector Clifford algebra generated from the basis vectors of Lorentz space–time. The product of an element at the side (on the left) multiplying an element at the top is given in the box common to their respective row and column. The ± sign in the top right corner of each box indicates whether the product is commutative (+) or anticommutative (−).

	1	e^0	e^1	e^2	e^3	e^{01}	e^{02}	e^{03}	e^{12}	e^{13}	e^{23}	e^{012}	e^{013}	e^{023}	e^{123}	e^{0123}
+	1	+ 1	+ e^0	+ e^1	+ e^2	+ e^3	+ e^{01}	+ e^{02}	+ e^{03}	+ e^{12}	+ e^{13}	+ e^{23}	+ e^{012}	+ e^{013}	+ e^{023}	+ e^{123}	+ e^{0123}
+	e^0	+ 1	− e^{01}	− e^{02}	− e^{03}	− e^1	− e^2	− e^3	+ e^{012}	+ e^{013}	+ e^{023}	+ e^{12}	+ e^{13}	+ e^{23}	− e^{0123}	− e^{123}
+	e^1	− e^{01}	− 1	+ e^{12}	+ e^{13}	− e^0	+ e^{012}	+ e^{013}	− e^2	− e^3	+ e^{123}	+ e^{02}	+ e^{03}	− e^{0123}	+ e^{23}	− e^{023}
+	e^2	− e^{02}	− e^{12}	+ 1	− e^{23}	+ e^{012}	− e^0	+ e^{023}	− e^1	+ e^{123}	− e^3	+ e^{01}	− e^{0123}	+ e^{03}	+ e^{13}	− e^{013}
+	e^3	− e^{03}	− e^{13}	− e^{23}	+ 1	+ e^{013}	+ e^{023}	− e^0	+ e^{123}	− e^1	− e^2	− e^{0123}	+ e^{01}	+ e^{02}	+ e^{12}	− e^{012}
+	e^{01}	− e^1	− e^0	+ e^{012}	+ e^{013}	+ 1	− e^{12}	− e^{13}	− e^{02}	− e^{03}	+ e^{0123}	+ e^3	+ e^3	− e^{123}	− e^{023}	+ e^{23}
+	e^{02}	− e^2	− e^{012}	− e^0	+ e^{023}	− e^{12}	+ 1	− e^{23}	− e^{01}	+ e^{0123}	− e^{03}	− e^1	+ e^{123}	− e^3	+ e^{013}	− e^{13}
+	e^{03}	− e^3	− e^{013}	+ e^{023}	− e^0	− e^{13}	− e^{23}	+ 1	+ e^{0123}	− e^{01}	− e^{02}	− e^{123}	+ e^1	+ e^2	− e^{012}	+ e^{12}
+	e^{12}	+ e^{012}	− e^2	− e^1	+ e^{123}	− e^{02}	− e^{01}	+ e^{0123}	+ 1	− e^{23}	− e^{13}	+ e^0	− e^{023}	− e^{013}	+ e^3	+ e^{03}
+	e^{13}	+ e^{013}	− e^3	+ e^{123}	− e^1	− e^{03}	+ e^{0123}	− e^{01}	− e^{23}	+ 1	− e^{12}	+ e^{023}	− e^0	+ e^{012}	− e^2	+ e^{02}
+	e^{23}	+ e^{023}	+ e^{123}	− e^3	− e^2	+ e^{0123}	− e^{03}	− e^{02}	− e^{13}	+ e^{12}	+ 1	− e^{013}	− e^{012}	+ e^0	+ e^1	+ e^{01}
+	e^{012}	+ e^{12}	+ e^{02}	+ e^{01}	− e^{0123}	+ e^2	+ e^1	− e^{123}	− e^0	+ e^{023}	+ e^{013}	+ 1	− e^{23}	− e^{13}	− e^{03}	− e^3
+	e^{013}	+ e^{13}	+ e^{03}	+ e^{0123}	+ e^{01}	+ e^3	+ e^{123}	+ e^1	+ e^{023}	+ e^0	+ e^{012}	+ e^{23}	+ 1	+ e^{12}	+ e^{02}	+ e^2
+	e^{023}	+ e^{23}	+ e^{0123}	+ e^{03}	+ e^{02}	+ e^{123}	+ e^3	+ e^2	+ e^{013}	+ e^{012}	+ e^0	+ e^{13}	+ e^{12}	+ 1	+ e^{01}	+ e^1
+	e^{123}	+ e^{0123}	+ e^{23}	+ e^{13}	+ e^{12}	+ e^{023}	+ e^{013}	+ e^{012}	+ e^3	+ e^2	+ e^1	+ e^{03}	+ e^{02}	+ e^{01}	+ 1	+ e^0
+	e^{0123}	+ e^{123}	+ e^{023}	+ e^{013}	+ e^{012}	+ e^{23}	+ e^{13}	+ e^{12}	+ e^{03}	+ e^{02}	+ e^{01}	+ e^3	+ e^2	+ e^1	+ e^0	+ 1

real and positive definite. We then have

$$(E_4)^2 = N\{[1 + (\varepsilon^{A_1})^2 + (\varepsilon^{A_2})^2 + (\varepsilon^{A_3})^2] + 2[\varepsilon^{A_1} + \varepsilon^{A_2} + \varepsilon^{A_3}] + [\varepsilon^{A_1}\varepsilon^{A_2} + \varepsilon^{A_2}\varepsilon^{A_1}] + [\varepsilon^{A_1}\varepsilon^{A_3} + \varepsilon^{A_3}\varepsilon^{A_1}] + [\varepsilon^{A_2}\varepsilon^{A_3} + \varepsilon^{A_3}\varepsilon^{A_2}]\}$$
$$\equiv N^{\frac{1}{2}}\{1 + \varepsilon^{A_1} + \varepsilon^{A_2} + \varepsilon^{A_3}\}. \quad (2.3)$$

From the scalar term it follows that

$$N^{-\frac{1}{2}} = 1 + (\varepsilon^{A_1})^2 + (\varepsilon^{A_2})^2 + (\varepsilon^{A_3})^2. \quad (2.4)$$

Hence either the $(\varepsilon^{A_j})^2$, $(j = 1, 2, 3)$ are all equal to $(+1)$ or at most one is equal to (-1). Then $N^{-\frac{1}{2}}$ is either equal to (a)4 or (b)2. Case (a) follows if $\varepsilon^{A_i}\varepsilon^{A_j}(i, j = 1, 2, 3; i \neq j)$ commute and belong to an algebra such that $\varepsilon^{A_i}\varepsilon^{A_j} = \varepsilon^{A_j}\varepsilon^{A_i} = \varepsilon^{A_k}$, $(i \neq j \neq k)$. Clearly then, case (a) corresponds to products of commuting two-component idempotents. The condition of case (b) implies $\varepsilon^{A_i}\varepsilon^{A_j}(i, j = 1, 2, 3; i \neq j)$ anticommute in order to ensure the validity of (2.3). In this case the ε^{A_j} may form a noncommutative algebra, i.e., the E_4 may consist of products of two-component idempotents whose nonscalar units anticommute. However, case (b) may also be satisfied by any three anticommuting unit ε^{A_j} not belonging to a mutually closed algebra (with the scalar unit), but such that only one of the ε^{A_j} has square (-1).

We show now that idempotents of type (a) are primitive whereas idempotents of type (b) are not. An idempotent of type (a) may be written in the form

$$E_{4a} = (\tfrac{1}{4})(1 + \varepsilon^{A_1})(1 + \varepsilon^{A_2}), \quad \varepsilon^{A_3} = \varepsilon^{A_1}\varepsilon^{A_2}. \quad (2.5)$$

We say an element is interior if it is one of $1, \varepsilon^{A_1}, \varepsilon^{A_2}, \varepsilon^{A_3}$, and it is exterior if it is proportional with square ± 1 to a basis element of the Clifford algebra independent of the interior elements. Now, suppose we consider an aggregate $C = \varepsilon^{A_4}$ where ε^{A_4} is any ηe^A. Then, because of the distributivity of multiplication, E_{4a} is primitive if

$$E_{4a}\varepsilon^{A_4}E_{4a} = tE_{4a} \quad (2.6)$$

for all independent ε^{A_4} where t is a scalar. Clearly if ε^{A_4} is proportional to any of the interior elements, Eq. (2.6) is valid with nonvanishing t. If ε^{A_4} is not proportional to an interior element, then from the multiplication table it anticommutes with either ε^{A_1} or ε^{A_2}, or both. Then Eq. (2.6) is valid since t vanishes for all ε^{A_4} which are exterior. This covers all cases so that idempotents of type (a) are primitive.

We consider the idempotents of type (b) in the two subtypes discussed above, i.e., (b1): $1, \varepsilon^{A_1}, \varepsilon^{A_2}, \varepsilon^{A_3}$ form a subalgebra in which $\varepsilon^{A_1}, \varepsilon^{A_2}, \varepsilon^{A_3}$ all anticommute and only one of the $\varepsilon^{A_j}(j = 1, 2, 3)$ has square -1; (b2): again all ε^{A_j} anticommute and only one has square (-1) but they do *not* form an algebra with the scalar unit.

We turn first to subtype (b1). It is readily shown that if ε^{A_4} is exterior it must commute with at least one of the $\varepsilon^{A_j}(j = 1, 2, 3)$, may commute with all three but not only two of them. Thus, it is clear that idempotents of type (b1) are not primitive for all exterior ε^{A_4} and therefore not primitive.

To show subtype (b2) is not primitive, we note the $\varepsilon^{A_1}\varepsilon^{A_2}\varepsilon^{A_3}$ must be exterior and it commutes with all three ε^{A_j}, $(j = 1, 2, 3)$. Hence idempotents of subtype (b2) cannot be primitive.

Now that we have established that idempotents of type (a) (hereafter designated by E_4) are primitive we should like to know how to generate other primitive aggregates from them. Clearly, any aggregate proportional to E_4 is also primitive. In particular, multiplication of E_4 by a scalar factor of magnitude 1 but not equal to 1 generates a primitive aggregate that we designate a pseudo-idempotent.

Now, we note from the multiplication table that any exterior element ε^{A_4} anticommutes with at least one of the nonscalar interior elements. However, it is readily shown that, if it anticommutes with one nonscalar interior element, it must anticommute with two but *not* all three nonscalar elements. Thus there are three distinct classes of exterior elements, i.e.,

ε^{C_1} commutes with ε^{A_1}
and anticommutes with ε^{A_2} and ε^{A_3},

ε^{C_2} commutes with ε^{A_2}
and anticommutes with ε^{A_3} and ε^{A_1},

ε^{C_3} commutes with ε^{A_3}
and anticommutes with ε^{A_1} and ε^{A_2}. (2.7)

The following three lemmas are obtained in a straightforward manner.

Lemma 1: Two exterior elements are in the same exterior class if and only if their product is proportional to an interior element.

Lemma 2: If two elements belong to two different exterior classes, their product belongs to the third exterior class.

Lemma 3: The product of an exterior element and an interior element belongs to the same class as the exterior element.

From Lemma 3 and the multiplication properties of the elements (Table I), it follows that there are at least four elements in each class. Since the exterior classes are mutually exclusive by Lemmas 1 and 2, it follows that the twelve exterior elements are equally divided among the three classes.

We now observe that the product of an element of an exterior class and a primitive idempotent (or pseudo-idempotent) is a primitive aggregate. This follows immediately from Lemmas 1 and 2, because if $C = \varepsilon^{A_4}$, then

$$(\varepsilon^{A_4}E_4)\varepsilon^{A_4}(\varepsilon^{A_4}E_4) = t\varepsilon^{A_4}E_4 = (E_4'\varepsilon^{A_4})\varepsilon^{A_4}(E_4'\varepsilon^{A_4}) = tE_4'\varepsilon^{A_4}, \text{ if } \varepsilon^{A_4} \text{ is in the same exterior class as } \varepsilon^{A_4}, \quad (2.8)$$
$$= 0, \text{ otherwise.}$$

Note that from Eqs. (2.7), E_4' is a primitive idempotent if E_4 is a primitive idempotent. Also, it should be noted from Lemmas 1 and 3 that the oriented product of any ε^{A_i} and a primitive idempotent is proportional to the similarly oriented product of another ε^{A_k} in the same class and the same primitive idempotent.

Thus, the primitive idempotent (or any scalar mutiple thereof) may be used to generate a complete "orthogonal" set of primitive aggregates by forming the products of an element from each exterior class and the primitive idempotent. These primitive aggregates, along with the idempotent itself, are complete in the sense that every ε^{A_i} has a nonvanishing t for at least one of them, and they are orthogonal in that ε^{A_i} has a nonvanishing t for only one of them. There are several possible such sets for each interior class. Thus we recall that there are four idempotents for each interior class, i.e., $(1 \pm \varepsilon^{A_1})(1 \pm \varepsilon^{A_2})$. Suppose we choose one and label it E_4^a. We then label E_4^b, E_4^c, E_4^d such that

$$\begin{aligned} E_4^a\varepsilon^{C_1} &= \varepsilon^{C_1}E_4^b, & E_4^b\varepsilon^{C_1} &= \varepsilon^{C_1}E_4^a, \\ E_4^a\varepsilon^{C_2} &= \varepsilon^{C_2}E_4^c, & E_4^c\varepsilon^{C_2} &= \varepsilon^{C_2}E_4^a, \\ E_4^a\varepsilon^{C_3} &= \varepsilon^{C_3}E_4^d, & E_4^d\varepsilon^{C_3} &= \varepsilon^{C_3}E_4^a, \end{aligned} \quad (2.9)$$

where the relations among the E_4^x are then determined by (2.7). For example, if $E_4^a = \frac{1}{4}(1 + \varepsilon^{A_1} + \varepsilon^{A_2} + \varepsilon^{A_3})$, then

$$\begin{aligned} E_4^b &= \tfrac{1}{4}(1 + \varepsilon^{A_1} - \varepsilon^{A_2} - \varepsilon^{A_3}), \\ E_4^c &= \tfrac{1}{4}(1 - \varepsilon^{A_1} + \varepsilon^{A_2} - \varepsilon^{A_3}), \\ E_4^d &= \tfrac{1}{4}(1 - \varepsilon^{A_1} - \varepsilon^{A_2} + \varepsilon^{A_3}). \end{aligned}$$

Clearly $E_4^xE_4^x = E_4^x$ and $E_4^yE_4^x = 0$, $(y \neq z;\ y, z = a, b, c, d)$. Also it is clear that all E_4^x are primitive and all products of exterior elements with any of these idempotents are primitive obeying Eq. (2.8).

A complete orthogonal primitive set can thus be formed by choosing any idempotent (or any scalar multiple thereof) and products of an element from each exterior class and any idempotent of the interior class.

Among these complete, orthogonal, primitive sets there is a class which may be used as a basis for an ideal. These are clearly only those for which the oriented product of any arbitrary elements of the Clifford algebra and a member of the primitive set written in the form of products of the same orientation yield again the same primitive set up to a scalar factor. This can only occur if the primitive set is made up of an idempotent and products of elements from the three exterior classes and the *same* idempotent.

Consider for specificity the set of left, normalized primitive aggregates

$$l_0^{\varphi_0} = e^{i\varphi_0}E_4, \quad l_1^{\varphi_1} = e^{i\varphi_1}\varepsilon^{C_1}E_4, \quad l_2^{\varphi_2} = e^{i\varphi_2}\varepsilon^{C_2}E_4,$$
$$\text{and} \quad l_3^{\varphi_3} = e^{i\varphi_3}\varepsilon^{C_3}E_4,$$

where normalized means l_0 is an idempotent (pseudo-idempotent) and the φ_μ are real. Then it is clear that

$$(l_0^{\varphi_0})^2 = l_0^{2\varphi_0}, \quad l_i^{\varphi_i}l_0^{\varphi_0} = l_i^{\varphi_0+\varphi_i}, \quad \text{and} \quad l_0^{\varphi_0}l_i^{\varphi_i} = 0 \quad (i = 1, 2, 3).$$

Now consider the linear combination

$$g^l = \alpha_0 l_0 + \alpha_1 l_1 + \alpha_2 l_2 + \alpha_3 l_3, \quad (2.10)$$

where for simplicity we have absorbed the phase factors into the coefficients α_μ. We multiply from the left by a general aggregate C of the entire Clifford algebra. It is clear from Lemmas 1, 2, and 3 that Cg^l is just another linear combination of the same form as g^l. Thus the l_μ's form a basis for a left ideal.

It follows that such an ideal contains no other ideal (other than null ideal). For suppose there was

TABLE II. The fifteen interior classes of the vector Clifford algebra generated from the basis vectors of Lorentz space–time.

A.	$(1,$	$e^0,$	$ie^{12},$	$ie^{012})$
	$(1,$	$e^0,$	$ie^{13},$	$ie^{013})$
	$(1,$	$e^0,$	$ie^{23},$	$ie^{023})$
B.	$(1,$	$ie^1,$	$ie^{23},$	$e^{123})$
	$(1,$	$ie^1,$	$e^{02},$	$ie^{012})$
	$(1,$	$ie^1,$	$e^{03},$	$ie^{013})$
C.	$(1,$	$ie^2,$	e^{03}	$ie^{023})$
	$(1,$	$ie^2,$	$e^{01},$	$ie^{013})$
	$(1,$	$ie^2,$	$ie^{13},$	$e^{123})$
D.	$(1,$	$ie^3,$	$e^{01},$	$ie^{013})$
	$(1,$	$ie^3,$	$e^{02},$	$ie^{023})$
	$(1,$	$ie^3,$	$ie^{12},$	$e^{123})$
E.	$(1,$	$ie^{0123},$	$e^{01},$	$ie^{23})$
	$(1,$	$ie^{0123},$	$e^{02},$	$ie^{13})$
	$(1,$	$ie^{0123},$	$e^{03},$	$ie^{12})$

a nonvanishing ideal whose basis could be made of nonindependent linear combinations of the l_μ so that there would be less than four basis elements. Multiply from the left by a Clifford aggregate proportional to one of the l_μ. This is nonvanishing only for a basis element containing l_0 and is proportional to l_μ. Hence the proposed combination basis elements do not arise in the same combinations when multiplied by an arbitrary aggregate and therefore are not an ideal basis. Thus the ideal basis formed from a class of complete, orthogonal, primitive sets is a minimal ideal basis.

From Table I (the commutation properties of the e^A), we see that any nonscalar element and the scalar element form a commuting algebra with just three pairs of nonscalar elements, each pair of which anticommutes with any element of any other pair. This provides fifteen different commuting algebras or interior classes. We then see a beginning of structure to the minimal ideals. There are fifteen interior classes from each of which we may form four idempotents. Each of these, up to factors of proportionality, determines a minimal-ideal basis. This follows since the exterior classes are determined once an interior class is chosen. The interior classes are listed in Table II. The interior classes have been enumerated in such a way to explicitly indicate that there are three classes containing the scalar, timelike vector, a bivector and a trivector basis elements; nine classes with the scalar, a spacelike vector, a bivector and a trivector basis elements; and finally three classes composed of the scalar, pseudoscalar, and two bivector basis elements. We discuss this grouping in connection with their corresponding representations in the next section.

III. THE MATRIX REPRESENTATION

We may use Riesz' prescription[4] to form the matrix representation of Clifford aggregates and the 4-spinor representation of elements of a minimal ideal. Thus, we write for the product of a general aggregate with a basis element of a left-minimal ideal

$$Cl_\nu = \sum_\mu c_{\mu\nu} l_\mu \tag{3.1}$$

so that

$$C \rightarrow \begin{bmatrix} c_{00} & c_{01} & \cdots \\ c_{10} & c_{11} & \cdots \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdots & c_{33} \end{bmatrix}. \tag{3.2}$$

For a minimal ideal of the form (2.10) we find

$$g^l l_\nu = \sum_\mu (\iota)_{\mu\nu} l_\mu \tag{3.3}$$

or

$$g^l \rightarrow \begin{bmatrix} \alpha_0 & 0 & 0 & 0 \\ \alpha_1 & 0 & 0 & 0 \\ \alpha_2 & 0 & 0 & 0 \\ \alpha_3 & 0 & 0 & 0 \end{bmatrix}, \tag{3.4}$$

where for definiteness we have chosen scalar factors such that

$$(l_0)^2 = l_0, \quad l_i l_0 = l_i, \quad l_0 l_i = 0, \qquad (j = 1, 2, 3). \tag{3.4a}$$

Similarly for a right ideal with basis elements r_μ, we may write

$$r_\mu C = \sum_\nu r_\nu c_{\mu\nu}, \tag{3.5}$$

where again $(c_{\mu\nu})$ has the form (3.2). We may write an element of a right-minimal ideal in the form

$$g^r = \beta_0 r_0 + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3, \tag{3.6}$$

where for definiteness we choose scalar factors such that

$$(r_0)^2 = r_0, \quad r_0 r_i = r_i, \quad r_i r_0 = 0, \qquad (j = 1, 2, 3). \tag{3.7a}$$

Then

$$g^r \rightarrow \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.7}$$

As we have indicated in I, the choice of left-minimal ideal basis which corresponds to the standard representation in matrix notation is given by

$$\begin{aligned} l_0^{\dot{a}a} &= \tfrac{1}{4}(1 - e^0 + ie^{12} - ie^{012}), \\ l_1^{\dot{a}a} &= \tfrac{1}{4}(-e^{13} + e^{013} + ie^{23} - ie^{023}), \\ l_2^{\dot{a}a} &= \tfrac{1}{4}(-ie^3 - ie^{03} + e^{123} + e^{0123}), \\ l_3^{\dot{a}a} &= \tfrac{1}{4}(-ie^1 - ie^{01} - e^2 - e^{02}). \end{aligned} \tag{3.8}$$

The right-minimal ideal basis corresponding to the adjoint in the standard representation is obtained (see, e.g., I) from the product oj e^0 and the conjugate reversion of the $l_\mu^{\dot{a}a}$. This has the form

$$\begin{aligned} r_0^{\dot{a}a} &= -\tfrac{1}{4}(1 - e^0 + ie^{12} - ie^{012}), \\ r_1^{\dot{a}a} &= -\tfrac{1}{4}(e^{13} - e^{013} + ie^{23} - ie^{023}), \\ r_2^{\dot{a}a} &= -\tfrac{1}{4}(ie^3 - ie^{03} - e^{123} + e^{0123}), \\ r_3^{\dot{a}a} &= -\tfrac{1}{4}(ie^1 - ie^{01} - e^2 + e^{02}). \end{aligned} \tag{3.9}$$

The exterior classes in this case may be delineated by $\mathcal{E}^{13}$ (i.e., $-e^{13}$ for left ideal and e^{13} for right ideal), $\mathcal{E}^3$ (i.e., $-ie^3$ for left and $+ie^3$ for right ideal), and $\mathcal{E}^1$ (i.e., $-ie^1$ for left and $+ie^1$ for right ideal). Note that the exterior class to which $\mathcal{E}^{13}$ belongs commutes with $\mathcal{E}^0$ but anticommutes with $\mathcal{E}^{12}$ and $\mathcal{E}^{012}$; $\mathcal{E}^3$ commutes with $\mathcal{E}^{12}$ but anticommutes with $\mathcal{E}^{012}$ and $\mathcal{E}^0$; and $\mathcal{E}^1$ commutes with $\mathcal{E}^{012}$ but anticommutes with $\mathcal{E}^0$ and $\mathcal{E}^{12}$.

Of course, equally valid forms for minimal-ideal bases can be obtained using any of the three other idempotents belonging to the interior class $(1, e^0, ie^{12}, ie^{012})$. Indeed, it is interesting to note that the sixteen elements composing the four distinct minimal-ideal bases belonging to a given interior class are complete in the sense that linear combinations may be used to describe a general Clifford aggregate.

Clearly a similar discussion can be made for any of the fifteen interior classes. Thus, for a given Lorentz frame we have at least fifteen distinct ways (spaces), each of which is fourfold, of describing a 4-spinor. In other words, given a Lorentz vector Clifford algebra, there are fifteen distinct four-component commuting algebras each of which may be used to form four primitive idempotents. Indeed, as we have indicated above, the standard-type representation corresponds to minimal ideals generated from an interior class containing the timelike vector basis element. In I, we showed that a Weyl-type representation corresponds to an interior class containing the pseudo-scalar element. In a similar way, it is readily shown that a Majorana-type representation corresponds to an interior class containing a spacelike vector basis element, a bivector with mixed indices and a spacelike pseudo-vector.

IV. HOMOGENEOUS LORENTZ TRANSFORMATIONS AND CHARGE CONJUGATION

In this section, we are interested in the effect of homogeneous transformations of the space–time basis e^μ which are used as generators of the Clifford algebra. As indicated in I and shown by Riesz,[6,8] the reflection of such a basis vector in a hyperplane normal to a real nonisotropic vector $\mathbf{v}$ may be expressed by

$$[e^\mu]_\mathbf{v} = -\mathbf{v}^{-1}e^\mu\mathbf{v}, \tag{4.1}$$

where nonisotropic means $\mathbf{v}$ has nonvanishing magnitude. As a generalization of Hamilton's well-known decomposition of rotations in ordinary space into two successive reflections, it may be shown that any homogeneous Lorentz transformation can be written in the form[5]

$$\begin{aligned} e^{\mu'} &= (-1)^q\mathbf{v}_q^{-1}\cdots\mathbf{v}_1^{-1}e^\mu\mathbf{v}_1\cdots\mathbf{v}_q \\ &\equiv (-1)^q V_q^{-1}e^\mu V_q, \end{aligned} \tag{4.2}$$

where $q \leq 4$ and the $\mathbf{v}_\mu$ are real nonisotropic vectors. Clearly q-even corresponds to proper transformations, whereas q-odd are improper.

For proper Lorentz transformation of the vector basis, the effect on the general Clifford aggregate may be written

$$C' = V^{-1}CV. \tag{4.3}$$

Here, V is most generally a member of the even subalgebra of the Clifford algebra consisting of a linear combination of scalar, all bivector and pseudoscalar terms. However, it is interesting to note that any proper Lorentz transformation can be expressed in terms of successive timelike (in general three) and spacelike rotations (in general three).[9]

A timelike rotation may be represented by

$$\begin{aligned} e^{0'} &= e^0\cosh\chi - e'\sinh\chi, \\ e^{1'} &= -e^0\sinh\chi + e'\cosh\chi, \\ e^{2'} &= e^2, \\ e^{3'} &= e^3. \end{aligned} \tag{4.4}$$

[8] See also P. K. Raševskiĭ, Transl. Amer. Math. Soc. **6**, 1 (1957), where reflections and rotations in Euclidean space are discussed.

[9] See, e.g., J. L. Synge, *Relativity, The Special Theory* (Interscience Publishers, Inc., New York, 1956), Chap. IV.

The V corresponding to (4.4) is

$$V_t = \cosh \tfrac{1}{2}\chi - e^{01} \sinh \tfrac{1}{2}\chi, \qquad V_t^{-1} = \cosh \tfrac{1}{2}\chi + e^{01} \sinh \tfrac{1}{2}\chi. \tag{4.5}$$

Note that V_t and V_t^{-1} are just reversions of one another and, since they are real, also conjugate reversions.

A spacelike rotation may be represented by

$$\begin{aligned} e^{0'} &= e^0, \\ e^{1'} &= e^1, \\ e^{2'} &= e^2 \cos\theta + e^3 \sin\theta, \\ e^{3'} &= -e^2 \sin\theta + e^3 \cos\theta. \end{aligned} \tag{4.6}$$

The corresponding V is given by

$$V_s = \cos \tfrac{1}{2}\theta - e^{23} \sin \tfrac{1}{2}\theta, \qquad V_s^{-1} = \cos \tfrac{1}{2}\theta + e^{23} \sin \tfrac{1}{2}\theta. \tag{4.7}$$

Again note V_s and V_s^{-1} are conjugate reversions of one another.

The proper Lorentz transformation transforms a minimal ideal basis to the corresponding minimal ideal basis expressed in terms of the new space–time basis. Indeed, since all the basis elements of the Clifford algebra are transformed, even the matrix representation in the transformed basis system is unchanged. This is one way to view the usual matrix discussion (however, see below) which specifies that the γ matrices do not change under Lorentz transformation. From that viewpoint the coefficients of the spinor, i.e., the coefficients multiplying the respective basis elements of the minimal ideal, must be "transformed." That is, they must be re-expressed in terms of point coordinates in the new space–time basis rather than in terms of the point coordinates of the old set in which they remain under a Lorentz basis transformation considered here.

However, the customary discussion of Lorentz transformations in Dirac theory is usually not so straightforward but actually centers on the probability current vector which in the present formalism has the form

$$j^\mu = (\psi^\dagger e^\mu \psi)_s, \tag{4.8}$$

where $(\ )_s$ is the scalar part of the Clifford product (see I), and $\psi^\dagger$ is the adjoint of ψ (i.e., the conjugate reversion of ψ multiplied by e^0).

A proper Lorentz transformation of the j^μ on the contravariant index μ may be written in the form

$$j^{\mu'} = (\psi^\dagger e^{\mu'} \psi)_s = (\psi^\dagger V^{-1} e^\mu V \psi)_s. \tag{4.9}$$

As indicated in I, we may, using the invariance of the scalar part under permutation of the factors (see, e.g., Riesz, Ref. 3), rewrite Eq. (4.9) in the form

$$j^{\mu'} = (V\psi^\dagger V^{-1} e^\mu V \psi V^{-1})_s = (\psi^{\dagger'} e^\mu \psi')_s, \tag{4.10}$$

where

$$\psi' \equiv V\psi V^{-1}, \qquad \psi^{\dagger'} \equiv V\psi^\dagger V^{-1} \tag{4.11}$$

are the transformed minimal ideal elements corresponding to the Lorentz transformation inverse to that expressed in (4.9) on e^μ. The element e^μ is left untransformed in (4.10). However, ψ' is not an element of a left-minimal ideal in the *old* untransformed basis system, but only one in the new primed minimal ideal basis since multiplication from right by V^{-1} destroys the left-ideal property in general. Thus, in a matrix form one would re-express e^μ in terms of the space–time basis corresponding to the Lorentz transformation that acted on the ψ.

Alternatively, one may take (4.9) at face value. Since ψ is a left ideal so is $(V\psi)$, and $\psi^\dagger V^{-1}$ is just the right-ideal adjoint of $(V\psi)$. Hence we may consider $V\psi$ as the transformed minimal-ideal element and define

$$\varphi' \equiv V\psi, \qquad \varphi^{\dagger'} \equiv \psi^\dagger V^{-1}, \tag{4.12}$$

and

$$j^{\mu'} = (\varphi^{\dagger'} e^\mu \varphi')_s. \tag{4.13}$$

This form is more in keeping with the usual approach inasmuch as the matrix representation of the e^μ is left unchanged when the minimal-ideal basis is unchanged. However, in our present context it is not entirely consistent to designate φ' the Lorentz-transformed spinor, such a designation should rather be reserved for the ψ' of Eq. (4.11). Note the lack of change of the matrix representation of e^μ here differs from our discussion above of the lack of change when both e^μ and the minimal-ideal basis were transformed.

We consider now time inversion, space inversion. and charge sign conjugation in the present formalism. Time inversion of space–time is just the change of e^0 to $-e^0$ and is accomplished by reflection in the time vector, i.e.,

$$e^\mu = -v_T^{-1} e^\mu v_T, \tag{4.14}$$

where $v_T = e^0$. Similarly, space inversion is given by an expression of the form (4.14) with $v_S = e^{123}$ corresponding to reflection in all three space vectors. The coefficients of any Clifford aggregate remain

unchanged in such a transformation. This means, for example, if the coefficient were expressed in terms of a general point (x^0, x^1, x^2, x^3) in the old coordinate basis, it would be expressed formally in the same way in the new coordinate basis. However, if the new coefficient were to be expressed in terms of the *old coordinate basis*, say, to give the effect of time inversion, they would be $(-x^0, x^1, x^2, x^3)$. Alternatively, one might seek to nullify the effect of the transformation of space–time basis, then, e.g., under time inversion the set $(-x^0, x^1, x^2, x^3)$ would again be used with the new basis. We note, however, if we do not change the coefficients but transform the space–time basis under space or time inversion, the coordinate arguments of the coefficients correspond to the inverted points since now the coordinates are measured along inverted axes.

Perhaps of more direct interest in the present investigation is the nature of charge sign conjugation since charge sign conjugation is not a Lorentz transformation but reveals the importance of other transformations in the minimal-ideal basis. Consider then the Dirac equation in the vector Clifford formalism

$$\sum_\mu (ie^\mu \partial_\mu - \kappa)\psi = -\sum_\mu qe^\mu A_\mu \psi, \qquad (4.15)$$

where $\kappa = mc/\hbar$, m is the mass of the particle, q is the (algebraic charge)/$\hbar c$, the A_μ are real, and ψ is an element of a minimal ideal. The effect of space and time inversion of the space–time basis in Eq. (4.15) is

$$\sum_\mu (-ie^\mu \partial_\mu - \kappa)\psi_{ST} = \sum_\mu qe^\mu A_\mu \psi_{ST}, \qquad (4.16)$$

where ψ_{ST} is the element of the minimal ideal expressed in terms of the inverted basis. Further, the operation of complex conjugation (i.e., the sign reversal $i \to -i$ in the complex scalar coefficients and in the basis elements) yields

$$\sum_\mu (ie^\mu \partial_\mu - \kappa)\bar{\psi}_{ST} = \sum_\mu qe^\mu A_\mu \bar{\psi}_{ST}, \qquad (4.17)$$

where $\bar{\psi}_{ST}$ is the complex conjugate of the ψ_{ST}. Comparison with Eq. (4.15) shows that Eq. (4.17) is its charge sign conjugate so that $\bar{\psi}_{ST}$ is the charge sign conjugate solution. Thus, in the Clifford formalism, the combined operation of space–time inversion and complex conjugation is equivalent to charge sign conjugation. In going from Eq. (4.16) to Eq. (4.17), it should be noted that the vectors e^μ are unaffected by complex conjugation. A discussion of complex conjugation in this formalism has been given in a paper written later but already published[10] and will also be further discussed in a forthcoming paper.

It is worthwhile to see the effect of these transformations in a specific case. Thus consider the equivalent of the standard representation Eqs. (3.8). For a proper homogeneous Lorentz transformation, we would just obtain the same equations in the primed frame. However, the effect of the improper inversions and complex conjugation is to change the basis to related ones. Thus, time inversion leads to

$$\begin{aligned} l_0^{sc} &= \tfrac{1}{4}(1 + e^0 + ie^{12} + ie^{012}), \\ l_1^{sc} &= \tfrac{1}{4}(-e^{13} - e^{013} + ie^{23} + ie^{023}), \\ l_2^{sc} &= \tfrac{1}{4}(-ie^3 + ie^{03} + e^{123} - e^{0123}), \\ l_3^{sc} &= \tfrac{1}{4}(-ie^1 + ie^{01} - e^2 + e^{02}), \end{aligned} \qquad (4.18)$$

so that for this basis it changes the generating idempotent but leaves the exterior classes unchanged. Space inversion leaves l_0^{sa} unchanged but affects the relative phase (signs) of the exterior class factors so that

$$\begin{aligned} Pl_0^{sa} &= l_0^{sa}, \qquad Pl_1^{sa} = l_1^{sa}, \\ Pl_2^{sa} &= -l_2^{sa}, \qquad Pl_3^{sa} = -l_3^{sa}. \end{aligned} \qquad (4.19)$$

Complex conjugation yields

$$\begin{aligned} l_0^{sb} &= \tfrac{1}{4}(1 - e^0 - ie^{12} + ie^{012}), \\ l_1^{sb} &= \tfrac{1}{4}(-e^{13} - e^{013} - ie^{23} - ie^{023}), \\ l_2^{sb} &= \tfrac{1}{4}(ie^3 - ie^{03} + e^{123} - e^{0123}), \\ l_3^{sb} &= \tfrac{1}{4}(ie^1 - ie^{01} - e^2 + e^{02}), \end{aligned} \qquad (4.20)$$

so that it changes the idempotent as well as the relative signs of the exterior class factors. The effect of time inversion and complex conjugation together on the standard basis is

$$\begin{aligned} l_0^{sd} &= \tfrac{1}{4}(1 + e^0 - ie^{12} - ie^{012}), \\ l_1^{sd} &= \tfrac{1}{4}(-e^{13} - e^{013} - ie^{23} - ie^{023}), \\ l_2^{sd} &= \tfrac{1}{4}(ie^3 - ie^{03} + e^{123} - e^{0123}), \\ l_3^{sd} &= \tfrac{1}{4}(ie^1 - ie^{01} - e^2 - e^{02}), \end{aligned} \qquad (4.21)$$

[10] S. Teitler, J. Math. Phys. **6**, 1976 (1965). It should be noted that the discussion of particle symmetry physics in the third paragraph of Sec. III of this reference is wrong. The regular decomposition symmetry for the Lorentz algebra does not properly provide for hypercharge. Thus, the SU_3 symmetry discussed there is not the usual extension of isotopic symmetry but provides instead a basis for higher-symmetry schemes once the appropriate lower symmetry is obtained. Further, the discussion of the reduction of higher symmetry there is over simplified and at best misleading. Recently, as this author reported in Bull. Am. Phys. Soc. Ser. II, 11, 469 (1966), a more complete marriage of the Lorentz and isotopic algebras was achieved, and a regular decomposition that does provide for a proper extension of isotopic symmetry was obtained.

which also changes the idempotent and the relative signs of the exterior class factors.

Finally, charge sign conjugation on Eqs. (3.18) yields a basis of the form

$$\begin{aligned} l_0^{\text{ccc}} &= \tfrac{1}{4}(1 + e^0 - ie^{12} - ie^{012}), \\ l_1^{\text{ccc}} &= \tfrac{1}{4}(-e^{13} - e^{013} - ie^{23} - ie^{023}), \\ l_2^{\text{ccc}} &= \tfrac{1}{4}(-ie^3 + ie^{03} - e^{123} + e^{0123}), \\ l_3^{\text{ccc}} &= \tfrac{1}{4}(-ie^1 + ie^{01} + e^2 - e^{02}), \end{aligned} \tag{4.22}$$

which is the same as Eqs. (4.21) except for the relative signs (phases) among the exterior class factors.

More pertinent perhaps is the comparison with the original basis l_μ^{ca}. The relative phase among exterior class multipliers remains unchanged, but in the example considered the basis now corresponds to the l_μ^{cd} basis and therefore the spinor is described in an associated but different spinor space. Further, it should be kept in mind that charge sign conjugation as described by Eq. (4.17) involves not only a basis change but also complex conjugation of the coefficients. Also the definition of charge sign conjugation we have used is somewhat different from the usual definition which leaves the underlying spinor space unchanged. This could be accomplished in the present formulation by multiplying from the right by an appropriate element to restore the basis to the original basis. In the example above, this would be achieved with an element belonging to exterior class 3. Of course, this has the effect of changing the labels (as well as, in general, the phases) of the coefficients in the sense that $(\bar{\psi}_0)_{\text{ST}}$ no longer multiplies l_0, etc. Inasmuch as this last step of re-expressing the basis arrived at after the combined operations of time inversion (T), space inversion (P), and complex conjugation (K) in terms of the original basis is representation-dependent, we prefer to use TPK as our definition of charge sign conjugation.

Thus, we have seen how proper, homogeneous Lorentz transformations continuously change a minimal-ideal basis into the same basis expressed in terms of the transformed space–time basis vectors. We have also seen how various combinations of space and time inversion of the space–time basis and complex conjugation transform the minimal-ideal basis to another basis belonging to the same interior class and in general changing the signs (relative phase) among the minimal ideal basis elements. Further, we are able to define a representation-independent operation of charge sign conjugation as the product of time inversion, space inversion, and complex conjugation.

Lorentz Equivalence, Unitary Symmetry, and Spin Unitary Symmetry

S. TEITLER
U. S. Naval Research Laboratory, Washington, D. C.
(Received 22 February 1965; final form received 4 March 1966)

The enumeration and classification of the minimal ideals (spinor spaces) of the vector Clifford algebra generated from the vector basis of Lorentz space–time obtained previously is used in a discussion of equivalence mappings among these minimal ideals. Lorentz equivalence is defined as an equivalence mapping which possesses Lorentz symmetry. The concept of regular decomposition is introduced by which any element of the complete algegra is expressed in terms of a complete set of minimal ideals or spinor spaces. A regular current density which is invariant under unitary symmetry U_4 among the spinor spaces of the regular decomposition is defined. The spin unitary symmetry, within a spinor space under which the spinor current density is invariant, is discussed.

I. INTRODUCTION

In previous works,[1,2] we have utilized Marcel Riesz's[3-5] idea of employing the vector Clifford algebra C_{16}, generated from the basis vectors of Lorentz space–time L_4, in the description of one-particle fields as linear combinations of Clifford basis elements termed Clifford aggregates. The algebra C_{16} has as basis a scalar unit, 1, four vector units of L_4, e^λ, six bivector units $e^{\lambda\mu} \equiv e^\lambda e^\mu$, $(\lambda < \mu)$, four trivector units $e^{\lambda\mu\nu}(\lambda < \mu < \nu)$, and the pseudoscalar unit e^{0123}. Here $\lambda, \mu, \nu = 0, 1, 2, 3$; the ascending order of indices has been chosen for specificity and the algebra is defined by the commutation relations

$$e^\mu e^\nu + e^\nu e^\mu = 2[g^{\mu\nu}], \qquad (\mu, \nu = 0, 1, 2, 3), \tag{1.1}$$

where $[g^{\mu\nu}]$ is the Lorentz metric which we take to be $g^{00} = -g^{11} = -g^{22} = -g^{33} = 1$. As we emphasize below, this algebra C_{16}, which is generated from a Lorentz basis, has particular value in that it provides an adequate description of the most common fields and contains further inherent symmetry properties.

In I,[1] we were able to show the unifying aspect of the vector Clifford algebra description of fields by characterizing the scalar, Proca, and spinor fields as particular Clifford aggregate solutions to the Dirac-like equation

$$\left(\kappa - i \sum_{\mu=0}^{3} e^\mu \partial_\mu\right)\psi = 0, \tag{1.2}$$

where $\kappa = mc/\hbar$, and ψ is a Clifford aggregate. The scalar field solution corresponds to a combination of a scalar part and a 4-vector part consisting of the 4-gradient of the scalar part. The Proca field solution is a particular combination of vector and bivector components. Finally, the usual Dirac-column 4-spinor corresponds to an element of a minimal left ideal.

More recently[6] we have enumerated and classified the minimal ideals of C_{16}. This, in effect, is an enumeration of the spinor spaces possible for the description of 4-spinors. We have shown that the primitive idempotents E_4, which may be used to generate minimal ideals, are composed of elements belonging to the commuting 4-component subalgebras (designated interior classes) of C_{16}, i.e.,

$$E_4 = \tfrac{1}{4}(1 + \eta_1 e^{A_1} + \eta_2 e^{A_2} + \eta_3 e^{A_3}), \tag{1.3a}$$

where

$$(\eta_i e^{A_i})(\eta_j e^{A_j}) = \begin{matrix} 1 & \text{when} \quad i = j \\ \eta_k e^{A_k} & \text{when} \quad i \neq j \neq k \end{matrix}\Bigg\}. \tag{1.3b}$$

Here the appropriate $\eta_i = \pm 1, \pm i$ and the e^{A_i} are nonscalar basis elements of C_{16}. Then it follows that, given an interior class, there are twelve other basis elements of C_{16} which belong to three mutually exclusive exterior classes composed of four elements each. The elements of exterior class 1 commute with e^{A_1}, but anticommute with e^{A_2} and e^{A_3}; the elements of exterior class 2 commute with e^{A_2}, but anticommute with e^{A_3} and e^{A_1} and similarly for exterior class 3. The interior classes and the related exterior classes are given in Table I.

A suitable left-(right-) minimal ideal basis is constituted of one basis element proportional to an

[1] S. Teitler, Nuovo Cimento Suppl. 3, 1 (1965), referred to as I.
[2] S. Teitler, Nuovo Cimento Suppl. 3, 15 (1965).
[3] M. Riesz in *Comptes Rendus du Dixième Congrès des Mathematiques des Pays Scandanaves* (Copenhagen 1946), pp. 123 ff.
[4] M. Riesz in *Comptes Rendus du Douzième Congrès des Mathematiques des Pays Scandanaves* (Lund, 1953), pp. 241 ff.
[5] M. Riesz, in *Lecture Series No. 38* (University of Maryland, 1958), Chaps. I–IV.
[6] S. Teitler, J. Math. Phys. 7, 1730 (1966), referred to as II.

idempotent derived from an interior class and three other elements respectively proportional to the left (right) multiplication of the primitive idempotent by an element from each of the exterior classes. For example, we may choose for a left-minimal ideal basis $l_0 = E_4$, $l_j = \varepsilon^{Cj} l_0 (j = 1, 2, 3)$, where ε^{Cj} is proportional to e^{Cj}. Thus $l_0^2 = l_0$, $l_0 l_j = 0$, $l_j l_0 = l_j$, $l_j l_k = 0$; $j, k = 1, 2, 3$. It follows in a straightforward way (see I or II) that an element of a minimal-left ideal may be represented by a column matrix or column 4-spinor in a spinor space whose basis may be identified with the left-minimal ideal basis.

Note that the minimal ideal basis element which we have defined are both primitive (i.e., $l_\mu C l_\mu = \alpha l_\mu$ or 0, where α is a scalar) and "orthogonal" (i.e., $l_\mu C l_\mu = 0$ unless C contains an element in l_μ). These requirements readily assure a column matrix representation for a spinor but are not really necessary from the point of view of the minimal ideal description of spinor fields. Indeed, one may also use the prescription in I or II to obtain a column matrix representation using a basis containing the idempotent as proportional to one of the basis elements and three independent linear combinations of the other three primitive "orthogonal" basis elements in the new basis. More generally, we may consider four independent linear combinations of all the primitive "orthogonal" basis elements as the new minimal-ideal basis. We speak of the equivalence (of the second kind) of primitive "orthogonal" bases and their independent linear combinations. This leads us to a discussion of spin unitary symmetry.

However, prior to that, we consider the more usual (first kind) equivalence mappings of primitive "orthogonal" bases which correspond to different minimal ideals or spinor spaces. We see that such equivalence mappings may violate Lorentz symmetry, and in Sec. II we discuss the distinction between equivalence and its subclassification Lorentz equivalence.

In Sec. III, we use the fact that the four minimal ideals generated from the idempotents of a given interior class are complete in the sense that any Clifford aggregate may be expressed as some linear combination of them. As is well known, this provides the basis for the usual discussion of the regular representation[7] and its irreducible components which correspond to the minimal ideals. In algebraic language, we speak of the regular decomposition of the C_{16} (which is its only two-sided ideal) into a complete set of equivalent minimal ideals. We use these concepts to discuss possible unitary symmetries among minimal ideals or spinor spaces.

This contrasts with the spin unitary symmetry mentioned above and discussed in Sec. IV. For spin unitary symmetry we are concerned with symmetry within a spinor space. This is distinguished from the unitary symmetry discussed in Sec. III which is among the spinor spaces of the regular decomposition.

II. LORENTZ EQUIVALENCE

Following Boerner[7] we define the equivalence mapping (of the first kind) from a minimal left ideal $\mathcal{I}_l$ onto another one $\mathcal{I}'_l$ as a one-to-one linear mapping which commutes with left multiplication, i.e., if ψ_l maps into ψ'_l, then $A\psi_l$ maps into $A\psi'_l$ or

$$(A\psi_l)' = A\psi'_l. \tag{2.1}$$

Here ψ_l is an element of $\mathcal{I}_l$, ψ'_l an element of $\mathcal{I}'_l$, and A any Clifford aggregate. As Boerner shows, all such equivalence mappings have the form of right multiplication by $l_0 C l'_0$, where l_0 and l'_0 are the respective idempotents of $\mathcal{I}_l$ and $\mathcal{I}'_l$ and C is a general Clifford aggregate. Such an equivalence mapping does not distinguish among minimal ideal bases, and for simplicity one takes a primitive "orthogonal" basis which has the idempotent as one basis element and three others generated from the idempotent.

Clearly, the minimal ideals generated with any of the interior classes are (first kind) equivalent to one another. However, we note the varied composition of these interior classes. Thus three, the standard type, contain a timelike vector, a bivector with space indices, and a trivector with mixed indices (i.e., a spacelike pseudo-vector). Six, the Majorana type, contain a spacelike vector, a bivector with mixed indices, and a mixed trivector; while three others, the pseudo-standard type, contain a spacelike vector, a bivector with space indices, and the trivector corresponding to a timelike pseudo-vector. Finally, there are three, the Weyl type which contain the pseudo-scalar and two complementary bivectors. Thus, for example, a minimal-left ideal generated by using one of the standard-type interior classes is (first kind) equivalent to one generated by using, say, one of the interior classes of the Majorana type. However, such a mapping requires a timelike vector e^0 to be mapped into a spacelike vector $ie^j (j = 1, 2, \text{or } 3)$, etc., hence it has no Lorentz symmetry. Thus, we may distinghish between (first kind) equivalence mappings which have Lorentz symmetry (Lorentz equivalence) and those which do not.

[7] H. Boerner, *Representations of Groups* (North-Holland Publishing Company, Amsterdam, 1963).

TABLE I. The interior and corresponding exterior classes for generating primitive "orthogonal" bases for minimal ideals.

Interior Class $(1, \eta_1 e^{A_1}, \eta_2 e^{A_2}, \eta_3 e^{A_3})$	Exterior Class 1	Exterior Class 2	Exterior Class 3
$(1, e^0, ie^{12}, ie^{012})$	$(e^{13}, e^{23}, e^{013}, e^{023})$	$(e^3, e^{03}, e^{123}, e^{0123})$	$(e^1, e^2, e^{01}, e^{02})$
$(1, e^0, ie^{13}, ie^{013})$	$(e^{12}, e^{23}, e^{012}, e^{023})$	$(e^2, e^{02}, e^{123}, e^{0123})$	$(e^1, e^3, e^{01}, e^{03})$
$(1, e^0, ie^{23}, ie^{023})$	$(e^{12}, e^{13}, e^{012}, e^{013})$	$(e^1, e^{01}, e^{123}, e^{0123})$	$(e^2, e^3, e^{02}, e^{03})$
$(1, ie^1, ie^{23}, e^{123})$	$(e^{02}, e^{03}, e^{012}, e^{013})$	$(e^0, e^{01}, e^{023}, e^{0123})$	$(e^2, e^3, e^{12}, e^{23})$
$(1, ie^1, e^{02}, ie^{012})$	$(e^{03}, e^{23}, e^{013}, e^{123})$	$(e^3, e^{13}, e^{023}, e^{0123})$	$(e^0, e^2, e^{01}, e^{12})$
$(1, ie^1, e^{03}, ie^{013})$	$(e^{02}, e^{23}, e^{012}, e^{123})$	$(e^2, e^{12}, e^{023}, e^{0123})$	$(e^0, e^3, e^{01}, e^{13})$
$(1, ie^2, ie^{13}, e^{123})$	$(e^{01}, e^{03}, e^{012}, e^{023})$	$(e^0, e^{02}, e^{013}, e^{0123})$	$(e^1, e^3, e^{12}, e^{23})$
$(1, ie^2, e^{01}, ie^{012})$	$(e^{03}, e^{13}, e^{023}, e^{123})$	$(e^3, e^{23}, e^{013}, e^{0123})$	$(e^0, e^1, e^{02}, e^{12})$
$(1, ie^2, e^{03}, ie^{023})$	$(e^{01}, e^{13}, e^{012}, e^{123})$	$(e^1, e^{12}, e^{013}, e^{0123})$	$(e^0, e^3, e^{02}, e^{23})$
$(1, ie^3, ie^{12}, e^{123})$	$(e^{01}, e^{02}, e^{013}, e^{023})$	$(e^0, e^{03}, e^{012}, e^{0123})$	$(e^1, e^2, e^{13}, e^{23})$
$(1, ie^3, e^{01}, ie^{013})$	$(e^{02}, e^{12}, e^{023}, e^{123})$	$(e^2, e^{23}, e^{012}, e^{0123})$	$(e^0, e^1, e^{03}, e^{13})$
$(1, ie^3, e^{02}, ie^{023})$	$(e^{01}, e^{12}, e^{013}, e^{123})$	$(e^1, e^{13}, e^{012}, e^{0123})$	$(e^0, e^2, e^{03}, e^{23})$
$(1, ie^{0123}, ie^{23}, e^{01})$	$(e^{02}, e^{03}, e^{12}, e^{13})$	$(e^0, e^1, e^{023}, e^{123})$	$(e^2, e^3, e^{012}, e^{013})$
$(1, ie^{0123}, ie^{13}, e^{02})$	$(e^{01}, e^{03}, e^{12}, e^{23})$	$(e^0, e^2, e^{013}, e^{123})$	$(e^1, e^3, e^{012}, e^{023})$
$(1, ie^{0123}, ie^{12}, e^{03})$	$(e^{01}, e^{02}, e^{13}, e^{23})$	$(e^0, e^3, e^{012}, e^{123})$	$(e^1, e^2, e^{013}, e^{023})$

While each minimal-ideal basis provides a spinor space in which a spinor may be described in a Lorentz-covariant way, the mapping from spinor space to spinor space does not in general possess Lorentz symmetry. Once the mapping is completed, however, the new spinor space is again covariant. This tells us that any interior class may, in general, be used in a covariant theory, but that we must be careful in discussing the relationship of spinor fields in spinor spaces generated from different interior classes.

Another interesting point is the possibility of symmetries compatible with Lorentz covariance among fields described in terms of spinors generated from Lorentz-equivalent interior classes. We briefly return to this point later, but first we turn to a discussion of unitary symmetries.

III. REGULAR DECOMPOSITION AND UNITARY SYMMETRY

As is well known[7] from the discussion of the regular representation in group theory, for an algebra which is its own only two-sided ideal may be decomposed into a linear combination of a complete set of equivalent minimal ideals. Such a complete set consists of the four (either right or left) minimal ideals generated from the four mutually annihilating idempotents of a given interior class. We call such a decomposition of an element of the complete algebra a regular decomposition.

$$C_{16} = \mathcal{g}_l^a + \mathcal{g}_l^b + \mathcal{g}_l^c + \mathcal{g}_l^d. \tag{3.1}$$

Thus, a regular decomposition of an element of C_{16} means expressing it in terms of elements of the minimal ideals of a given interior class or, in other words, expressing it as a linear combination of 4-spinors in the four spinor spaces of a given interior class.

A column 4-spinor or minimal-left ideal may be written in the form

$$\psi_l = \psi_0 l_0 + \psi_1 l_1 + \psi_2 l_2 + \psi_3 l_3. \tag{3.2}$$

From I we recall that, for any ψ_l satisfying the Dirac-like equation, we may define a divergence-free current density

$$j^\mu = (\psi^\dagger e^\mu \psi)_s, \tag{3.3}$$

where the subscript s means scalar part, and $\psi^\dagger$ is the adjoint of ψ (or the conjugate reversion of ψ multiplied on the left by e^0). The possibility of unitary symmetry arises here in a relation among the spinor spaces. Thus, we may consider a regular current density J^μ

$$J^\mu = j_a^\mu + j_b^\mu + j_c^\mu + j_d^\mu \tag{3.4}$$

arising from the four 4-spinors $\psi_l^\alpha (\alpha = a, b, c, d)$. Now, let us make a physical assumption that for a class of spinors the regular current density J^μ is invariant under transformations among spinor spaces, e.g.,

$$\psi_l^{\alpha'} = \sum_{\beta=a,b,c,d} u_\beta^\alpha \psi_l^\beta, \qquad \psi_l^{\alpha\dagger'} = \sum_\beta \overline{u_\beta^\alpha} \psi_l^{\beta\dagger}, \tag{3.5}$$

where $\overline{u_\beta^\alpha}$ is the complex conjugate of u_β^α. Note Eqs.

(3.5) represent combinations of entire spinor spaces, and *not* transformations among the spin components of a given spinor. These latter are the subject of the next section.

Returning to the transformations [Eq. (3.5)], we see that if these transformations are unitary, we do indeed have

$$\sum_\alpha j_\alpha^\mu = \sum_\alpha j_\alpha^{\mu'}. \tag{3.6}$$

The compatibility of this U_4 regular decomposition symmetry varies somewhat with minimal-ideal type. For the standard and pseudo-standard types, one can obtain all four minimal ideal bases from one of them by the operations of 1, TPK, TP, and K, where T is time inversion, P is 3-space inversion, and K is complex conjugation. For the Weyl and Majorana types the above combinations do not suffice to generate the complete set and improper operations such as P, PK, T, and TK are used in completing the set. Hence, the U_4 regular decomposition symmetry may be said to be compatible with homogeneous Lorentz transformations for a standard or pseudo-standard-type minimal ideal basis but only a U_2 regular decomposition symmetry is compatible for Weyl and Majorana types. Further, as emphasized in a recent publication,[8] the U_4 symmetry is also reduced when the spinor coefficients are constrained to obey the Dirac-like equation. Thus, for example, the projection of all the regular decomposition minimal ideal spinors onto one basis encompasses the spinor and its (particle) charge conjugate, mass conjugate, and mass–charge conjugate counterparts so that the mass term is not left invariant.

IV. SPIN UNITARY SYMMETRY

We reduce our view now from the decomposition of the complete algebra and symmetry among the component spinor spaces of its regular decomposition to a particular spinor space. Thus, we consider an element of a given minimal-left ideal as a solution to the Dirac-like equation (1.2) which may be expressed in terms of a primitive "orthogonal" basis as in Eq. (3.2). The physical content of this spinor field is expressed in the divergence-free current density, Eq. (3.3). Now, we observe that the content of Eq. (3.2) and Eq. (3.3) remains unchanged if we consider the minimal ideal basis to be transformed by a unitary transformation

$$\lambda_\mu = \sum_{\nu=0}^{3} \alpha_\mu^\nu l_\nu, \qquad \mu = 0, 1, 2, 3$$

or

$$\lambda = Al, \qquad A^\dagger A = 1. \tag{4.1}$$

We say that the λ basis is (second kind) equivalent to the l basis. However, in general, the (unnecessary) relationship between the vector Clifford minimal left-ideal basis and column matrices is no longer straightforward for the λ basis. More important, the unitary combinations (4.1) are in general not Lorentz equivalent for the spinor field, for they may require, for example, combinations of scalar and vector, etc. However, for standard, pseudo-standard, and Majorana-type minimal ideals there is a nontrivial unitary combination which is compatible in a straightforward way with proper homogeneous Lorentz transformations. This refers to combinations involving the scalar and pseudo-scalar elements. Thus, consider for specificity a spinor in the standard basis

$$\begin{aligned} 4\psi_l = {} & \psi_0(1 - e^0 + ie^{12} - ie^{012}) \\ & + \psi_1(-e^{13} + e^{013} + ie^{23} - ie^{023}) \\ & + \psi_2(-ie^3 - ie^{03} + e^{123} + e^{0123}) \\ & + \psi_3(-ie^1 - ie^{01} - e^2 - e^{02}) \end{aligned}$$

or

$$\psi_l = \psi_0 l_0^S + \psi_1 l_1^S + \psi_2 l_2^S + \psi_3 l_3^S. \tag{4.2}$$

We note that

$$\begin{aligned} l_2^S &= e^{0123} l_0^S, \\ l_3^S &= e^{0123} l_1^S. \end{aligned} \tag{4.3}$$

One might suppose that such an SU_2 symmetry (or a similar one) is involved in higher particle symmetry in the spirit of Gürsey and Radicati, and Pais[9] but this oversimplifies their view. Actually, as Gürsey has indicated[10] one should consider the particle "spin" unitary symmetry in the literature to be a regular decomposition unitary symmetry. For a standard-type minimal ideal with a basis used in Eq. (4.2), this turns out to be a unitary symmetry over two minimal ideal bases obtained from one another by complex conjugation. This may be readily

[8] S. Teitler, J. Math. Phys. **6**, 1976 (1965). However, the discussion of particle symmetry physics in the third paragraph of Sec. III of this reference is incorrect. The regular decomposition symmetry for the Lorentz algebra does not properly provide for hypercharge. Hence, the extension of the isotopic symmetry discussed there does not contain the usual lowest SU_3 symmetry of particle schemes but provides a basis for higher-symmetry schemes once the appropriate lower symmetry is obtained. Recently, as this author reported in Bull. Am. Phys. Soc. Ser. II 11, 469 (1966), a more complete combination of the Lorentz and isotopic algebras was achieved and a regular decomposition symmetry that does provide for a proper extension of isotopic symmetry as well as higher symmetry was obtained.

[9] F. Gürsey and L. A. Radicati, Phys. Rev. Letters **13**, 173 (1964); A. Pais, *ibid.* **13**, 175 (1964).

[10] F. Gürsey, Phys. Letters **14**, 330 (1965). I thank a referee for pointing out this reference.

seen since the operation of complex conjugation on these standard minimal ideal basis elements yields the same results as an ordinary (in contrast to spinor) 180° rotation about the 2 axis. Namely,

$$Kl_\mu = -e^{13}l_\mu e^{13}, \tag{4.4}$$

where

$$l_0 = (\tfrac{1}{4})(1 - e^0)(1 + ie^{12}), \qquad l_1 = -e^{13}l_0,$$
$$l_2 = -ie^3 l_0, \qquad l_3 = -ie^1 l_0.$$

Finally, we may note one other symmetry inherent in the present formulation and alluded to at the end of Sec. II. This is a possible symmetry among the interior classes of a given type. Thus, for a standard-type this corresponds to a rotational symmetry in ordinary space. Summarizing, we may consider a heirarchy of symmetries. Within a spinor space we may have spin unitary symmetry of spin components or basis elements. Next there may be a unitary symmetry among the spinor spaces of the regular decomposition for a given interior class. Further, there may be a symmetry among the interior classes of a particular type.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 7, NUMBER 9 SEPTEMBER 1966

Numerical Solutions of Singular Fredholm Equations

NELLY ULLMAN
Department of Mathematics, Eastern Michigan University, Ypsilanti, Michigan
AND
ROBERT ULLMAN
Scientific Laboratory, Ford Motor Company, Dearborn, Michigan

A singular Fredholm equation of the second kind is solved numerically by a Fourier series analysis in which the singularity is removed naturally, and by a Gaussian quadrature procedure in which the singularity was eliminated by an approximation using the law of the mean. In addition, an analytic solution of an associated Fredholm equation of the first kind was used for comparison with the numerical results. Some properties of these numerical solutions are indicated, and a brief discussion of the errors is given.

INTRODUCTION

RECENTLY, one of the authors (R.U.) solved a number of integral equations of the form

$$\varphi(x) = f(x) - \lambda \int_{-1}^{1} \varphi(t)K(x, t)\, dt \tag{1}$$

by numerical approximations in the course of treating a well-known problem in polymer chemistry.[1,2] Correct analytical solutions had not been found (a solution presented in Ref. 2 was later shown to be in error[3]). The point of special interest was that $K(x, t)$ contained a singularity of the form $|x - t|^{-\alpha}$, $0 < \alpha < 1$, and the technique we had used (an application of the law of the mean) has not, to our knowledge, been discussed in the mathematical literature.[4]

It seemed interesting to us to look into the problem of numerical solutions of singular integral equations by comparing numerical results for a specific integral equation of this form.

EQUATION AND METHODS OF SOLUTION

The integral equation studied was

$$\varphi(x) = x^2 - \lambda \int_{-1}^{1} \varphi(t)\ |x - t|^{-\frac{1}{2}}\, dt. \tag{2}$$

The first method of solution was based on approximation of the integral in (2) by a finite sum. This leads to a set of simultaneous linear equations which can be solved. In doing this, the removable singularity at $t = x$ was evaluated by setting

$$\int_{x-\frac{1}{2}\epsilon}^{x+\frac{1}{2}\epsilon} \varphi(t)\ |x - t|^{-\frac{1}{2}}\, dt$$
$$\approx \varphi(x) \int_{x-\frac{1}{2}\epsilon}^{x+\frac{1}{2}\epsilon} |x - t|^{-\frac{1}{2}}\, dt = 2\sqrt{2}\ \epsilon^{\frac{1}{2}}\varphi(x). \tag{3}$$

The approximation procedure used was a "best" polynomial approximation using a Gaussian quad-

[1] R. Ullman, J. Chem. Phys. **40**, 2193 (1964).
[2] J. G. Kirkwood and J. Riseman, J. Chem. Phys. **16**, 565 (1948).
[3] J. G. Kirkwood and J. Riseman, J. Chem. Phys. **22**, 1626 (1954).
[4] Procedures using expansions in terms of functions which are themselves singular have been used by F. B. Hildebrand, Proc. Am. Acad. Arts Sci. **74**, 287 (1941).

seen since the operation of complex conjugation on these standard minimal ideal basis elements yields the same results as an ordinary (in contrast to spinor) 180° rotation about the 2 axis. Namely,

$$Kl_\mu = -e^{13}l_\mu e^{13}, \tag{4.4}$$

where

$$l_0 = (\tfrac{1}{4})(1 - e^0)(1 + ie^{12}), \qquad l_1 = -e^{13}l_0,$$
$$l_2 = -ie^3 l_0, \qquad l_3 = -ie^1 l_0.$$

Finally, we may note one other symmetry inherent in the present formulation and alluded to at the end of Sec. II. This is a possible symmetry among the interior classes of a given type. Thus, for a standard-type this corresponds to a rotational symmetry in ordinary space. Summarizing, we may consider a heirarchy of symmetries. Within a spinor space we may have spin unitary symmetry of spin components or basis elements. Next there may be a unitary symmetry among the spinor spaces of the regular decomposition for a given interior class. Further, there may be a symmetry among the interior classes of a particular type.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 7, NUMBER 9 SEPTEMBER 1966

Numerical Solutions of Singular Fredholm Equations

NELLY ULLMAN
Department of Mathematics, Eastern Michigan University, Ypsilanti, Michigan
AND
ROBERT ULLMAN
Scientific Laboratory, Ford Motor Company, Dearborn, Michigan

A singular Fredholm equation of the second kind is solved numerically by a Fourier series analysis in which the singularity is removed naturally, and by a Gaussian quadrature procedure in which the singularity was eliminated by an approximation using the law of the mean. In addition, an analytic solution of an associated Fredholm equation of the first kind was used for comparison with the numerical results. Some properties of these numerical solutions are indicated, and a brief discussion of the errors is given.

INTRODUCTION

RECENTLY, one of the authors (R.U.) solved a number of integral equations of the form

$$\varphi(x) = f(x) - \lambda \int_{-1}^{1} \varphi(t)K(x, t)\, dt \tag{1}$$

by numerical approximations in the course of treating a well-known problem in polymer chemistry.[1,2] Correct analytical solutions had not been found (a solution presented in Ref. 2 was later shown to be in error[3]). The point of special interest was that $K(x, t)$ contained a singularity of the form $|x - t|^{-\alpha}$, $0 < \alpha < 1$, and the technique we had used (an application of the law of the mean) has not, to our knowledge, been discussed in the mathematical literature.[4]

It seemed interesting to us to look into the problem of numerical solutions of singular integral equations by comparing numerical results for a specific integral equation of this form.

EQUATION AND METHODS OF SOLUTION

The integral equation studied was

$$\varphi(x) = x^2 - \lambda \int_{-1}^{1} \varphi(t)\ |x - t|^{-\frac{1}{2}}\, dt. \tag{2}$$

The first method of solution was based on approximation of the integral in (2) by a finite sum. This leads to a set of simultaneous linear equations which can be solved. In doing this, the removable singularity at $t = x$ was evaluated by setting

$$\int_{x-\frac{1}{2}\epsilon}^{x+\frac{1}{2}\epsilon} \varphi(t)\ |x - t|^{-\frac{1}{2}}\, dt$$
$$\approx \varphi(x) \int_{x-\frac{1}{2}\epsilon}^{x+\frac{1}{2}\epsilon} |x - t|^{-\frac{1}{2}}\, dt = 2\sqrt{2}\ \epsilon^{\frac{1}{2}}\varphi(x). \tag{3}$$

The approximation procedure used was a "best" polynomial approximation using a Gaussian quad-

[1] R. Ullman, J. Chem. Phys. **40**, 2193 (1964).
[2] J. G. Kirkwood and J. Riseman, J. Chem. Phys. **16**, 565 (1948).
[3] J. G. Kirkwood and J. Riseman, J. Chem. Phys. **22**, 1626 (1954).
[4] Procedures using expansions in terms of functions which are themselves singular have been used by F. B. Hildebrand, Proc. Am. Acad. Arts Sci. **74**, 287 (1941).

rature formula.[5] Accordingly, (2) becomes

$$\sum_{i=1}^{n} \varphi_G(x_j) + \lambda_i \varphi_G(x_i) K(x_i, x_i) W(x_i) = x_i^2, \qquad j = 1, n, \tag{4a}$$

$$K(x_i, x_j) = |x_i - x_j|^{-\frac{1}{2}}, \qquad i \neq j, \tag{4b}$$

$$K(x_j, x_j) = 2\sqrt{2}\,[W(x_j)]^{-\frac{1}{2}}. \tag{4c}$$

The coefficients $W(x_i)$ and the coordinates x_i are both functions of n and were obtained from published tables[6] for $n = 20$, 40, and 80. The simultaneous equations were solved for 16 different values of λ ranging from 0.5 to 200. The subscript G means that the solution is obtained by Gaussian quadratures. Values of $\varphi_G(x_i)$ were computed for all values of x_i used in (4a) for each λ.

The second method was a Fourier series solution which was carried through in a standard manner. By this procedure, the singularity is automatically removed, and an approximation similar to that of (3) is unnecessary. The specific operations are as follows:

$$\varphi(x) = \sum_{p=-\infty}^{\infty} \varphi_p e^{i\pi px}, \tag{5a}$$

$$x^2 = \sum_{p=-\infty}^{\infty} f_p e^{i\pi px}, \tag{5b}$$

$$|x - t|^{-\frac{1}{2}} = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} k_{pq} e^{i\pi qt} e^{i\pi px}. \tag{5c}$$

One obtains a set of n simultaneous equations by keeping only those terms in (5) where p and q are less than or equal to $n - 1$ and greater than or equal to $-n + 1$. Since $f_p = f_{-p}$ and $k_{pq} = k_{-p,-q}$, $\varphi_p = \varphi_{-p}$, the numerical work is simplified.

The equations for the Fourier coefficients φ_p are

$$\varphi_p = f_p - 2\lambda k_{p0}\varphi_0 - 2\lambda \sum_{q=1}^{n-1} (k_{pq} + k_{p,-q})\varphi_q, \qquad p = 0 \text{ to } n - 1, \tag{6}$$

$$f_p = (2/\pi^2 p^2)(-1)^p, \qquad p \neq 0, \tag{7a}$$

$$f_0 = \tfrac{1}{3}, \tag{7b}$$

$$k_{pq} + k_{p,-q} = \frac{\sqrt{2}\,(-1)^{p+q-1}}{\pi(p^2 - q^2)} [p^{\frac{1}{2}} S(2p^{\frac{1}{2}}) - q^{\frac{1}{2}} S(2q^{\frac{1}{2}})], \qquad p \neq q, \tag{7c}$$

$$k_{pp} + k_{p,-p} = \left(\frac{2}{p}\right)^{\frac{1}{2}} \left[C(2p)^{\frac{1}{2}} - \frac{1}{4p\pi} S(2p^{\frac{1}{2}}) \right], \qquad p \neq 0, \tag{7d}$$

$$k_{00} = 2^{5/2}/3. \tag{7e}$$

The functions $C(x)$ and $S(x)$ are the Fresnel integrals defined by

$$S(x) = \int_0^x \sin \frac{\pi t^2}{2}\, dt, \tag{8a}$$

$$C(x) = \int_0^x \cos \frac{\pi t^2}{2}\, dt. \tag{8b}$$

The integrals $S(x)$ and $C(x)$ oscillate rapidly as x becomes large. We were able to obtain values accurate to six decimal places by linear interpolation using a tabulation spaced at values of x 0.001 units apart.[7]

The coefficients φ_p were obtained by simultaneous solution of (6) for 16 different values of λ, the same values used in the Gaussian quadrature procedure. Values of φ_F (φ_F is the function φ obtained from a Fourier series solution) were computed from the equation

$$\varphi_F(x) = \sum_{p=-n+1}^{n-1} \varphi_p e^{i\pi px} \tag{9}$$

for $n = 20$, 40, 80, and each value of λ and every value of x used in the $n = 20$, 40, and 80 Gaussian quadrature procedures.

The function $\varphi(x)$ may be obtained analytically for infinite λ only, by setting the left-hand side of (2) equal to zero. This result based on a previous analysis[8] is

$$\varphi_A(x) = \sqrt{2}\,(4x^2 - 1)/3\pi\lambda(1 - x^2)^{\frac{1}{2}} \tag{10}$$

and is useful for comparison with the results obtained from the approximations by the Fourier series and Gaussian quadrature methods. $\lambda\varphi_A(x)$ was tabulated for the 70 values of x between 0 and 1 used in the $n = 20$, 40, and 80 Gaussian quadrature calculations.

The numerical work was performed by using a FORTRAN program and a Philco 212 computer.

RESULTS

The extensive array of numerical data obtained cannot be conveniently presented here, but we have chosen some typical results which illustrate the

[5] See, for example, Z. Kopal, *Numerical Analysis* (John Wiley & Sons, Inc., New York, 1955), Chap. VII.

[6] P. Davis and P. Rabinoqitz, J. Res. Natl. Bur. Std. (U. S.) **56**, 35 (1956); *ibid.* **60**, 613 (1958).

[7] *Tables of Fresnel Integrals* (Academy of Sciences, USSR, Moscow, 1953).

[8] P. L. Auer and C. S. Gardner, J. Chem. Phys. **23**, 1545 (1955); *ibid.* **23**, 1546 (1955).

advantages and disadvantages of the two methods. The notations $\varphi_F(20)$, $\varphi_F(40)$, $\varphi_F(80)$, $\varphi_G(20)$, $\varphi_G(40)$, $\varphi_G(80)$, and φ_a are used. The subscript F stands for a solution by the Fourier series procedure, the subscript G for a solution by Gaussian quadrature and the subscript a for an asymptotic solution at infinite λ. The indices 20, 40, and 80 stand for solutions obtained by simultaneous solution of 20, 40, and 80 equations, respectively. Data are presented for positive values of x only since $\varphi(x)$ is an even function.

In Fig. 1, plots of $\varphi_G(20)$, $\varphi_G(40)$, $\varphi_G(80)$, and $\varphi_F(80)$ are presented for $\lambda = 200$, $0 < x < 0.37$. The lines are drawn to connect the points of $\varphi_G(80)$ and also $\varphi_F(80)$. The oscillations of the Fourier series result are clearly demonstrated, and, in fact, the agreement between $\varphi_G(20)$, $\varphi_G(40)$, and $\varphi_G(80)$ is good, while the percentage deviation of $\varphi_F(80)$ is high. In Fig. 2, $\varphi_F(20)$, $\varphi_F(40)$, and $\varphi_F(80)$ are presented. The oscillatory character of the results is apparent and decreases in magnitude as n goes from 20 to 80. It is evident from (10) that $\varphi(x)$ diverges at $x = 1$ for infinite λ. Accordingly, it is a critical test of the approximate calculation to examine $\varphi(x)$ for large λ in the neighborhood of $x = 1$. In Fig. 3, graphs of $\varphi_G(20)$, $\varphi_G(40)$, $\varphi_G(80)$, and $\varphi_F(80)$ at $\lambda = 200$ are shown in the range $0.96 < x < 1$. It is instructive to note that $\varphi_G(80)$ increases sharply with x in the neighborhood of unity while $\varphi_F(80)$ does not. The fact that $\varphi_G(80)$ is very nearly correct is evident from Table I. These differences between $\varphi_G(80)$ and the exact value of φ for λ infinite are too small to be detectable on a graph with the scale used in Fig. 3. Table I presents data on $\varphi_G(80)$ for $\lambda = 100$ and 200. It is clear from

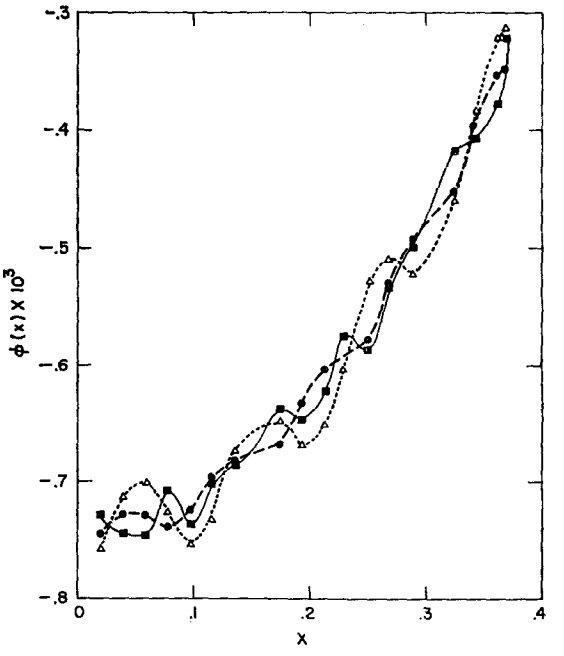


FIG. 2. A plot of $\varphi(x)$ vs x for $0 < x < .37$ and $\lambda = 200$. $\varphi_F(20)$, △; $\varphi_F(40)$, ■; $\varphi_F(80)$, ●.

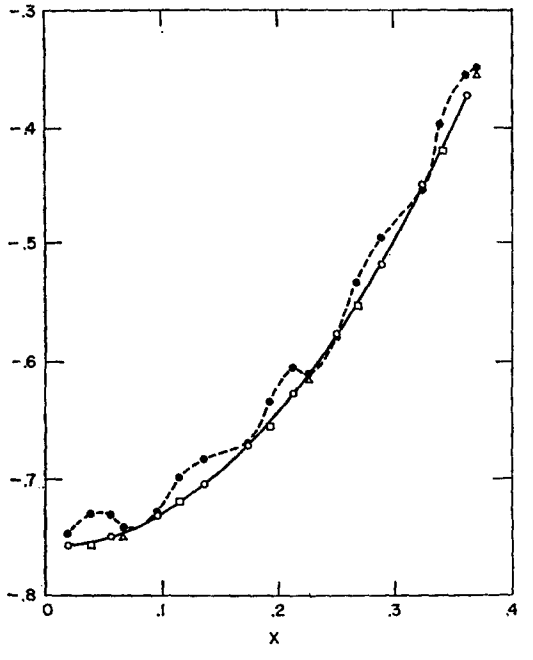


FIG. 1. A plot of $\varphi(x)$ vs x for $0 < x < 0.37$ and $\lambda = 200$. $\varphi_G(20)$, △; $\varphi_G(40)$, □; $\varphi_G(80)$, ○; $\varphi_F(80)$, ●. The solid line connects the points of $\varphi_G(80)$, the dashed line, $\varphi_F(80)$.

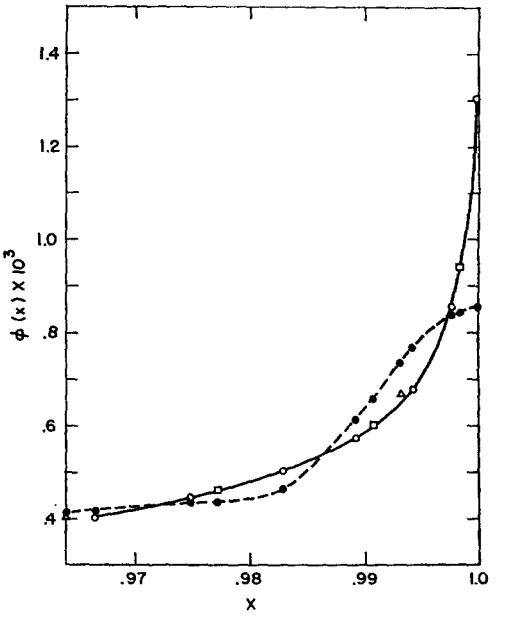


FIG. 3. A plot of $\varphi(x)$ vs x for $0.96 < x < 1.00$ and $\lambda = 200$. $\varphi_G(20)$, △; $\varphi_G(40)$, □; $\varphi_G(80)$, ○; $\varphi_F(80)$, ●. The solid line connects the points of $\varphi_G(80)$, the dashed line, $\varphi_F(80)$.

TABLE I. A comparison of the approximate solution of (1) by Gaussian quadrature for $\lambda = 100$ and 200 with the exact solutions of (1) for λ infinite.

x	$\lambda\varphi_G(80)$, $\lambda = 100$	$\lambda\varphi_G(80)$, $\lambda = 200$	$\lambda\varphi_a$
0.019511	−0.15059	−0.15134	−0.14984
0.136164	−0.14027	−0.14099	−0.13958
0.250952	−0.11462	−0.11529	−0.11409
0.362305	−0.07410	−0.07468	−0.07382
0.468697	−0.01933	−0.01977	−0.01936
0.568671	0.04898	0.04871	0.04857
0.660860	0.13020	0.13165	0.12938
0.744000	0.22413	0.22438	0.22288
0.816954	0.33155	0.33217	0.32991
0.878723	0.45554	0.45668	0.45362
0.913263	0.55121	0.55284	0.54924
0.954591	0.72872	0.73154	0.72714
0.982849	1.00001	1.00574	1.00072
0.989291	1.14206	1.14990	1.14483
0.994228	1.34603	1.35812	1.35321
0.997650	1.69033	1.71358	1.70897
0.999553	2.52160	2.60904	2.60159

TABLE II. A comparison of the Fourier series and Gaussian quadrature solutions of (1) for $\lambda = 0.5$, and $n = 20$, 40, and 80.

x	φ_F $n = 20$	$n = 40$	$n = 80$	φ_G $n = 20$	$n = 40$	$n = 80$
0.019511	−0.081353	−0.080763	−0.081022			−0.081611
0.116084	−0.074417	−0.073676	−0.073681		−0.074555	
0.227786	−0.052577	−0.052066	−0.052520	−0.053243		
0.413779	0.013979	0.014963	0.015277		0.014998	
0.502804	0.061399	0.062126	0.062386			0.062666
0.636054	0.15347	0.15443	0.15394	0.15498		
0.778306	0.28619	0.28468	0.28370		0.28551	
0.912234	0.44826	0.45203	0.45436	0.45718		
0.999554	0.61713	0.63934	0.65425			0.67470

TABLE III. A comparison of the Fourier series and Gaussian quadrature solutions of (1) for $\lambda = 5$ and $n = 20$, 40, and 80.

x	φ_F $n = 20$	$n = 40$	$n = 80$	φ_G $n = 20$	$n = 40$	$n = 80$
0.019511	−0.025170	−0.024621	−0.024921			−0.025238
0.116084	−0.024067	−0.023441	−0.023418		−0.024002	
0.227786	−0.019659	−0.019142	−0.019692	−0.020066		
0.413779	−0.0075476	−0.0068064	−0.0064108		−0.0068272	
0.502804	0.0017947	0.0022485	0.0024800			0.0026375
0.636054	0.021022	0.021985	0.021436	0.021620		
0.778306	0.052164	0.051056	0.049849		0.050836	
0.912234	0.091805	0.094123	0.096774	0.098252		
0.999554	0.17826	0.20045	0.22038			0.25978

TABLE IV. A comparison of the Fourier series and Gaussian quadrature solutions of (1) for $\lambda = 20$ and $n = 20$, 40, and 80.

x	φ_F $n = 20$	$n = 40$	$n = 80$	φ_G $n = 20$	$n = 40$	$n = 80$
0.019511	−0.0072429	−0.0070077	−0.0071484			−0.0072413
0.116084	−0.0069817	−0.0067257	−0.0067065		−0.0069158	
0.227786	−0.0057505	−0.0055208	−0.0057875	−0.0058668		
0.413779	−0.0026111	−0.0023277	−0.0021308		−0.0023099	
0.502804	−0.000090573	0.000061074	0.00014928			0.00024289
0.636054	0.0051870	0.0056115	0.0053615	−0.0053895		
0.778306	0.014036	0.013624	0.013034		0.013422	
0.912234	0.024606	0.025353	0.026582	0.027030		
0.999554	0.054916	0.064667	0.074849			0.101630

TABLE V. A comparison of the Fourier series and Gaussian quadrature solutions of (1) for $\lambda = 200$ and $n = 20$, 40, and 80.

x	φ_F $n = 20$	$n = 40$	$n = 80$	φ_G $n = 20$	$n = 40$	$n = 80$
0.019511	−0.00075804	−0.00072890	−0.00074714			−0.00075671
0.116084	−0.00073315	−0.00070222	−0.00069900		−0.00072363	
0.227786	−0.00060475	−0.00057582	−0.00061105	−0.00061661		
0.413779	−0.00028978	−0.00025680	−0.00023038		−0.00025321	
0.502804	−0.000031732	−0.000015562	−0.0000057949			0.0000079162
0.636054	0.00051253	0.00056579	0.00053403	0.00053548		
0.778306	0.0014387	0.0013917	0.0013132		0.0013618	
0.912234	0.0024989	0.0025763	0.0027342	0.0027810		
0.999554	0.0059301	0.0071496	0.0085312			0.013045

Table I that the extrapolation of $\lambda\varphi_G(80)$ to λ infinite leads to results which are greater in magnitude than the correct results, the errors generally being of the order of 1%.

Tables II through V contain a sample of results in which $\varphi_F(n)$ and $\varphi_G(n)$ are compared for a range of values of x and λ. Two principal trends are shown. First, the agreement between φ_F and φ_G is best at low λ, and systematically worsens as λ increases. Second, the difference between φ_F and φ_G is worse, the closer x is to one. It should also be noticed that $\varphi_F(n)$ may sometimes be in closer accord with φ_G for small n rather than for large n. This is due to the fact that φ_F oscillates about some mean curve as a function of x regardless of the value of n. This trend is also apparent in Figs. 1 and 2.

DISCUSSION OF ERRORS

The Fourier series solution introduces approximations in two ways. First of all, the expansion of $|x - t|^{-\frac{1}{2}}$ and x^2 are terminated after n terms, and the quantities φ_n obtained from the truncated set of equations. Following this, φ_F is constructed for a particular value of x using (5a) which is also truncated, p taking on values in the range $-n + 1 \leq p \leq n - 1$. The result suffers from the fact that the quantities φ_n are not quite correct, and that the expansion of φ is limited. The advantage of the method is that $\varphi_F(x)$ can be easily computed for any value of x.

The coefficients f_p and k_{pq} of (7) do not drop off very rapidly with p and q. If an expansion of x^2 and $|x - t|^{-\frac{1}{2}}$ were made in terms of a more rapidly convergent orthogonal set of functions, the Fourier-type solution would be better. The Legendre polynomials would probably satisfy this condition, but unfortunately the Fourier–Legendre coefficients of $|x - t|^{-\frac{1}{2}}$ are not easily expressed in terms of known functions.

One can write the integral equation in general notation

$$\mathbf{A}\cdot\mathbf{\Phi} = \mathbf{f} \tag{11a}$$

with its representation as a set of simultaneous equations as

$$\bar{\mathbf{A}}\cdot\bar{\mathbf{\Phi}} = \mathbf{f}. \tag{11b}$$

$\bar{\mathbf{A}}$ differs from $\mathbf{A}$ because the integral in (2) is replaced by a sum, and because the diagonal terms of $\bar{\mathbf{A}}$, in particular, contains the approximation of (3). Whether the error committed is serious or not depends on how a small error in a particular element of the matrix propagates in the inversion process. The solution of the simultaneous equation is unstable if a small error perturbs the solution greatly. Some estimate of the instability can be obtained by examining the determinant of the coefficients. If (4a) is divided by λ, it may be written as

$$\mathbf{B}\cdot\mathbf{\Phi} = \mathbf{f}/\lambda, \tag{12a}$$

$$b_{ij} = \delta_{ij}/\lambda + W(x_i)K(x_i, x_j), \tag{12b}$$

b_{ij} is an element of the matrix $\mathbf{B}$. The determinant of $\mathbf{B}$ for the Gaussian quadrature procedure using 80 simultaneous equations is given for 16 values of λ in Table VI. The large values of $|\mathbf{B}|$ for low λ

TABLE VI. The determination of $|\mathbf{B}|$ for the Gaussian quadrature procedure for 16 values of λ.

λ	$\lvert\mathbf{B}\rvert$
0.5	2.69203×10^{30}
1.	2.57299×10^{11}
2.	2.51109×10^{-5}
3.	5.21611×10^{-13}
5.	8.07563×10^{-21}
7.5	1.34191×10^{-25}
10	2.64995×10^{-28}
20	6.78439×10^{-33}
30	1.31227×10^{-34}
40	1.65478×10^{-35}
50	4.61425×10^{-36}
60	1.93940×10^{-36}
80	6.44523×10^{-37}
100	3.29460×10^{-37}
150	1.33013×10^{-37}
200	8.40579×10^{-38}

are due to the first term on the right of (12b). However, as the value of λ increases, this term is negligible and yet the value of the determinant steadily decreases. This is the evidence (not proof) that the errors in φ_G would be expected to be greater for large λ.

It is probable that the device for removal of the singularity (3) in the Gaussian procedure is the weakest step in the process, since $K(x_i, x_j)$ is in error by terms of the order of $W(x_j)^{\frac{1}{2}}$. Let us assume that the only errors in the analysis arise from this approximation. The solution of Eq. (12a) is treated as the exact result, but the approximate solution differs because of errors on the main diagonal $\mathbf{B}$. One has

$$\bar{\mathbf{B}} \cdot \bar{\mathbf{\Phi}} = \mathbf{f}/\lambda, \tag{13a}$$

where $\bar{\mathbf{B}}$ is related to $\mathbf{B}$ by

$$\bar{b}_{ij} = b_{ij} + \epsilon_i \, \delta_{ij}. \tag{13b}$$

The solution of Eq. (12a) is obtained by matrix multiplication with the inverse matrix $\mathbf{B}^{-1}$. Multiplying Eqs. (12a) and (13a) by $\mathbf{B}^{-1}$ leads to

$$\varphi_i = \frac{1}{\lambda} \sum_j b_{ij}^{-1} f_j, \tag{14a}$$

$$\bar{\varphi}_i = \varphi_i - \frac{1}{\lambda} \sum_j b_{ij}^{-1} \epsilon_j \bar{\varphi}_j, \tag{14b}$$

which, upon substitution of φ_j for $\bar{\varphi}_j$, yields

$$\bar{\varphi}_i = \varphi_i - \frac{1}{\lambda} \sum_j b_{ij}^{-1} \epsilon_j \varphi_j. \tag{14c}$$

Note that the errors in (14c) increase as the terms b_{ij}^{-1} become large; generally, this is true for small $|\mathbf{B}|$.

An improved approximation over that of (3) would be obtained by representing $\varphi(t)$ as a Taylor series about $t = x$ keeping higher derivatives of $\varphi(x)$ in the expansion. These derivatives could be expressed as difference formulas in terms of $\varphi(x)$ at neighboring points. This was not done by us, and such a procedure would be interesting to examine.

The Gaussian quadrature method yields results only at special values of x, and in order to obtain values of φ at an arbitrary value of x, appropriate interpolation methods must be applied. This is not difficult except perhaps in the neighborhood of a singularity of the function φ.

ACKNOWLEDGMENT

The authors wish to thank M. S. Klamkin of the Ford Scientific Laboratory for helpful discussions.