Quantum axiomatics, representation theorem, and communicability of observations in physics

Adolf Miklavc

Physik Department der Technischen Universität München, Theoretische Physik, 8046 Garching bei München, West Germany[a)]

(Received 13 January 1976)

We show how a fundamental assumption in the Dirac formulation of quantum mechanics, namely that the states of a physical system at a particular time are mathematically represented by unit vectors in Hilbert space, can be deduced from certain aspects of our experimental procedures and of the observed outcome of quantum mechanical experiments. Our assumptions have clear empirical meaning and the results hold true for any dimensionality of the system, without anomalies in low dimensions which exist in the two well-known axiomatic approaches to quantum mechanics. The propositional logic approach of Birkhoff and von Neumann does not work for quantum systems of dimension less than four and requires an assumption which does not have an empirical basis. Jordan algebra axioms, on the other hand, also lead to anomalies in low dimensions and, moreover, are formal and cannot be directly physically interpreted. In our work it was possible to avoid these shortcomings.

INTRODUCTION

Dirac in his classic work[1] developed quantum theory starting from the basic assumption that the states of a physical system at a particular time can be mathematically represented by vectors in Hilbert space. The assumption holds true even in relativistic quantum mechanics and appears to be very deep and far-reaching. It is, however, rather formal since it is not easy to see how it can be related to the known empirical facts.

Birkhoff and von Neumann,[2] with their work on the "propositional logic," were the first who attempted to build quantum mechanics from physically motivated rather than formal axioms. Plausible arguments led them to the conclusion that the set of experimentally verifiable propositions of a physical system form a complete, orthocomplemented lattice. Their proposition system is an unique direct union of irreducible proposition systems and each of them can be imbedded into a projective geometry. The representation theorem[3] then tells us that if the projective geometry G has dimension $n \geq 3$ there exists a linear vector space L over some field and a one-to-one correspondence between the elements of G and the linear manifolds of L. The property of orthocomplementation defines a definite Hermitian form in the vector space L and restricts the field to a real, complex, and quaternionic field. The vector space L thus becomes a metric space with positive definite metric.

Much work has subsequently been done[4] to improve and complete the approach of Birkhoff and von Neumann, and especially to derive their axioms from physically more defensible assumptions. Two serious difficulties associated with their approach, however, remain. First, they must assume that the greatest lower bound exists for any, even an infinite, family of propositions. There is no empirical basis to support this assumption. The second difficulty is that the strong representation theorem mentioned above exists only for projective geometries of dimension $n > 2$ (in conventional quantum mechanics they would correspond to systems with a basis of dimension $\nu = n+1 > 3$). In the case $n = 2$ the theorem becomes so weak that it is of little value[8,9] *and when* $n=1$, *there is no representation theorem at all*. The axioms of Birkhoff and von Neumann therefore can say nothing about the one- and two-dimensional systems which, in fact, occur in nature (e.g., spin systems with $j = \frac{1}{2}$ or 1), even if they are derived from the empirically well-based assumptions.[10]

The Jordan algebra axioms,[11] on the other hand, which give a strong representation theorem for all finite dimensions, also lead to anomalies in low dimensions and, moreover, are formal and cannot be directly physically interpreted.[10]

Our goal will be to obtain the representation theorem for all dimensions from physically well-based assumptions. We shall do this in two steps. First we introduce the states of a physical system as an abstract set over which one can define a topology on the basis of certain empirical observations. As we shall see, the requirement that the observations in physics be communicable will have a profound effect on the topology, making it satisfy the second axiom of countability. One can then show that, with this topology, the set of quantum states is homeomorphic to a subset of the unit sphere in the complex Hilbert space. It will become clear that some of the vectors representing the states must be superpositions of others.

QUANTUM MECHANICAL STATES AND THEIR TOPOLOGY

There is a profound difference in our observations of classical and quantum physical systems. A classical system, after we carried through the measurements, is essentially still the same as it was before. Measurements perturb a classical system so little that we can normally neglect their effect on it. It is not so with quantum systems. Here a measurement usually perturbs the system strongly and in an uncontrolled way. The outcome of the measurement is thus unpredictable and

[a)] This work was initiated and in part performed when the author was at the Department of Physics and Astronomy, University of Maryland.

so is the change which the system itself undergoes in this process. A measurement like in classical physics can therefore give us hardly any information at all in the case of quantum systems. One therefore creates a large number of identical systems, performs the same measurement on each of them and then looks at the probability for a certain outcome of the measurement to occur. But how do we know that the systems are indeed "identical"? We have no way of controlling the large number of the microscopic constituents of the apparatus which we use to prepare the system. Also there could be unknown "hidden variables" which we would need to specify. We call the systems "identical" simply because we prepare them using (macroscopically) the same procedure.

Thus, to be able to say anything at all about a quantum mechanical system and also to be able to communicate the observations to others, *we must be able to specify uniquely the system which we investigate*. We could not create "identical" copies of it otherwise. Suppose that we have assembled all the necessary equipment needed to generate the physical system which we want to study. Different states of the system must then correspond to different settings of the meters with which we control our instruments. If the states of the physical system are introduced as an abstract set S, we can say that to each element of this set there corresponds a unique set of numbers which we read on our meters, that is, a point in R^∞, to be completely general. Later on we shall discuss this correspondence a little more in detail. Let us emphasize again that such a correspondence must exist since otherwise we would have no way of reproducing the state and therefore we could not perform quantum mechanical measurements on it and we could not communicate our observations to others.

Suppose now that we prepare our system in a state a by passing it through the appropriate apparatus, often called "filter." If the system so prepared is then made to pass through the filter corresponding to some other state b, we can either get our system in the state b or we can get nothing. It cannot be predicted which of these two possibilities will actually occur. Sometimes one occurs, sometimes the other. However, if the experiment is repeated many times, one finds that there exists the limit

$$\lim_{n\to\infty}[n(a,b)/n] \equiv p(a,b),$$

where $n(a,b)$ is the number of those experiments which have for the outcome the system in the state b and n is the total number of the experiments performed. The limit is, of course, a function of both states a and b.

There exists therefore, a function p, usually called transition probability, which maps the elements $(a,b)\in S\times S$ onto the points $p(a,b)\in[0,1]$. Transition probability has the well-known properties:

$$\begin{aligned}&(a)\quad p(a,b)=1 \Longleftrightarrow a=b,\\ &(b)\quad p(a,b)=p(b,a).\end{aligned}\tag{1}$$

Note that $p(x,y)=1$ means that x and y are indistinguishable. One can further assume that if $1-p(x,y)$ is "small" then x and y are "close," i.e., similar to each other in physical properties, in particular that $|p(z,x)-p(z,y)|$ is "small" for all z. Thus, one can assume that, in addition to the properties (a) and (b) given above, transition probability also has the following property:

(c) Given an $\epsilon>0$, there exists a $\delta>0$ such that, for all $z\in S$,

$$|p(z,x)-p(z,y)|<\epsilon \text{ if } 1-p(x,y)<\delta. \tag{2}$$

For any $a\in S$ we can now define the family $U(a)$ of sets $U_\nu(a)$

$$U_\nu(a)=\{x\,|\,x\in S,\ 1-p(a,x)<1/\nu\},\quad \nu=2,3,4,\cdots. \tag{3}$$

It is not difficult to show that the sets $U(a)$ satisfy the three axioms of Hausdorff[12]:

(H1) Every element $x\in S$ possesses at least one set in $U(x)$ and is contained in all sets of $U(x)$.

(H2) The intersection of two sets of $U(x)$ contains a set of $U(x)$.

(H3) If y is in the set $V\in U(x)$ then there exists a set $W\in U(y)$ such that $W\subset V$.

The first two axioms are obviously satisfied. To show that the sets $U(x)$ satisfy also the third axiom, we choose a set $U_\nu(a)\subset U(a)$ and a state $b\neq a$ such that $b\in U_\nu(a)$. We then have

$$1-p(a,b)=1/\nu-\epsilon,\quad \text{where } 0<\epsilon<1/\nu.$$

Because of the property (2) of the transition probability, there exists an integer ρ such that the inequality

$$|p(a,b)-p(a,x)|<\epsilon/2$$

holds true for all x which satisfy

$$1-p(b,x)<1/\rho.$$

This means that for all $x\in U_\rho(b)$ we have

$$1-p(a,x)\leq 1-p(a,b)+\epsilon/2=1/\nu-\epsilon/2.$$

Therefore, for any $b\in U_\nu(a)$ there exists a $U_\rho(b)$ such that

$$U_\rho(b)\subset U_\nu(a),$$

which was to be shown.

Since the sets $U(x)$ satisfy the axioms of Hausdorff, they form a fundamental system of neighborhoods (i.e., a neighborhood basis) at each point $x\in S$ and therefore define uniquely a topology in S.[12,13] We shall now show that this topology, call it τ, has a countable basis, so that the set S of physical states satisfies the second axiom of countability.[14]

We pointed out already earlier in this section that to each state of a physical system there corresponds a unique sequence of numbers, i.e., a unique point in R^∞. We came to this conclusion by imagining that we have all the instruments assembled which are needed to generate the system we want to study so that different states of this system then correspond to different readings on the meters. Even if infinite number of instruments were needed, each of them having an arbitrarily large (could be infinite) number of meters, the set of all the readings would still be countable and thus could be represented by a point in R^∞. (This is because the union of a countable family of countable sets is countable.[15]) We also pointed out that such a correspondence must exist since otherwise we would have no way of reproducing a state, and therefore we could not perform quantum mechanical measurements. It is clear that a state can have physical meaning only if it can be characterized by a finite number of data or if it can be approximated by a state of such property to any desired degree of accuracy. It is also clear that we should be able to approximate any state, to any desired degree of accuracy, by a state characterized by rational numbers only. Let R^+ be the set of all points in R^∞ which have different from zero only finite number of components, all of them being rational numbers. The set R^+ is countable. The physical states which correspond to points in R^+ form then a countable subset $S^+ \subset S$. It follows, from what we just said about physical states, that for any $a \in S$ and an arbitrarily small $\epsilon > 0$ there must exist a $b \in S^+$ such that $1 - p(a,b) < \epsilon$.

The family of sets

$$B = \{U_\nu(x) \mid x \in S^+,\ \nu = 2, 3, 4, \cdots\} \tag{4}$$

with $U_\nu(x)$ as defined in (3) form a basis of topology τ in S. To show this, we choose an arbitrary set $G \subset S$ which is open with respect to the topology τ. For each point $a \in G$ there is a member of its neighborhood basis $U_\nu(a)$ such that $U_\nu(a) \subset G$. Suppose that $a \notin S^+$. We can find a basic set $U_{4\nu}(b) \subset B$, where b is chosen so that

$$1 - p(a,b) < 1/4\nu,$$

$$|p(a,x) - p(b,x)| < 1/4\nu \quad \text{for all } x \in U_{4\nu}(b).$$

Such a choice is possible because of the properties of the set S^+ introduced above, and because of the property (2) of transition probability. It follows then that

$$a \in U_{4\nu}(b) \quad \text{and} \quad 1 - p(a,x) < 1/2\nu \quad \text{for all } x \in U_{4\nu}(b)$$

so that $U_{4\nu}(b) \subset U_\nu(a)$. We have therefore the relation

$$a \in U_{4\nu}(b) \subset G. \tag{5}$$

If $a \in S^+$, a similar relation obviously exists. Thus, since G and $a \in G$ are arbitrary, the relation (5) implies that the family of sets B, defined in (4), form a basis of topology τ. It is important to note that this basis is countable since it is a union of a countable family of countable sets.[15]

The space of quantum states S is regular in the topology τ. To see this, we note first that the closure $\bar{U}_\nu(a)$ of a set $U_\nu(a)$ defined in (3) is contained in

$$\bar{U}_\nu(a) \subset \{x \mid 1 - p(a,x) \leq 1/\nu\}.$$

Indeed, for any point y such that

$$1 - p(a,y) = 1/\nu + \epsilon, \quad \epsilon > 0$$

we can find a $\nu' = \nu'(\epsilon)$ so that if $x \in U_{\nu'}(y)$, then

$$|p(a,x) - p(a,y)| < \epsilon/2.$$

This means that $1 - p(a,x) > \nu + \epsilon/2$ for all $x \in U_{\nu'}(y)$, that is,

$$U_{\nu'}(y) \cap U_\nu(a) = \emptyset.$$

The point y is therefore an exterior point of $U_\nu(a)$, no matter how small ϵ is. Clearly then, every neighborhood of a point x contains a closed neighborhood of x, that is, the space S is regular.

We see that the well-known empirical facts about the quantum states, together with the requirement that our observations be communicable, enable us to define a topology τ over the set of states S which has a countable base and which makes S into a regular space.

QUANTUM MECHANICAL STATES AND HILBERT SPACE

We shall now show that the space of quantum states S, with the topology τ constructed in the preceding section, is indeed, homeomorphic to a set in Hilbert space. The problem of metrization of topological spaces that is, the problem of finding necessary and sufficient conditions for a topological space to be metrizable (i.e., homeomorphic to a certain metric space), has been regarded as one of the basic and most important problems in mathematics. The most general solution to this problem was given by Smirnov in 1951.[16,17] We shall base our work on his solution which, for our topology, can be simplified somewhat.

In the preceding section we showed that topology τ in the space S has a countable base B

$$B = \{B_n\}, \quad n = 1, 2, 3, \cdots,$$

and that the space is regular. It is not difficult to show that every open set G in the space S is of type F_σ, i.e., it can be expressed as union of a sequence of closed sets. In fact, since S is regular, for every $x \in G$, there exists a neighborhood $B_{n(x)}$ belonging to the basis B, the closure of which is contained in G

$$\bar{B}_{n(x)} \subset G.$$

It is therefore possible to find a subsequence $B' = \{B_{n'}\} \subseteq B$ such that

$$G = \bigcup_{n'} \bar{B}_{n'}, \quad B_{n'} \in B' \quad \text{for all } n'.$$

G is therefore of type F_σ.

The space S is normal. To prove this, we select two disjoint closed sets C and D of the space S. For every point $x \in C$ we select a neighborhood $B_{n(x)}$ belonging to the basis B, whose closure is disjoint from the set D and, analoguously, for every point $y \in D$ we select a neighborhood $B_{n(y)} \in B$ whose closure is disjoint from the set C. Such neighborhoods exist in view of the regularity of the space S. Furthermore, for every $x \in C$ and every $y \in D$ we can define the sets

$$V_{n(x)} = B_{n(x)} \Big\backslash \bigcup_{\substack{y \in D \\ k(y) \leq n(x)}} \bar{B}_{k(y)},$$

$$\tilde{V}_{n(y)} = B_{n(y)} \Big\backslash \bigcup_{\substack{x \in C \\ k(x) \leq n(y)}} \bar{B}_{k(x)}.$$

Clearly then the sets

$$V = \bigcup_{x \in C} V_{n(x)} \quad \text{and} \quad \tilde{V} = \bigcup_{y \in D} \tilde{V}_{n(y)}$$

are disjoint neighborhoods of the sets C and D. The normality of the space S is thus proved.

We can now construct a topological mapping of the space of states into Hilbert space H. Since S is normal and since every open set of the space S is of the type F_δ, there exist[16] continuous functions $q_n(x)$ satisfying the conditions $0 \leqslant q_n(x) \leqslant 1$ for all $x \in S$ and vanishing at all points $x \in S \setminus B_n$, and only at these points. With the help of the functions $q_n(x)$ one can then define continuous, complex-valued functions $\tilde{\xi}_n(x)$, for example, by

$$\tilde{\xi}_n(x) \equiv 2^{-n}(1 - a^{q_n(x)})/(1 + |a|), \quad \operatorname{Im} a > 0,$$

so that $F(x) \equiv \sum_n |\tilde{\xi}_n(x)|^2 \leqslant 2$. The functions $\xi_n(x) \equiv \tilde{\xi}_n(x)/F^{1/2}(x)$ then satisfy the following conditions:

(a) $0 \leqslant |\xi_n(x)| \leqslant 1$ for all $x \in S$ and vanishing at all points $x \in S \setminus B_n$; and only at these points.

(b) $\sum_n |\xi_n(x)|^2 = 1$.

This means that the family $\{\xi_n(x)\}$, where x is an arbitrary point of the space S, is an element of the Hilbert space H, which we denote by $f(x)$. In this way, we obtain a mapping f of the space S onto a certain subset $f(S)$ of the Hilbert space H. We shall now prove that this mapping is one-to-one. In fact, if x and y are two distinct points of the space S, there exists a certain B_n containing the point x and not containing the point y. Then $\xi_n(x) \neq 0$, but $\xi_n(y) = 0$, from which it follows that $f(x) \neq f(y)$. We show next that the mapping f is continuous. Let $x \in S$ and $\epsilon > 0$ be chosen arbitrarily, and let N be a natural number such that $2^{-N} < \epsilon^2/8$. We can select a neighborhood O_x of the point x so small that for every $y \in O_x$ the inequality

$$|\xi_{n_i}(x) - \xi_{n_i}(y)| < \epsilon/\sqrt{2n}$$

is satisfied for all $n_i \leqslant N$. From this we have the inequality

$$\sum_{n \leqslant N} |\xi_n(x) - \xi_n(y)|^2 < \epsilon^2/2.$$

In accordance with the choice of the number N, we have, furthermore,

$$\sum_{n > N} |\xi_n(x) - \xi_n(y)|^2 \leqslant 4/2^n < \epsilon^2/2.$$

The two inequalities together give us

$$\rho[f(x), f(y)] = \left(\sum_n |\xi_n(x) - \xi_n(y)|^2\right)^{1/2} < \epsilon,$$

by which the continuity of f is proved.

We shall finally prove that the one-to-one and continuous mapping f is continuous also in the other direction. Let the point $\xi \in f(S)$ and the neighborhood O_x of the point $x = f^{-1}(\xi)$ be chosen arbitrarily. We select $B_n \in B$ such that $x \in B_n \subset O_x$ and set $\epsilon = |\xi_n(x)| \neq 0$. Then, for every point $\eta \in f(S)$ such that $\rho(\xi, \eta) < \epsilon$, we shall have $|\xi_n(y)| \neq 0$, where $y = f^{-1}(\eta)$. This implies that $y \in B_n \subset O_x$. In other words,

$$f^{-1}[O(\xi, \epsilon)] \subset O_x,$$

by which all the necessary proofs are completed.

We have thus shown that the states of a quantum system are mathematically vectors in Hilbert space, more precisely, in l_2. Since the basis in this space is either finite or at most countably infinite, not all of these state vectors can be linearly independent, that is, some of them are linear combinations of others. (No two state vectors are proportional to each other, though.) We have thus arrived at the celebrated "superposition principle" for quantum states. The present derivation is based solely on a few well-known facts about our experimental procedures and about the observed outcome of experiments performed on quantum systems. The requirement that our observations be communicable is found to have a profound effect on the topology of quantum states, making it satisfy the second axiom of countability. In our analysis we are not limited in any way by the dimensionality of quantum systems. We mentioned already that the propositional logic approach of Birkhoff and von Neumann does not work for systems of dimension less than four. None of the two approaches, however, seems to enable us to determine the Hilbert space completely. Propositional logic allows Hilbert spaces over real complex, or quaternionic fields. In our case, as can be seen easily, a homeomorphism of the space of quantum states into a real Hilbert space is also possible. We do not know as yet what are all the possible Hilbert spaces allowed by the assumptions upon which our approach is based. Thus, the empirical basis for the complex numbers in conventional quantum mechanics remains still unknown.

ACKNOWLEDGMENTS

I wish to express my gratitude to Professor D. I. Fivel whose colloquium, given at the University of Maryland, stimulated my interest in this subject and who also otherwise contributed much to my knowledge of it. I also wish to thank Professor H. Busemann and Professor H. C. Wang for the correspondence and Professor C. H. Woo for discussions and his many-sided support.

[1] P.A.M. Dirac. *Principles of Quantum Mechanics* (Clarendon, Oxford, 1958), 4th ed.

[2] G. Birkhoff and J. von Neumann, Ann. Math. 37, 823 (1936).

[3] R. Baer, *Linear Algebra and Projective Geometry* (Academic, New York, 1952).

[4] The reader may refer to Refs. 5—7, which list some of the relevant work in this field (e.g., of G. Emch, S. Gudder, J.M. Jauch, G. Ludwig, M.D. Mac Laren, C. Piron, and J.C.T. Pool).

[5] G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley-Interscience, New York, 1972).

[6] J.M. Jauch, *Foundations of Quantum Mechanics* (Addison-Wesley, Reading, Mass., 1968).

[7] V.S. Varadarajan, *Geometry of Quantum Theory* (Van Nostrand, Princeton, N.J., 1968).

[8] R.J. Bumcrot, *Modern Projective Geometry* (Holt, Rinehart, Winston, New York, 1969), Chap. 3.

[9] D. Pedoe, *Introduction to Projective Geometry* (MacMillan, New York, 1969), Chaps. 5 and 6.

[10] D.I. Fivel, "A New Approach to the Axiomatic Foundations of Quantum Mechanics," Univ. of Maryland Technical Report No. 73-102, March 1973.

[11] P. Jordan, J. Von Neumann, and E. Wigner, Ann. Math. 35, 29 (1934).

[12] G. Köthe, *Topologische Lineare Räume I* (Springer-Verlag, Berlin, 1966), p. 3.

[13] A similar topology has been already introduced by D.I. Fivel in his work "A New Approach to the Axiomatic Founda-

tions of Quantum Mechanics," Univ. of Maryland Technical Report No. 73-102, March 1973. There is an error in his definition of topology which, however, can be corrected without difficulties. The two topologies, although looking quite similar to each other, in fact differ greatly. To obtain our results, we have to make much stronger requirements on the transition probability $p(a, b)$ than Fivel. Our assumptions can nevertheless be easily justified.

[14] S. T. Hu, *Elements of General Topology* (Holden-Day, San Francisco, 1964), Chap. 2.

[15] J. L. Kelley, *General Topology* (Van Nostrand, Princeton, N.J., 1955), p. 26.

[16] Yu. M. Smirnov, Usp. Mat. Nauk (N.S.) 6, no. 6 (46), 100—111 (1951), also published by Amer. Math. Soc. in 1953 as Translation No. 91.

[17] Our topology is such that we could also use the classical metrization theorem of Uryson, completed by a theorem of A. N. Tihonov, which states that every regular space with a countable basis is normal (Refs. 18, 19).

[18] F. Hausdorff, *Mengenlehre* (de Gruyter, Berlin-Leipzig, 1935), 3 Aufl., p. 139.

[19] P. S. Aleksandrov, *Introduction to the General Theory of Sets and Functions* (OGIZ, Moscow-Leningrad, 1948).

On the field of the quantum mechanical Hilbert space

Adolf Miklavc

Physik Department der Technischen Universität München, Theoretische Physik, 8046 Garching bei München, West Germany

(Received 12 July 1976)

The well-known limitations on the field of the quantum mechanical vector space, which, within the framework of the propositional logic of Birkhoff and von Neumann, can be obtained only for systems of dimension greater than or equal to four, are obtained here for all dimensions. The propositional logic fails to provide any information about the quantum systems of dimension less than four and, moreover, one does not know the empirical meaning of one of its basic assumptions. The Jordan algebra approach, on the other hand, which provides the same limitations on the field, also suffers from anomalies in low dimensions and is altogether formal rather than physical. In the present work, which is based on a recent study of the topological properties of quantum states, there are no low-dimensional anomalies and the assumptions have clear empirical meaning.

INTRODUCTION

Within both of the well-known axiomatic approaches to quantum mechanics, i.e., the propositional logic,[1,2] originated by Birkhoff and von Neumann, and the Jordan algebra,[3] there exists possibility of determining partially the quantum mechanical vector space. In the propositional calculus, one first shows that an irreducible system of propositions can be imbedded canonically into a projective geometry. There exists then the representation theorem which tells us that projective spaces of dimension > 2 (which in conventional quantum mechanics would correspond to systems with a basis of dimension $d = n + 1 > 3$) can be represented by vector spaces over division rings.[4] If one analyzes which of the division rings fulfil the requirements involved by the representation theorem, one is left with real, complex, or quaternionic fields.[5] The original work of Birkhoff and von Neumann was later clarified and improved by many authors. (See, e.g., Ref. 2 and the works mentioned therein.) However, a fundamental difficulty cannot be removed.[6] The representation theorem mentioned above exists only for systems of dimension ≥ 4 and so the propositional logic cannot tell us anything about the quantum mechanics of some systems which do occur in nature, e.g., spin systems with $j = \frac{1}{2}$ and 1.

The Jordan algebra axioms[3] lead to a much stronger representation theorem than the propositional logic.[6] One obtains real, complex, and quaternionic vector spaces for all finite dimensions and a small number of low-dimensional "anomalies," namely, one non-Desarguesian plane (Cayley—Moufang plane) correspond ing to the Jordan algebra M_3^8 for dimension $n = 2$, and the N-spheres for $n = 1$, corresponding to the Jordan algebras S_N. The axioms, however, are formal rather than physical.

It was recently shown,[7] on the basis of the well-known empirical facts about measurements, that there exists a topology in the set of quantum states and that the space with such a topology is homeomorphic to a subset of the unit sphere in the complex (or real) Hilbert space. These results hold true for any dimensionality of the quantum system. There are no "anomalies" in low dimensions, and the assumptions have clear empirical meaning. The question not fully discussed so far is whether the homeomorphism of the quantum states into the complex (or real) Hilbert space is the only one possible. We will continue the studies of the Ref. 7 by proving that such a homeomorphism cannot exist unless the Hilbert space is over real, complex, or quaternionic field. Even more can be shown to be true: A one-to-one correspondence between the quantum states $x \in S$ and any sequences $\{f_1(x), f_2(x), \cdots\}$, where f_n, $n = 1, 2, \cdots$, are continuous functions over S with the range in some continuous field F, is possible only if F is a real, complex, or quaternionic field.

CONNECTEDNESS OF QUANTUM STATES

First, let us us recall some of the important findings, obtained in the Ref. 7. Let S be the set of all quantum states of a physical system. Empirical observations lead us to the conclusion that there exists a function p, usually called transition probability, which maps the elements $(a,b) \in S \times S$ unto the points $p(a,b) \in [0,1]$ and has the following well-known properties:

(a) $p(a,b) = 1 \Longleftrightarrow a = b$,

(b) $p(a,b) = p(b,a)$, (1)

(c) given an $\epsilon > 0$, there exists a $\delta > 0$ such that, for all $z \in S$, $|p(z,x) - p(z,y)| < \epsilon$ if $1 - p(x,y) < \delta$.

One can then show that the sets

$$U_n(a) = \{x : x \in S,\ 1 - p(a,x) < 1/n\}, \quad n = 2, 3, \cdots, \tag{2}$$

form a fundamental system of neighborhoods (i.e., a neighborhood basis) at each point $a \in S$ and therefore define uniquely a topology τ in S. The requirement that our experiments be communicable, so that they can be repeated, makes this topology τ satisfy the second axiom of countability. The assumptions, introduced above, have clear empirical basis and are sufficent to prove that the set of quantum states S, with the topology τ, is homeomorphic to a subset of the unit sphere in the complex (or real) Hilbert space. These assumptions by themselves, however, do not seem to be complete enough to prove that homeomorphisms of S into the Hilbert spaces over other fields are impossible. We will now show that, together with another, apparently innocent property of quantum systems, they do restrict the field of the vector space to real, complex, or quaternionic numbers.

To make clear what the essence is of some of the arguments, we will first discuss the simple case of polarization states of a photon. We will then show that similar arguments can be made for any quantum system so that the subsequent analysis therefore will be of general validity.

Let S be the space of all polarization states of a photon, with the topology τ as introduced and described in Ref. 7 and above, and let A be the subspace of S, consisting of linearly polarized states only. A being a subspace of S means, of course, that its topology τ_A is the relative topology[8] with respect to τ. If $a \in A \subset S$, then the following class $V_A(a)$ of subsets of A is a τ_A-local (i.e., neighborhood) basis at $a \in A$ [8]

$$V_A(a) = \{A \cap U_n(a) \colon U_n(a) \in V(a)\}, \tag{3}$$

where $V(a) = \{U_n(a)\}$ is the τ-neighborhood basis at $a \in S$, defined in (2). It is clear, of course, that

$$A \cap U_n(a) = \{x \colon x \in A, 1 - p(a,x) < 1/n\}, \quad n = 2, 3, \cdots. \tag{4}$$

The elements of $V_A(a)$ we shall designate by $U_{A,n}(a)$.

Recall now a simple, but very important empirical observation concerning the linearly polarized states: If our polarizer at the moment produces the state a, we can obtain any other state $b \in A$ by simply rotating the apparatus by an angle ϕ_b around the appropriate axis. (By ϕ_b I mean the smallest such angle.) Even more: To each ϕ, $0 \leq \phi \leq \phi_b$, there corresponds one, and only one state in A. Thus, when we rotate the polarizer from the state a to the state b, the corresponding quantity ϕ/ϕ_b takes on all the values from 0 to 1 exactly once. These observations can be recast into a more formal language. A continuous function s of the unit interval $I = [0,1]$ into an arbitrary space X, i.e.,

$$s \colon I \to X$$

is usually called a path in the space X. It follows then, from what is said above, that for any two states a and b in A there exists a path $s \colon I \to A$ such that $s(0) = a$ and $s(1) = b$. The function s is continuous with respect to the relative topology τ_A. The subspace A is therefore pathwise connected.[9] Since A is pathwise connected, it is also connected. To see this, we pick an element $x_0 \in A$ and choose for each element $x \in A$ a path

$$s_x \colon I \to A$$

such that $s_x(0) = x_0$ and $s_x(1) = x$. As continuous image of a connected space I, $s_x(I)$ is a connected set for every $x \in A$. Furthermore, $x_0 \in s_x(I)$ for every $x \in A$ and

$$A = \bigcup_{x \in A} s_x(I),$$

and, since the union of any family of connected sets with a common point is connected, it follows that A is connected in the relative topology τ_A. This, however, implies that the subspace A is connected also with respect to the topology τ on S, since a subspace of any topological space (X, τ) is connected with respect to τ if and only if it is connected with respect to its own relative topology.[8]

We have thus established that the topological space (S, τ) of all polarization states of a photon contains at least one nonempty connected subset A, which is not a singleton. The same, however, can be said about the states of any quantum system. In principle, at least, we can always rotate a quantum system about an axis and then use the arguments given above to show that the states which correspond to particular angles of rotation form a connected subset in the space of all states of the system. Other operations may also be possible which can be used to show that there are connected subsets within the space of quantum states. Rotations of the system, however, are always possible and, as we just showed, they are sufficient to establish the existence of connected subsets of quantum states.

THE POSSIBLE VECTOR SPACES

It remains now to show that, because of the connected subsets existing in the space of quantum states S, only certain one-to-one maps are possible over S. In the Ref. 7, it was found that to every quantum state $x \in S$ there corresponds a vector $h(x)$ in the Hilbert space H, i.e.,

$$x \xrightarrow{h} \{f_1(x), f_2(x), \cdots\},$$

$$\sum_n |f_n(x)|^2 = 1, \text{ for every } x \in S, \tag{5}$$

and that this correspondence h is a homeomorphism. f_n, $n = 1, 2, \cdots$, are continuous, complex, or real valued functions over S. Continuity of the functions f_n is necessary for the function h to be continuous. This can be seen easily. The neighborhood $S_r[h(x)]$ of a $h(x)$ in H is defined by

$$S_r[h(x)] = \left\{ h(y) \colon \left[\sum_n |f_n(x) - f_n(y)|^2 \right]^{1/2} < r \right\}.$$

If h is continuous then for every neighborhood $S_r[h(x)]$ of $h(x) \in H$ there exists a neighborhood O_x of $x \in S$ such that

$$h(O_x) \subset S_r[h(x)].$$

It follows then immediately that the functions f_n, $n = 1, 2, \cdots$, are all continuous.

Being a homeomorphism, h is, of course, a one-to-one map. However, a one-to-one correspondence between quantum states $x \in S$ and the set of sequences $\{f_1(x), f_2(x), \cdots\}$ where f_n, $n = 1, 2, \cdots$, are continuous functions over S with their range in some continuous field F, can exist only if F is a real, complex, or quaternionic field. To show that this is indeed so, we first note that a topological field can only be either connected or totally disconnected.[10] Suppose first that F is a totally disconnected field. Single elements are then the largest connected sets in F. In the preceding section we found that there exist connected subsets in the space of quantum states S which are not singletons. Let $A \subset S$ be one of these connected subsets. Since every continuous image of a connected set is connected,[9] it follows that each function f_n maps all elements of the set A into the same element $a_n \in F$, i.e.,

$$f_n(A) = a_n \in F.$$

To every quantum state $x \in A$ there corresponds therefore the same sequence $\{a_1, a_2, \cdots\}$ of elements in F. It is clear that a one-to-one correspondence between the quantum states $x \in S$ and the set of sequences $\{f_1(x), f_2(x), \cdots\}$ cannot exist in this case. The values of the functions f_n, $n = 1, 2, \cdots$, can therefore lie only in a connected field. There are, however, only three continuous connected fields: the real, the complex, and the quaternionic field.[10] Herewith is our assertion proven.

It was shown in the Ref. 7 that homeomorphism exists between the space of quantum states S and a subset of the unit sphere in complex Hilbert space. Here, we found that quantum states can be mapped homeomorphically into Hilbert space only if this space is over a real, complex, or quaternionic field. Even more, a one-to-one correspondence between the states $x \in S$ and any sequences $\{f_1(x), f_2(x), \cdots\}$ where f_n, $n=1,2,\cdots$, are continuous functions over S with the range in some continuous field F, is not possible unless F is a real, complex, or quaternionic field. These results hold true for any dimensionality of the quantum system. There are no irregularities in low dimensions as they are known to exist in other well-known approaches. Also, the assumptions upon which our considerations are based have clear empirical meaning.

An interesting question emerges now naturally: Are the three possible quantum mechanics, i.e., the real, the complex, and the quaternionic quantum mechanics, equivalent? This question was already studied by a number of authors. (See, e.g., Ref. 2 and Refs. given therein.) So far it has been answered only partially. The real quantum mechanics was found[11] to be, up to a superselection rule, equivalent to the complex quantum mechanics, at least for simple systems. It has also been found[12] for the relativistic form of quaternionic quantum mechanics that it is equivalent to complex quantum mechanics in case of systems consisting of just one particle. It is not known if the same is true in case of more general systems.

ACKNOWLEDGMENTS

I wish to thank Professor D. I. Fivel for all I have learned from him on this subject. Furthermore, I wish to thank Professor Willi Rinow for helpful correspondence and Professor C. H. Woo for his comments and encouragment.

[1] G. Birkhoff and J. von Neumann, Ann. Math 37, 823 (1936).
[2] J.M. Jauch, *Foundations of Quantum Mechanics* (Addison-Wesley, Reading, Mass., 1968).
[3] P. Jordan, J. von Neumann and E. Wigner, Ann. Math. 35, 29 (1934).
[4] R.J. Bumcrot, *Modern Projective Geometry* (Holt, Rinehart, Winston, New York, 1969), Chap. 3.
[5] E.G. Betrametti and G. Cassinelli, Z. Naturforsch. **28a**, 1516 (1973), and the references therein.
[6] D.I. Fivel, "A New Approach to the Axiomatic Foundations of Quantum Mechanics," Univ. of Maryland Technical Report No. 73-102 March 1973.
[7] A. Miklavc, J. Math. Phys. **18**, 1521 (1977).
[8] S. Lipschutz, *General Topology*, Schaum Outline Series (McGraw-Hill, New York, 1965), Chaps. 5 and 6.
[9] S.T. Hu, *Elements of General Topology* (Holden-Day, San Francisco, 1965).
[10] L.S. Pontrjagin, *Topologische Gruppen* (Teubner-Verlag, Leipzig, 1957), Teil 7, Chap. 4.
[11] E.C.G. Stueckelberg, Helv. Phys. Acta 33, 727 (1960).
[12] G. Emch, Helv. Phys. Acta 36, 739, 770 (1963).

Symmetries of the stationary Einstein–Maxwell field equations. I*

William Kinnersley

Department of Physics, Montana State University, Bozeman, Montana 59715

(Received 29 November 1976)

The Einstein equations for stationary axially symmetric gravitational fields are written in several extremely simple forms. Using a tensor generalization of the Ernst potential, we give forms that are manifestly covariant under (i) the external group **G** of coordinate transformations, (ii) the internal group **H** of Ehlers transformations and gage transformations, and (iii) the infinite parameter group **K** of Geroch which combines both. We then show how the same thing can be done to the Einstein–Maxwell equations. The enlarged internal group **H**′ now includes the Harrison transformations, and is isomorphic to SU(2,1). The enlarged group **K**′ contains even more parameters, and generates even more potentials and conservation laws.

1. INTRODUCTION

The static axially symmetric class of vacuum gravitational fields was completely solved by Weyl in 1917. Even today it still represents the largest collection of realistic exact solutions of the Einstein field equations that we have available. The *stationary* axially symmetric fields are the ones which should logically be considered next, both from a mathematical and a physical point of view. The stationary field equations have indeed been the target of many recent interesting investigations. Despite their apparent simplicity, the equations have not yet yielded more than a handful of solutions.

They *have* been found to contain a remarkable amount of symmetry and internal structure. Moreover, when the stationary Einstein—Maxwell equations are considered, the symmetry group persists, and is considerably enlarged. In the belief that such a beautiful and unexpected structure cannot be a mere accident, we have tried to systematically explore this approach to the stationary problem. In as much as the symmetries may be used to produce new solutions from old ones, it is entirely possible that a complete understanding of the symmetry group may eventually prove the key to finding the general stationary solution.

Since the existing publications on this topic have been somewhat compressed, one aim of the present paper is to give a simpler and more detailed discussion of our current understanding. We hope thereby to attract a wider interest to this important problem.

2. NOTATION

Stationary axially symmetric gravitational fields are spacetimes which contain two commuting Killing vectors. We may therefore choose the metric to be independent of two of the coordinates, say $x^1=t$ and $x^2=\varphi$. We require also that the spacetime possess "orthogonal transitivity." Physically, this assumes the existence of a reflection symmetry, or motion reversal,

$$(t,\varphi)\to(-t,\,-\varphi).$$

Mathematically it means that the metric must be block diagonal,

$$\begin{aligned}&ds^2=ds_1^2-ds_2^2,\\&ds_1^2=f_{AB}dx^Adx^B,\quad A,B=1,2,\\&ds_2^2=h_{MN}dx^Mdx^N,\quad M,N=3,4.\end{aligned}\tag{2.1}$$

To raise indices in a two-dimensional space we may use either the inverse metric $(f_{AB})^{-1}$ or the alternating symbol $\epsilon^{AB}=\pm1$, and both choices have certain advantages. Unless otherwise stated, we will raise indices using ϵ^{AB} and $(h_{MN})^{-1}$. When it becomes necessary to raise an index using $(f_{AB})^{-1}$ or ϵ^{MN}, that index will be marked with a tilde.

For example, since

$$(f_{AB})^{-1}=-\rho^{-2}\epsilon^{AC}\epsilon^{BD}f_{CD},$$

where

$$\rho^2\equiv-\det(f_{AB})$$

we will have

$$f^{AB}=-\rho^2f^{\tilde A\tilde B},\quad f^{AB}f_{BC}=-\rho^2\delta^A{}_C.\tag{2.2}$$

To express derivatives we may use ∇_2, the two-dimensional covariant derivative associated with ds_2^2. Thus

$$\nabla_2\bullet\mathbf{V}=h^{-1/2}(h^{1/2}h^{MN}V_M)_{,N}.$$

However in two dimensions the expression $h^{-1/2}h^{MN}$ is conformally invariant. Therefore, we might also consider some other 2-metric $\overline{ds}_2^2$ which is conformally related to ds_2^2,

$$ds_2^2=e^{2\Gamma}\overline{ds}_2^2.\tag{2.3}$$

Then $\nabla_2\bullet\mathbf{V}=0$ and $\overline{\nabla}_2\circ\mathbf{V}=0$ are equivalent. If desired, we can also write the equations in terms of a 3-metric,

$$ds_3^2=\overline{ds}_2^2+\rho^2d\varphi^2,$$

in which case

$$\nabla_3\bullet\mathbf{V}=\rho^{-1}\overline{\nabla}_2\circ(\rho\mathbf{V}).\tag{2.4}$$

The simplest approach is to now choose $\overline{h}_{MN}=\delta_{MN}$. Acting on a vector

$$\mathbf{V}=(V_M)=(V_3,V_4)$$

the tilde operation becomes

$$\tilde{\mathbf{V}}=(\overline{h}_{MN}\epsilon^{NP}V_P)=(V_4,-V_3)\tag{2.5}$$

and satisfies

$$\tilde{\tilde{\mathbf{V}}} = -\mathbf{V}, \tag{2.6}$$

$$\tilde{\mathbf{V}} \circ \mathbf{W} = -\mathbf{V} \circ \tilde{\mathbf{W}}. \tag{2.7}$$

We can use it to define a second set of ∇'s, e.g.,

$$\tilde{\nabla}_2 = (\partial_4, -\partial_3). \tag{2.8}$$

For all scalars U we have the identity

$$\nabla_2 \circ (\tilde{\nabla}_2 U) = 0, \quad \nabla_3 \circ (\rho^{-1}\tilde{\nabla}_3 U) = 0. \tag{2.9}$$

The vacuum field equations fall into three groups: R_{AB}, R_{AM}, R_{MN}, of which R_{AM} vanishes automatically for our block diagonal metric. We find that

$$R^A{}_C = -\tfrac{1}{2}\rho\nabla_2 \circ (\rho^{-1} f^{AB} \nabla_2 f_{BC}). \tag{2.10}$$

Hence this portion of the field equations may be written in the form of a vanishing divergence

$$\nabla_2 \circ (\rho^{-1} f^{AB} \nabla_2 f_{BC}) = 0. \tag{2.11}$$

The remaining field equations are

$$\begin{aligned} R_{MN} = 0 = {}^{(2)}R_{MN} &+ \nabla_2 \nabla_2 \ln\rho \\ &+ \tfrac{1}{4}\rho^{-4} f^{BC} f^{AD} (\nabla_2 f_{AB})(\nabla_2 f_{CD}). \end{aligned} \tag{2.12}$$

They may be written in terms of the flat 2-metric $\overline{ds_2^2}$, using

$${}^{(2)}R_{MN} = \delta_{MN} \overline{\nabla}_2^2 \Gamma$$

and eventually lead to integrable equations for $\overline{\nabla}_2 \Gamma$, which can be solved once f_{AB} is known. They are not needed in the rest of this paper.

3. THE GROUP G

The field equations Eq. (2.11) have been written in a form which is manifestly covariant under linear transformations[1] of the coordinates. One such transformation is the rescaling

$$t \to \Lambda t, \quad \varphi \to \Lambda \varphi. \tag{3.1}$$

The rest of the transformations form a three-parameter group **G**, isomorphic to SL(2,R). For its three generators we may take[2]

$$t \to t + a\varphi, \quad \varphi \to \phi, \tag{3.2}$$

$$t \to t, \quad \varphi \to \varphi + bt, \tag{3.3}$$

$$t \to ct, \quad \varphi \to c^{-1}\varphi. \tag{3.4}$$

As is well known, SL(2,R) is isomorphic to SO(2,1), the three-dimensional Minkowski group, and it is often convenient to recast the equations in language appropriate to SO(2,1). An SL(2,R) symmetric tensor T_{AB} is equivalent to SO(2,1) vector T_a. The correspondence is given by a connecting quantity $\sigma_a{}^{AB}$, analogous to the Pauli matrices,

$$T_a = \sigma_a{}^{AB} T_{AB}, \quad T_{AB} = \sigma^a{}_{AB} T_a. \tag{3.5}$$

SO(2,1) indices are raised and lowered by means of the metric tensor

$$G_{ab} = \sigma_a{}^{AB} \sigma_{bAB}. \tag{3.6}$$

In particular, to obtain the desired correspondence

$$T_{11} \longleftrightarrow T_1, \; T_{12} \longleftrightarrow T_2, \quad T_{22} \longleftrightarrow T_3,$$

we use the matrices

$$\sigma_1^{AB} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \sigma_2{}^{AB} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}, \quad \sigma_3{}^{AB} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.7}$$

This implies the metric tensor

$$G_{ab} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -\frac{1}{2} & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{3.8}$$

The invariant norm of an SO(2,1) vector is then

$$G^{ab} T_a T_b = \epsilon^{AC} \epsilon^{BD} T_{AB} T_{CD},$$

$$2(T_1 T_3 - T_2 T_2) = 2(T_{11} T_{22} - T_{12} T_{12}).$$

We can now translate Eq. (2.11) into SO(2,1) language, putting

$$f_{AB} = f_a \sigma^a{}_{AB}$$

and using the relationship[3]

$$\sigma^{aAB} \sigma^b{}_{BC} = \tfrac{1}{2} G^{ab} \delta^A{}_C - \varepsilon^{abc} \sigma_c{}^A{}_C. \tag{3.9}$$

We are left with two terms which must vanish independently: the trace[4]

$$\nabla_3 \circ [\rho^{-2} f^a \nabla f_a] = \nabla_3 \circ [\rho^{-1} \nabla \rho] = 0, \tag{3.10}$$

and

$$\begin{aligned} &\nabla_3 \cdot \mathbf{v}^c = 0, \\ &\mathbf{v}^c = \rho^{-2} \varepsilon^{abc} f_a \nabla f_b. \end{aligned} \tag{3.11}$$

For any scalar function U, we have the identity

$$\nabla_3 \circ [\rho^{-1} \tilde{\nabla} U] = 0. \tag{3.12}$$

Conversely, any conservation law

$$\nabla_3 \circ \mathbf{V} = 0$$

may be regarded as the integrability condition for the existence of a potential U such that

$$\mathbf{V} = \rho^{-1} \tilde{\nabla} U. \tag{3.13}$$

Our field equations therefore imply the existence of a set of "twist potentials" $\psi^A{}_C$,

$$\begin{aligned} &\rho^{-1} \tilde{\nabla} \psi^A{}_C = \rho^{-2} f^{AB} \nabla f_{BC}, \\ &\nabla \psi_{AC} = -\rho^{-1} f_A{}^B \tilde{\nabla} f_{BC}. \end{aligned} \tag{3.14}$$

This equation can also be split into its trace,

$$\nabla \psi^A{}_A = 2 \tilde{\nabla} \rho, \tag{3.15}$$

and a symmetric part, equivalent to an SO(2,1) vector

$$\nabla \psi^a = \rho^{-1} \varepsilon^{abc} f_b \tilde{\nabla} f_c. \tag{3.16}$$

Equations (3.14) and (3.16) may be inverted and solved for ∇f_{AB},

$$\begin{aligned} &\nabla f_{AB} = \rho^{-1} f_A{}^C \tilde{\nabla} \psi_{CB}, \\ &\nabla f_a = -\tfrac{1}{2} \rho^{-1} \varepsilon_{abc} f^b \tilde{\nabla} \psi^c - f_a \rho^{-1} \nabla \rho. \end{aligned} \tag{3.17}$$

The inverted forms imply further field equations, obeyed by $\psi^A{}_C$,

$$\begin{aligned} &\nabla_3 \cdot [\rho^{-2} f_A{}^B \nabla \psi_{BC}] = 0, \\ &\nabla_3 \cdot [\rho^{-2} \varepsilon_{abc} f^b \nabla \psi^c - 2\rho^{-2} f_a \tilde{\nabla} \rho] = 0. \end{aligned} \tag{3.18}$$

These are of course nothing more than the integrability conditions for the existence of the original metric functions f_{AB}, given ψ_{AC}.

In terms of complex Ernst potentials

$$\mathcal{E}_{AB}=f_{AB}+i\psi_{AB}\,, \tag{3.19}$$

Eqs. (3.14) and (3.17) may be combined into a single complex equation,

$$\nabla\mathcal{E}_{AB}=-i\rho^{-1}f_A{}^C\tilde{\nabla}\mathcal{E}_{CB}. \tag{3.20}$$

4. THE GROUP H

Following Lewis and others,[5] we now focus on a particular parametrization for f_{AB},

$$f_{11}=f,\quad f_{12}=-f\omega,\quad f_{22}=f\omega^2-\rho^2 f^{-1}. \tag{4.1}$$

We have

$$\mathbf{v}^1=\rho^{-1}\tilde{\nabla}\psi^1=2\omega f^{-1}\nabla f+\nabla\omega + \rho^{-2}f^2\omega^2\nabla\omega-2\rho^{-1}\omega\nabla\rho, \tag{4.2}$$

$$\mathbf{v}^2=\rho^{-1}\tilde{\nabla}\psi^2=2(f^{-1}\nabla f + \rho^{-2}f^2\omega\nabla\omega)-2\rho^{-1}\nabla\rho, \tag{4.3}$$

$$\mathbf{v}^3=\rho^{-1}\tilde{\nabla}\psi^3=\rho^{-2}f^2\nabla\omega. \tag{4.4}$$

Because of the identities

$$f_a\mathbf{v}^a=\nabla f_a\bullet\mathbf{v}^a=0,$$

the three field equations $\nabla\circ\mathbf{v}^a=0$ are not algebraically independent. Any two of them suffice to determine f,ω and hence the metric. E.g., for $a=2,\ 3$,

$$\nabla_3\circ[f^{-1}\nabla f+\rho^{-2}f^2\omega\nabla\omega]=0, \tag{4.5}$$

$$\nabla_3\circ[\rho^{-2}f^2\nabla\omega]=0. \tag{4.6}$$

In this f,ω formalism the G covariance of the equations is obscured, since the transformations of f,ω under G are not especially simple. The reason for introducing it is that it enables us to find a second invariance group H. We must now proceed in a way which is not G covariant, eliminating ω from the field equations in favor of one of the ψ^a's.

Let $\psi\equiv\psi_1$. Equation (4.4) is

$$\nabla\psi=-\rho^{-1}f^2\tilde{\nabla}\omega,\quad \nabla\omega=\rho f^{-2}\tilde{\nabla}\psi \tag{4.7}$$

and using this in Eq. (4.3) we obtain

$$\begin{aligned}\nabla_3\circ&[f^{-1}\nabla f+\rho^{-1}\omega\tilde{\nabla}\psi]\\ &=\nabla_3\circ[f^{-1}\nabla f+\rho^{-1}\tilde{\nabla}(\omega\psi)-\rho^{-1}\psi\tilde{\nabla}\omega]\\ &=\nabla_3\circ[f^{-1}\nabla f+f^{-2}\psi\nabla\psi]=0.\end{aligned} \tag{4.8}$$

This result, taken together with the integrability condition for ω,

$$\nabla_3\circ[f^{-2}\nabla\psi]=0, \tag{4.9}$$

gives us a set of two divergence equations for the two variables f,ψ.

The close analogy of this pair with the original pair, Eqs. (4.5) and (4.6) was pointed out by Neugebauer and Kramer,[6] who discovered that the mapping

$$f\to\rho f^{-1},\quad \omega\to i\psi \tag{4.10}$$

transforms one pair into the other *directly*. What this means is that Eqs. (4.8) and (4.9) *also* possess an invariance group H, isomorphic to SO(2,1). The images of f_a under the map must form a 3-vector F_a under **H**. Removing an invariant factor ρ, they are

$$\begin{aligned}&F_1=f^{-1},\quad F_2=f^{-1}\psi,\\ &F_3=f^{-1}(f^2+\psi^2),\quad F^aF_a=+2.\end{aligned} \tag{4.11}$$

The field equations, Eqs. (4.8) and (4.9), are now seen to be the $a=2$, 3 components of an H covariant set

$$\nabla_3\bullet\mathbf{V}^a=0,\quad \mathbf{V}^a=\varepsilon^{abc}F_b\nabla F_c. \tag{4.12}$$

The $a=1$ component,

$$\nabla_3\bullet[2f^{-1}\psi\nabla f+f^{-2}(\psi^2-f^2)\nabla\psi]=0, \tag{4.13}$$

can easily be shown to follow as an algebraic consequence of the other two.

While H acts linearly among the three particular expressions F_a, it produces a nonlinear action on f,ψ themselves. Three particular symmetry transformations which may be taken as the three generators of H are

$$f\to f,\quad \psi\to\psi-\alpha, \tag{4.14}$$

$$f\to\beta f,\quad \psi\to\beta\psi, \tag{4.15}$$

$$f\to\frac{f}{(1-\gamma\psi)^2+\gamma^2f^2}\,,\quad \psi\to\frac{\psi-\gamma(f^2+\psi^2)}{(1-\gamma\psi)^2+\gamma^2f^2}. \tag{4.16}$$

The first of these is a gage transformation, originating in the fact that Eq. (4.7) defines ψ only up to a constant. The second is a rescaling,[7] and the third is the gravitational duality rotation discovered by Ehlers.[8,9]

We continue to apply to H the same procedure we used for G. Equation (4.12) permits us to define a set of potentials Ψ^a which form an H vector,

$$\nabla\Psi^a=\rho\varepsilon^{abc}F_b\tilde{\nabla}F_c. \tag{4.17}$$

Inverting,

$$\nabla F_a=\tfrac{1}{2}\rho^{-1}\varepsilon_{abc}F^b\tilde{\nabla}\Psi^c, \tag{4.18}$$

we find necessary field equations for Ψ^a,

$$\nabla_3\bullet[\rho^{-2}\varepsilon_{abc}F^b\nabla\Psi^c]=0. \tag{4.19}$$

However the potentials Ψ^a are not entirely new. We can show that a partial identification of Ψ^a with some of the previous variables is possible. Writing Eq. (4.17) out explictly,

$$\begin{aligned}&\nabla\Psi_1=\rho f^{-2}\tilde{\nabla}\psi,\\ &\nabla\Psi_2=\rho f^{-2}(f\tilde{\nabla}f+\psi\tilde{\nabla}\psi),\\ &\nabla\Psi_3=\rho f^{-2}(2f\psi\tilde{\nabla}f+(\psi^2-f^2)\tilde{\nabla}\psi),\end{aligned}$$

and comparing with Eqs. (4.4) and (4.3), we find

$$\begin{aligned}&\nabla\Psi_1=\nabla\omega,\\ &\nabla\Psi_2=\nabla(\omega\psi+\psi_2)+\tilde{\nabla}\rho.\end{aligned} \tag{4.20}$$

Recall that $\nabla_2^2\rho=0$. We let z denote the harmonic function conjugate to it. Then

$$\tilde{\nabla}\rho=-\nabla z$$

and by a proper choice of the integration constants we can put

$$\Psi_1=\omega,\quad \Psi_2=\omega\psi+\psi_2-z. \tag{4.21}$$

Knowledge of Ψ_a is indispensable if these symmetries are to be used to generate new solutions. It is the relation between Ψ_1 and ω that permits us to carry the action of H back to the original metric variables. For example under the Ehlers transformation, Eq. (4.16), we can now write

$$\omega \rightarrow \omega - 2\gamma\Psi_2 + \gamma^2\Psi_3. \qquad (4.22)$$

Discovery of an internal symmetry group like **H** takes great ingenuity if one must construct *ab initio* a nonlinear transformation like Eq. (4.16). It is far more reasonable to begin by seeking variables such as F_a which make the action linear.

We might mention here another possible approach to linearizing H. This approach requires greatly enlarging the number of variables. Equations (4.14)—(4.16) show that f, ψ form a two-dimensional *nonlinear realization* of SO(2, 1). We will indicate how to convert this realization into an infinite-dimensional linear one. In terms of the Ernst potential

$$\mathcal{E} = f + i\psi, \qquad (4.23)$$

Eq. (4.16) is

$$\mathcal{E} \rightarrow \frac{\mathcal{E}}{1 + i\gamma\mathcal{E}}. \qquad (4.24)$$

Expanding in a power series,

$$\mathcal{E} \rightarrow \sum_0^\infty (-i\gamma)^n \mathcal{E}^{n+1}$$

suggests that the set of functions $\{\mathcal{E}^n, n = 1, 2, \cdots\}$ forms the basis of a linear representation. In fact, when further infinitesimal transformations are applied to $\{\mathcal{E}^n\}$ we find

$$\mathcal{E}^n \rightarrow \mathcal{E}^n - \alpha n\mathcal{E}^{n-1}, \quad \mathcal{E}^n \rightarrow \mathcal{E}^n + \gamma n \mathcal{E}^{n+1}. \qquad (4.25)$$

Comparison with a list of representations[10] of SO(2,1) shows that $\{\mathcal{E}^n\}$ transforms according to $\mathcal{D}^+(0)$, the tail of the scalar representation.

As a further illustration of this approach, apply the Ehlers tranformation to Eq. (4.7),

$$\nabla\left(\frac{\psi - \gamma(f^2 + \psi^2)}{(1 - \gamma\psi)^2 + \gamma^2 f^2}\right) = -\rho^{-1} f^2 \frac{\tilde{\nabla}(\omega - 2\gamma\Psi_2 + \gamma^2\Psi_3)}{[(1 - \gamma\psi)^2 + \gamma^2 f^2]^2}.$$

Expand as a power series in γ and equate coefficients. In this way an infinite sequence of potentials is generated, and corresponding to each one of them is a conservation law. For example, from γ^1,

$$\nabla(\psi^2 - f^2) = -2\rho^{-1} f^2 (2\psi\tilde{\nabla}\omega - \tilde{\nabla}\Psi_2)$$
$$\Rightarrow \nabla_3 \cdot [\rho^{-2} f^2 (2\psi\nabla\omega - \nabla\Psi_2)] = 0.$$

Of course these potentials will be functionally dependent; their main interest lies in the linear action produced upon them by **H**.

To avoid any possible confusion which might arise, we would like to show how to translate our own approach to the one used by Geroch.[11] Geroch describes the Ehlers transformation as

$$g_{\mu\nu} \rightarrow \tilde{g}_{\mu\nu} = \tilde{\lambda}^{-1}\tilde{\xi}_\mu\tilde{\xi}_\nu + \lambda\tilde{\lambda}^{-1}(g_{\mu\nu} - \lambda^{-1}\xi_\mu\xi_\nu), \qquad (4.26)$$

where

$$\tilde{\lambda} = \frac{\lambda}{(\cos\theta - \omega\sin\theta)^2 + \lambda^2\sin^2\theta},$$
$$\tilde{\xi}_\mu = \xi_\mu + 2\tilde{\lambda}\alpha_\mu \sin\theta\cos\theta - \tilde{\lambda}\beta_\mu\sin^2\theta. \qquad (4.27)$$

In our notation, λ is f, ω is ψ, and

$$\xi_\mu = (f, -f\omega, 0, 0),$$
$$\alpha_\mu = (\psi, -\omega\psi + \Psi_2, 0, 0), \qquad (4.28)$$
$$\beta_\mu = (f^2 + \psi^2 - 1,\ -\omega(f^2 + \psi^2) + \Psi_3, 0, 0).$$

We get Geroch's transformation with the choice of parameters

$$\gamma = \tan\theta, \quad \beta = \sec^2\theta \qquad (4.29)$$

performed in that order.

5. THE GROUP K

Since the preceding discussion of **H** was carried out in a non-**G**-covariant way, we must now consider what happens when the coordinate transformations of **G** are applied in combination with the internal symmetries of **H**. Conjugation with elements of **G** should lead to other internal groups, $\hat{\mathbf{H}} = g\mathbf{H}g^{-1}$, but it is not clear *a priori* how many such groups exist. What is needed is a description of the total symmetry group **K** of the stationary vacuum Einstein equations, in which the subgroups **G** and **H** have not been given preferential treatment.

As Geroch has shown,[12] the infinitesimal part of **K** can be built up by induction. The starting point is to write the infinitesimal transformations of **H** in a form which suggests their **G**-covariant generalization. Thus for the Ehlers transformation, we have, to first order in γ,

$$f \rightarrow f + 2\gamma f\psi, \quad \omega \rightarrow \omega - 2\gamma\Psi_2. \qquad (5.1)$$

and this suffices to determine the action on all three components of f^a. The three transformations can be written together as

$$f^a \rightarrow f^a + 2\gamma(f^3\psi^a - G^{a3}(f_b\psi^b) + 2\varepsilon^{ab3}f_b z).$$

The parameter γ is evidently the 3-component of a **G**-vector γ_a, and the generalized form is

$$f^a \rightarrow f^a + 2\gamma_c(f^c\psi^a - G^{ac}(f_b\psi^b) + 2\varepsilon^{abc}f_b z). \qquad (5.2)$$

The number of subgroups conjugate to **H** is therefore three. We denote them $\mathbf{H}_a$.

As Geroch has also pointed out, one of these other subgroups (H_1, corresponding to γ_1) could have been easily anticipated. H_1 will be obtained in place of H_3 if we interchange the two Killing vectors. The transformations of H_2, on the other hand, are an essentially new and unexpected feature. Unfortunately they cannot be used for the practical generation of new solutions until they have been written down for finite values of the parameters, and so far this has not been accomplished.

Proceeding in similar fashion we obtain the transformation of ψ^a under **H**,

$$\psi^a \rightarrow \psi^a + \gamma_c(\psi^c\psi^a - f^c f^a + 2\varepsilon^{acb}\pi_b), \qquad (5.3)$$

where π^a is defined by

$$\nabla\pi^a = \varepsilon^{abc}(\psi_b\nabla\psi_c + f_b\nabla f_c) + 4z\nabla\psi^a. \tag{5.4}$$

We can also consider the action produced upon F_a, Ψ_a by the infinitesimal transformations of **G**. Under Eq. (3.2),

$$f\to f,\quad \omega\to\omega+a. \tag{5.5}$$

This too can be regarded as a gage transformation connected with the arbitrary additive constant in Eq. (4.7).

Under Eq. (3.3), to first order in b,

$$\begin{aligned} &f\to f+2bf\omega,\\ &\omega\to\omega-b(\omega^2+\rho^2 f^{-2}),\\ &\psi\to\psi-2b\psi_2,\end{aligned} \tag{5.6}$$

and we find

$$F^a\to F^a - 2b_c(F^a\Psi^c - G^{ac}(F_b\Psi^b) + 2\varepsilon^{acb}F_b z), \tag{5.7}$$

$$\Psi^a\to\Psi^a - b_c(\Psi^c\Psi^a + \rho^2F^cF^a + 2\varepsilon^{acb}\Pi_b z), \tag{5.8}$$

with Π^a defined by

$$\begin{aligned}\nabla\Pi^a = {}&\varepsilon^{abc}(\Psi_b\nabla\Psi_c + \rho^2F_b\nabla F_c)\\ &+4z\nabla\Psi^a.\end{aligned} \tag{5.9}$$

A great deal more remains to be done to elucidate the complete structure of **K**, but in the present paper we will not pursue this topic further. We turn instead to the discussion of a more general problem.

6. EINSTEIN-MAXWELL EQUATIONS

Suppose now that the spacetime also contains an electromagnetic field,

$$F_{\mu\nu} = A_{\nu,\mu} - A_{\mu,\nu}. \tag{6.1}$$

If the spacetime is stationary and axially symmetric, then $F_{\mu\nu}$ is independent of t and φ. We further assume that A_μ can be chosen to be independent of t, φ as well. The only surviving field components will be

$$F_{AM} = -A_{A,M},\quad F_{MN} = A_{N,M} - A_{M,N}.$$

The Maxwell equations which remain to be solved are

$$F^{\mu\nu}{}_{;\nu} = 0 = (\sqrt{-g}\,F^{\mu\nu})_{,\nu}$$

and fall into two sets. The first is

$$0 = (\sqrt{-g}\,F^{MN})_{,N} = \nabla_2\circ(\rho F^{MN}) \Rightarrow F^{MN} = C\rho^{-1}\epsilon^{MN}. \tag{6.2}$$

Such a term in $F^{\mu\nu}$ corresponds to a magnetic field in the φ direction, falling off as ρ^{-1}. It would be produced by a line current along the symmetry axis, which is a situation we want to exclude. We assume $C=0$, and hence also $A_N=0$. The final set of equations is

$$\begin{aligned} &0 = (\sqrt{-g}\,F^{AM})_{,M}\\ &\;\;= \nabla_2\circ(\rho f^{\tilde{A}\tilde{B}}\nabla A_B)\\ &\Rightarrow 0 = \nabla_3\circ(\rho^{-2}f^{AB}\nabla A_B).\end{aligned} \tag{6.3}$$

Now we need the stress-energy tensor

$$4\pi T_{\mu\nu} = F_\mu{}^\sigma F_{\nu\sigma} - \tfrac{1}{4}g_{\mu\nu}F_\sigma{}^\tau F^\sigma{}_\tau.$$

we find that

$$\begin{aligned}4\pi T^A{}_B = {}&f^{AC}\nabla A_C\cdot\nabla A_B\\ &-\tfrac{1}{2}\delta^A{}_B f^{CD}\nabla A_C\cdot\nabla A_D.\end{aligned}$$

Making use of Eq. (6.3), this can be written as a divergence,

$$\begin{aligned}4\pi T^A{}_B = {}&\rho\nabla_2\cdot[\rho^{-1}(f^{AC}A_B\nabla A_C\\ &-\tfrac{1}{2}\delta^A{}_B f^{CD}A_C\nabla A_D)].\end{aligned} \tag{6.4}$$

Since $T^\mu{}_\mu = T^A{}_A = 0$, the Einstein–Maxwell equations are

$$R^A{}_B = -8\pi T^A{}_B.$$

Using Eq. (2.10) we once again have a field equation in the form of a total divergence,

$$\begin{aligned}&\nabla_2\circ[\rho^{-1}(f^{AC}\nabla f_{BC} - 4f^{AC}A_B\nabla A_C\\ &\quad+2\delta^A{}_B f^{CD}A_C\nabla A_D)] = 0.\end{aligned} \tag{6.5}$$

Equivalently, using the identity

$$V^A{}_B - V_B{}^A = \delta^A{}_B V^C{}_C,$$

we can write it as

$$\begin{aligned}&\nabla_3\circ[\rho^{-2}(f^{AC}\nabla f_{BC} - 2f^{AC}A_B\nabla A_C\\ &\quad-2f_B{}^C A^A\nabla A_C)] = 0.\end{aligned} \tag{6.6}$$

As before, the field equations imply the existence of potentials B_A, ψ_{AB},

$$\nabla B_A = -\rho^{-1}f_A{}^B\tilde{\nabla}A_B, \tag{6.7}$$

$$\begin{aligned}\nabla\psi_{AC} = {}&-\rho^{-1}(f_A{}^B\tilde{\nabla}f_{BC} - 2f_A{}^BA_C\tilde{\nabla}A_B\\ &-2f_C{}^BA_A\tilde{\nabla}A_B).\end{aligned} \tag{6.8}$$

The inverse relations

$$\nabla A_A = \rho^{-1}f_A{}^B\tilde{\nabla}B_B, \tag{6.9}$$

$$\nabla f_{AC} = \rho^{-1}f_A{}^B(\tilde{\nabla}\psi_{BC} + 2A_C\,\tilde{\nabla}B_B + 2A_B\tilde{\nabla}B_C), \tag{6.10}$$

yield further field equations

$$\nabla_3\circ[\rho^{-2}f_A{}^B\nabla B_B] = 0, \tag{6.11}$$

$$\nabla_3\circ[\rho^{-2}f_A{}^B(\nabla\psi_{BC} + 2A_C\nabla B_B + 2A_B\nabla B_C)] = 0. \tag{6.12}$$

We also have the condition necessary to maintain the symmetry of f_{AC},

$$f^{AB}(\nabla\psi_{BA} + 2A_A\nabla B_B + 2A_B\nabla B_A) = 0. \tag{6.13}$$

Just as in the vacuum case, a simpler form for the field equations can be obtained by introduction of appropriate complex potentials. Letting

$$\Phi_A = A_A + iB_A, \tag{6.14}$$

Eqs. (6.7) and (6.9) can be combined,

$$\nabla\Phi_A = -i\rho^{-1}f_A{}^B\tilde{\nabla}\Phi_B. \tag{6.15}$$

To do the same for Eq. (6.8), we must first bring f_{AB} outside the parenthesis. Let

$$\Omega_{AC} = \psi_{AC} + 2A_AB_C \tag{6.16}$$

$$\begin{aligned}\Rightarrow\nabla\Omega_{AC} = {}&-\rho^{-1}f_A{}^B(\tilde{\nabla}f_{BC} - 2A_C\tilde{\nabla}A_B\\ &-2B_C\tilde{\nabla}B_B).\end{aligned} \tag{6.17}$$

Equations (6.10) and (6.17) can now be written as one. Let

$$\mathcal{E}_{AC}=f_{AC}+i\Omega_{AC}. \tag{6.18}$$

Then

$$\nabla\mathcal{E}_{AC}=-i\rho^{-1}f_A{}^B(\tilde{\nabla}\mathcal{E}_{BC}-2\Phi_C\tilde{\nabla}\Phi_B{}^*), \tag{6.19}$$

$$\nabla\mathcal{E}_{AC}-\Phi_C\nabla\Phi_A{}^*=-i\rho^{-1}f_A{}^B(\tilde{\nabla}\mathcal{E}_{BC}-\Phi_C\tilde{\nabla}\Phi_B{}^*). \tag{6.20}$$

Alternatively in terms of the potential

$$\mathcal{G}_{AC}=\mathcal{E}_{AC}-\Phi_A{}^*\Phi_C, \tag{6.21}$$

$$\nabla\mathcal{G}_{AC}+\Phi_A{}^*\nabla\Phi_C=-i\rho^{-1}f_A{}^B(\tilde{\nabla}\mathcal{G}_{BC}+\Phi_B{}^*\nabla\Phi_C). \tag{6.22}$$

Finally, in terms of

$$\mathcal{H}_{AC}=\mathcal{G}_{AC}+\epsilon_{AC}K, \tag{6.23}$$

where K is given by Eq. (8.34), we obtain the Einstein–Maxwell equations in a most attractive form, analogous to Eq. (6.15), and the vacuum case Eq. (3.20),

$$\nabla\mathcal{H}_{AC}=-i\rho^{-1}f_A{}^B\nabla\mathcal{H}_{BC}. \tag{6.24}$$

These equations are all invariant under an enlarged external symmetry group $\mathbf{G}'$. $\mathbf{G}'$ consists of the coordinate transformations $\mathbf{G}$, plus the electromagnetic and gravitational gage transformations

$$\Phi_A\to\Phi_A+a_A, \tag{6.25}$$

$$\mathcal{G}_{AC}\to\mathcal{G}_{AC}-a_A{}^*\Phi_C-\Phi_A a^*{}_C-a^*{}_A a_C+i\alpha_{AC}, \tag{6.26}$$

where the arbitrary constants α_{AC} and a_A are respectively real and complex. (These are the most general linear transformations of Φ_A and $\mathcal{G}_{AC}$ that preserve f_{AC}.)

7. THE GROUP H′

Just as we did in the vacuum case, we now abandon $\mathbf{G}'$ covariance. We transform to an "internal" set of variables, using only *some* of the B_A, ψ_{AC} to eliminate *some* of the A_A, f_{AC}.

Written out in terms of f and ω, Eqs. (6.7) and (6.8) are:

$$\rho^{-1}\tilde{\nabla}B_1=-\rho^{-2}f(\nabla A_2+\omega\nabla A_1), \tag{7.1}$$

$$\rho^{-1}\tilde{\nabla}B_2=-f^{-1}\nabla A_1+\rho^{-2}f\omega(\nabla A_2+\omega\nabla A_1), \tag{7.2}$$

$$\rho^{-1}\tilde{\nabla}\psi_{11}=\rho^{-2}f^2\nabla\omega+4\rho^{-2}fA_1(\nabla A_2+\omega\nabla A_1), \tag{7.3}$$

$$\begin{aligned}\rho^{-1}\tilde{\nabla}\psi_{21}=&-f^{-1}\nabla f-\rho^{-2}f^2\omega\nabla\omega+2f^{-1}A_1\nabla A_1\\&+2\rho^{-2}f(A_2-\omega A_1)(\nabla A_2+\omega\nabla A_1),\end{aligned} \tag{7.4}$$

$$\rho^{-1}\tilde{\nabla}\psi_{12}=\rho^{-1}\tilde{\nabla}\psi_{21}+2\rho^{-1}\nabla\rho, \tag{7.5}$$

$$\begin{aligned}\rho^{-1}\tilde{\nabla}\psi_{22}=&\,2f^{-1}\omega\nabla f+\nabla\omega+\rho^{-2}f^2\omega^2\nabla\omega\\&-2\rho^{-1}\omega\nabla\rho+4f^{-1}A_2\nabla A_1\\&-4\rho^{-2}f\omega A_2(\nabla A_2+\omega\nabla A_1),\end{aligned} \tag{7.6}$$

and the vanishing of the divergence of each of these expressions constitutes the field equations. Only four of them are independent.

Restricting attention to the simplest ones, Eqs. (7.1)–(7.4), we can derive the following:

$$\rho^{-1}\tilde{\nabla}(A_2+\omega A_1)=-f^2(-f\nabla B_1+A_1\nabla\psi_{11}+4A_1{}^2\nabla B_1), \tag{7.7}$$

$$\rho^{-1}\tilde{\nabla}(B_2+\omega B_1)=-f^{-2}(f\nabla A_1+B_1\nabla\psi_{11}+4A_1B_1\nabla B_1), \tag{7.8}$$

$$\rho^{-1}\tilde{\nabla}\omega=-f^{-2}(\nabla\psi_{11}+4A_1\nabla B_1), \tag{7.9}$$

$$\begin{aligned}\rho^{-1}\tilde{\nabla}(\psi_{21}+\omega\psi_{11}+2B_1A_2+2\omega B_1A_1)\\=-f^{-2}(f\nabla f+\psi_{11}\nabla\psi_{11}-2fA_1\nabla A_1\\-2fB_1\nabla B_1+4\psi_{11}A_1\nabla B_1+2A_1B_1\nabla\psi_{11}\\+8A_1A_1B_1\nabla B_1).\end{aligned} \tag{7.10}$$

The vanishing of the divergence of each of *these* expressions constitutes a complete set of field equations for f, ψ_{11}, A_1, B_1.

We can return now to the complex potentials, and let $\Phi=\Phi_1$ and $\mathcal{G}=\mathcal{G}_{11}$ for short. Equations (7.7)–(7.10) become

$$\nabla_3\circ[f^{-2}(f\nabla\Phi-\Phi\{\nabla\mathcal{G}+2\Phi^*\nabla\Phi-\nabla f\})]=0, \tag{7.11}$$

$$\nabla_3\circ[f^{-2}(\nabla\mathcal{G}+2\Phi^*\nabla\Phi-\nabla f)]=0, \tag{7.12}$$

$$\begin{aligned}\nabla_3\circ[f^{-2}(f\nabla\mathcal{G}-if\nabla\Omega\\-i\Omega\{\nabla\mathcal{G}+2\Phi^*\nabla\Phi-\nabla f\})]=0,\end{aligned} \tag{7.13}$$

or equivalently

$$f\nabla^2\Phi=(\nabla\mathcal{G}+2\Phi^*\nabla\Phi)\cdot\nabla\Phi, \tag{7.14}$$

$$f\nabla^2\mathcal{G}=(\nabla\mathcal{G}+2\Phi^*\nabla\Phi)\cdot\nabla\mathcal{G}, \tag{7.15}$$

the form given by Ernst.[13]

Equations (7.14) and (7.15) are invariant under the gravitational and electromagnetic gage transformations,

$$\Phi\to\Phi,\quad \mathcal{G}\to\mathcal{G}+i\alpha, \tag{7.16}$$

where α is real, and

$$\Phi\to\Phi+a,\quad \mathcal{G}\to\mathcal{G}-2a^*\Phi-aa^*, \tag{7.17}$$

where a is complex. [They follow from the arbitrariness of the integration constants in Eqs. (7.1) and (7.3), and the definitions of Φ, $\mathcal{G}$.]

As may be easily verified, the equations are also invariant under a discrete symmetry

$$\Phi\to\mathcal{G}^{-1}\Phi,\quad \mathcal{G}\to\mathcal{G}^{-1}. \tag{7.18}$$

Under this involution, the gage transformations must map into still other symmetry operations, e.g., $\mathcal{G}^{-1}\to\mathcal{G}^{-1}+i\gamma$, or

$$\Phi\to\frac{\Phi}{1+i\gamma\mathcal{G}},\quad \mathcal{G}\to\frac{\mathcal{G}}{1+i\gamma\mathcal{G}} \tag{7.19}$$

and

$$\begin{aligned}\Phi&\to\frac{\Phi+c\mathcal{G}}{1-2c^*\Phi-cc^*\mathcal{G}},\\ \mathcal{G}&\to\frac{\mathcal{G}}{1-2c^*\Phi-cc^*\mathcal{G}}.\end{aligned} \tag{7.20}$$

Equation (7.19) is actually the Ehlers tranformation, but now it has been generalized to allow for the presence of an electromagnetic field. Equation (7.20) is the transformation discovered by Harrison[14] that mixes electromagnetism with gravity.

Neither of them commutes with the gage transformations. Consideration of the commutators leads to another two-parameter symmetry,

$$\Phi\to\beta e^{i\alpha}\Phi,\quad \mathcal{G}\to\beta^2\mathcal{G}, \tag{7.21}$$

where α and β are real. It is a combination of the

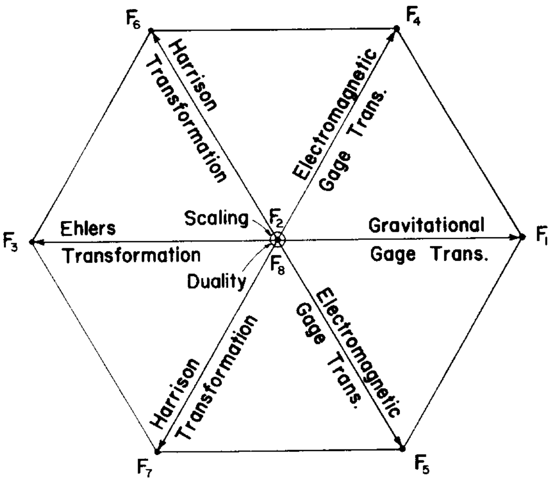


FIG. 1. The generators of H, and their action on F_i.

electromagnetic duality rotation and the rescaling of Eq. (4.15). Further commutators lead to nothing new. The continuous transformations of Eqs. (7.16), (7.17), and (7.19)–(7.21) therefore close on themselves to form a Lie group H′ with eight real parameters.[15] A root diagram for the eight generators of H′ is given in Fig. 1. H′ is isomorphic to SU(2,1), and contains the vacuum group H as an SU(1,1) subgroup.

The Ernst equations, Eqs. (7.14) and (7.15), are *individually* invariant under H′, but the equivalent divergence forms, Eqs. (7.11)–(7.13) are not. The action of H′ on them produces other divergence equations. It is therefore interesting to find all of the equations that can be generated in this way, and to write the entire collection in a form which is manifestly H′ covariant. We can most easily do this by discovering a set of variables on which the action of H′ is linear. It is thus appropriate at this point to digress on some of the representations of SU(2,1).

8. SU(2, 1) FORMULATION

SU(2,1) is defined on a complex 3-space of variables u, v, w. It is the group of unimodular 3×3 matrices which leaves invariant the Hermitian form $uu^* + vv^* - ww^*$. It has a complex three-dimensional spinor representation, typified by the position vector

$$u^\alpha = (u, v, w).$$

We may consider also complex conjugate spinors, such as

$$u^{\dot\alpha} = (u^\alpha)^*$$

and we may raise and lower any index using the invariant Hermitian metric

$$\eta^{\alpha\dot\beta} = \eta_{\alpha\dot\beta} = \mathrm{diag}(1,1,-1). \tag{8.1}$$

Higher rank objects such as $T_{\alpha\dot\beta}{}^{\gamma}$ may be reduced by means of $\eta^{\alpha\dot\beta}$, $\delta^\alpha{}_\beta$, or the alternating symbol $\varepsilon_{\alpha\beta\gamma}$.

The octet representation consists of objects $P^{\alpha\dot\beta}$ which are Hermitian and traceless. It can be written using connecting quantities $\lambda^i{}_{\alpha\dot\beta}$, $i = 1, \ldots, 8$ which generalize the Pauli matrices,

$$P^i = \lambda^i{}_{\alpha\dot\beta} P^{\alpha\dot\beta}. \tag{8.2}$$

The identity analogous to Eq. (3.9),

$$\lambda^i{}_\alpha{}^\beta \lambda^j{}_\beta{}^\gamma = \tfrac{2}{3} g^{ij} \delta_\alpha{}^\gamma + d^{ijk} \lambda_{k\alpha}{}^\gamma + i f^{ijk} \lambda_{k\alpha}{}^\gamma, \tag{8.3}$$

defines the matrices g^{ij}, d^{ijk}, and f^{ijk}. Here g^{ij} is symmetric, d^{ijk} totally symmetric and traceless, and f^{ijk} is totally antisymmetric. Octet indices are raised and lowered by means of g^{ij} and its inverse.

The explicit representation of $\lambda_{i\alpha}{}^\beta$ most convenient for our purposes is

$$\lambda_1 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & -1 \end{pmatrix}, \quad \lambda_2 = \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \lambda_3 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix},$$
$$\lambda_4 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, \quad \lambda_5 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_6 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \tag{8.4}$$
$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_8 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

implying

$$\begin{aligned} &g_{13} = 2, \quad g_{47} = g_{56} = -g_{22} = 1, \quad g_{88} = 3, \\ &d_{138} = 2, \quad d_{167} = d_{345} = -d_{228} = 1, \quad d_{478} = d_{568} = -\tfrac{1}{2}, \\ &\qquad d_{256} = -d_{247} = \tfrac{1}{2} i, \quad d_{888} = -3, \\ &f^{123} = -f^{247} = -f^{256} = \tfrac{1}{2}, \quad f^{167} = f^{345} = f^{478} = -f^{568} = \tfrac{1}{2} i. \end{aligned} \tag{8.5}$$

The linear representation of H′ that we need will be an extension of the vector F_a, Eq. (4.11). Starting with the first component f^{-1}, we let H′ act, and thereby generate an octet of field variables

$$\begin{aligned} &F_1 = f^{-1}, \quad F_5 = f^{-1}\Phi^*, \\ &F_2 = f^{-1}\Omega, \quad F_6 = f^{-1}\Phi\mathcal{G}^*, \\ &F_3 = f^{-1}\mathcal{G}\mathcal{G}^*, \quad F_7 = f^{-1}\Phi^*\mathcal{G}, \\ &F_4 = f^{-1}\Phi, \quad F_8 = f^{-1}(f - 3\Phi\Phi^*). \end{aligned} \tag{8.6}$$

The field equations Eqs. (7.11)–(7.13) are then seen to be only the $i = 2, 3, 6, 7$ components of a covariant set of conservation laws,

$$\nabla_3 \cdot [f^{ijk} F_j \nabla F_k] = 0, \tag{8.7}$$

a direct generalization of Eq. (4.12). We will have an octet of potentials defined by

$$\rho^{-1}\tilde\nabla\Psi^i = -f^{ijk} F_j \nabla F_k. \tag{8.8}$$

Explicitly,

$$\begin{aligned} &\rho^{-1}\tilde\nabla\Psi^1 = f^{-2}[f(g\nabla\Omega - \Omega\nabla g) - \tfrac{1}{2}\mathcal{G}\mathcal{G}^*(\nabla\Omega - J)], \\ &\rho^{-1}\tilde\nabla\Psi^2 = f^{-2}[f\nabla g + \Omega(\nabla\Omega - J)], \\ &\rho^{-1}\tilde\nabla\Psi^3 = -\tfrac{1}{2} f^{-2}[\nabla\Omega - J], \\ &\rho^{-1}\tilde\nabla\Psi^4 = f^{-2}[if(\mathcal{G}\nabla\Phi^* - \Phi^*\nabla\mathcal{G}) \\ &\qquad\qquad - \Phi^*\mathcal{G}(\nabla\Omega - J)], \\ &\rho^{-1}\tilde\nabla\Psi^5 = (\rho^{-1}\tilde\nabla\Psi^4)^*, \\ &\rho^{-1}\tilde\nabla\Psi^6 = f^{-2}[if\nabla\Phi^* - \Phi^*(\nabla\Omega - J)], \\ &\rho^{-1}\tilde\nabla\Psi^7 = (\rho^{-1}\tilde\nabla\Psi^6)^*, \\ &\rho^{-1}\tilde\nabla\Psi^8 = f^{-2}[\Phi\Phi^*\nabla\Omega + gJ], \end{aligned} \tag{8.9}$$

where $g = f - \Phi\Phi^*$, $\Omega = \Omega_{11}$, and

$$J=i(\Phi^*\nabla\Phi-\Phi\nabla\Phi^*). \tag{8.10}$$

The inverse relation,

$$\nabla F^i=4\rho^{-1}f^{ijk}F_j\tilde{\nabla}\Psi_k \tag{8.11}$$

implies field equations for Ψ_i,

$$\nabla_3\circ[\rho^{-2}f^{ijk}F_j\nabla\Psi_k]=0. \tag{8.12}$$

Some interesting covariant relations which can now be derived are:

$$g^{ij}F_iF_j=\tfrac{4}{3}, \tag{8.13}$$

$$g^{ij}F_i\nabla\Psi_j=0, \tag{8.14}$$

$$d^{ijk}F_jF_k=\tfrac{2}{3}F^i, \tag{8.15}$$

$$d^{ijk}F_j\nabla\Psi_k=\tfrac{1}{3}\nabla\Psi^i, \tag{8.16}$$

$$f^{ijk}\nabla\Psi_j\cdot\tilde{\nabla}\Psi_k=\rho^2f^{ijk}\nabla F_j\cdot\tilde{\nabla}F_k. \tag{8.17}$$

The last of these implies the existence of another octet potential,

$$\nabla\Pi^i=f^{ijk}\Psi_j\nabla\Psi_k-\rho^2f^{ijk}F_j\nabla F_k+2z\nabla\Psi^i, \tag{8.18}$$

which is the enlargement of Eq. (5.9).

Comparison of Eqs. (7.7)–(7.10) with Eq. (8.9) shows that we may identify

$$\Psi_1=\omega, \tag{8.19}$$

$$\Psi_2=\Omega_{21}+\omega\Omega_{11}, \tag{8.20}$$

$$\Psi_4=\Psi_5{}^*=\Phi_2+\omega\Phi_1. \tag{8.21}$$

Since we began with just two complex field variables, not all of the eight components of F_i can be functionally independent. Hence the existence of nonlinear constraints like Eqs. (8.13) and (8.15) is not at all surprising. F_i is thus a rather special sort of octet, and this leads us to suspect that it can be expressed in terms of a simpler covariant object. Writing F_i in its equivalent spinor form

$$F_i=\lambda_i{}^{\alpha\dot\beta}F_{\alpha\dot\beta} \tag{8.22}$$

and using the identities

$$g^{ij}\lambda_i{}^{\alpha\dot\beta}\lambda_j{}^{\gamma\dot\delta}=2(\eta^{\alpha\dot\delta}\eta^{\gamma\dot\beta}-\tfrac{1}{3}\eta^{\alpha\dot\beta}\eta^{\gamma\dot\delta}), \tag{8.23}$$

$$\begin{aligned}d^{ijk}\lambda_j{}^{\alpha\dot\beta}\lambda_k{}^{\gamma\dot\delta}&=\eta^{\alpha\dot\delta}\lambda_i{}^{\gamma\dot\beta}+\eta^{\gamma\dot\beta}\lambda_i{}^{\alpha\dot\delta}\\&\quad-\tfrac{2}{3}(\eta^{\alpha\dot\beta}\lambda_i{}^{\gamma\dot\delta}+\eta^{\gamma\dot\delta}\lambda_i{}^{\alpha\dot\beta}),\end{aligned} \tag{8.24}$$

we find that Eqs. (8.13) and (8.15) can be identically satisfied if we assume that $F_{\alpha\dot\beta}$ *factors*,

$$F_{\alpha\dot\beta}=F_\alpha F_{\dot\beta}, \tag{8.25}$$

$$F_\alpha F^\alpha=1. \tag{8.26}$$

The inverse of Eq. (8.22) is then

$$F_\alpha F_{\dot\beta}=\tfrac{1}{2}F_i\lambda^i{}_{\alpha\dot\beta}+\tfrac{1}{3}\eta_{\alpha\dot\beta} \tag{8.27}$$

which enables us to determine F_α up to a phase. Using Eqs. (8.4) and (8.6), we find

$$\begin{aligned}F_1&=\tfrac{1}{2}e^{i\chi}f^{-1/2}(1+\mathcal{G}),\\F_2&=e^{i\chi}f^{-1/2}\Phi,\\F_3&=\tfrac{1}{2}e^{i\chi}f^{-1/2}(1-\mathcal{G}),\end{aligned} \tag{8.28}$$

or[15]

$$\mathcal{G}=\frac{F_1-F_3}{F_1+F_3},\quad \Phi=\frac{F_2}{F_1+F_3}\cdot \tag{8.29}$$

Using the further identity

$$\begin{aligned}&f^{ijk}\lambda_{i\alpha}{}^{\beta}\lambda_{j\gamma}{}^{\delta}\lambda_{k\epsilon}{}^{\zeta}\\&\quad=2i(\delta_\alpha{}^\zeta\delta_\gamma{}^\beta\delta_\epsilon{}^\delta-\delta_\alpha{}^\delta\delta_\gamma{}^\zeta\delta_\epsilon{}^\beta)\end{aligned} \tag{8.30}$$

in Eq. (8.7) gives the field equations also in terms of F_α,

$$\nabla_3\cdot[F_\beta\nabla F_\alpha-F_\alpha\nabla F_\beta+2F_\alpha F_\beta(F_\gamma\nabla F^\gamma)]=0. \tag{8.31}$$

Unfortunately the determination (and even the existence) of the required phase remains problematic.[16]

Finally, we can extend the results of Sec. 5 to the Einstein–Maxwell case. Under an infinitesimal Ehlers transformation, the octet experiences

$$\begin{aligned}&F_1\to F_1-2\gamma F_2,\quad F_2\to F_2-\gamma F_3,\\&F_4\to F_4-i\gamma F_6,\quad F_5\to F_5+i\gamma F_7,\end{aligned} \tag{8.32}$$

and F_3, F_6, F_7, F_8 remain unchanged. Similarly for Ψ_i. From this we can deduce

$$\begin{aligned}&f\to f+2\gamma f\Omega,\\&\omega\to\omega-2\gamma(\Omega_{21}+\omega\Omega),\\&\Omega\to\Omega-\gamma(\mathcal{G}\mathcal{G}^*-2\Omega^2),\\&\Omega_{21}\to\Omega_{21}-\gamma(\Psi_3-\omega\mathcal{G}\mathcal{G}^*-2\Omega\Omega_{21}),\\&\Phi\to\Phi-i\gamma\Phi\mathcal{G},\\&\Phi_2\to\Phi_2+\gamma(-i\Psi_6+2\Phi\Omega_{21}+i\omega\Phi\mathcal{G}^*).\end{aligned}$$

This leads at once to the **G**-covariant generalization

$$f_{AB}\to f_{AB}+2(\gamma_{CD}f^{CD}\Omega_{(AB)}-\gamma_{AB}f^{CD}\Omega_{CD}). \tag{8.33}$$

To obtain the transformation of the remaining components of Ω_{AB} we must appeal to the invariance of Eqs. (6.15) and (6.19). We find that new potentials are required, all quadratic in the field variables:

$$\begin{aligned}&\nabla K=\Phi_C{}^*\nabla\Phi^C,\quad \nabla L_B=\Phi_C{}^*\nabla\mathcal{H}^C{}_B,\\&\nabla M_A=\mathcal{H}^*{}_{CA}\nabla\Phi^C,\quad \nabla N_{AB}=\mathcal{H}^*{}_{CA}\nabla\mathcal{H}^C{}_B.\end{aligned} \tag{8.34}$$

[The integrability conditions $\nabla_3\cdot(\rho^{-1}\tilde{\nabla}K)=0$, etc. are easily verified.] The results are

$$\begin{aligned}\mathcal{H}_{AB}\to\mathcal{H}_{AB}&-i\gamma^{XY}\mathcal{H}_{AX}\mathcal{H}_{YB}-i\gamma_A{}^YN_{YB}\\&-i\gamma^X{}_B(N_{AX}+2\Phi_AL_X+\mathcal{H}_{AY}\mathcal{H}^Y{}_X),\end{aligned} \tag{8.35}$$

$$\Phi_A\to\Phi_A-i\gamma^{BC}\Phi_B\mathcal{G}_{AC}-i\gamma_A{}^CL_C. \tag{8.36}$$

Under an infinitesimal Harrison transformation

$$\begin{aligned}&F_1\to F_1-2c^*F_4+\text{c.c.},\\&F_2\to F_2-ic^*F_6+\text{c.c.},\\&F_4\to F_4+cF_8+icF_2,\\&F_6\to F_6+cF_3,\\&F_8\to F_8-3c^*F_6+\text{c.c.},\end{aligned} \tag{8.37}$$

where c.c. denotes a corresponding infinitesimal term in c. F_3 remains unchanged, and F_5 and F_7 transform like $F_4{}^*$ and $F_6{}^*$. Hence

$$\begin{aligned}&f\to f+2c^*f\Phi+\text{c.c.},\\&\omega\to\omega-2c^*(\Phi_2+\omega\Phi_1)+\text{c.c.},\end{aligned}$$

$$\Omega \to \Omega - ic^*\Phi\mathcal{G} + \text{c.c.},$$

$$\Omega_{21} \to \Omega_{21} - ic^*(\Psi_6 + 2i\Omega\Phi_2 - \omega\Phi\mathcal{G}^*),$$

$$\Phi \to \Phi + 2c^*\Phi\Phi + c\mathcal{G},$$

$$\Phi_2 \to \Phi_2 + 2c^*\Phi\Phi_2 + c(\Psi_8 + i\Omega_{21} - \omega\mathcal{G} + 2\Phi\Phi_2{}^* + 2\omega\Phi\Phi^*),$$

leading to the generalization

$$f_{AB} \to f_{AB} + 2(c^{*C}f_{CB}\Phi_A + c^*{}_B f_{AC}\Phi^C) + \text{c.c.} \tag{8.38}$$

Invariance of Eqs. (6.15) and (6.19) implies

$$\mathcal{H}_{AB} \to \mathcal{H}_{AB} + 2c^{*C}\mathcal{H}_{CB}\Phi_A - 2c_A L_B + 2c^*{}_B(M_A + 2\Phi_A K + \mathcal{H}_{AX}\Phi^X), \tag{8.39}$$

$$\Phi_A \to \Phi_A + 2c^{*C}\Phi_C\Phi_A + c^C\mathcal{H}_{AC} - 2c_A K. \tag{8.40}$$

*Supported in part by National Science Foundation Grant PHY76-12246.

[1] Nonlinear transformations would produce t or φ derivatives.

[2] We are considering here only the local properties of the spacetime as determined by its line element ds^2, and ignoring all questions of a global nature, such as the periodicity of the φ coordinate.

[3] We are taking for the Levi-Civita symbols the values $\varepsilon^{abc} = \varepsilon_{abc} = \pm 1$, thereby avoiding many appearances of $\sqrt{2}$. The distinction between tensors and tensor densities is irrelevant here anyway, since the allowed transformations all have determinant unity. This notation implies that $\varepsilon^{abc}G_{ad}G_{be}G_{cf} = \frac{1}{2}\varepsilon_{def}$.

[4] Equation (3.10) may be identically satisfied by imposing the canonical coordinate condition $x^3 = \rho$. However we shall not make this restriction.

[5] T. Lewis, Proc. Roy. Soc. (London) **A136**, 176 (1932); A. Papapetrou, Ann. Phys. **12**, 309 (1953); F.J. Ernst, Phys. Rev. **167**, 1175 (1968).

[6] G. Neugebauer and D. Kramer, Ann. Phys. **24**, 62 (1969).

[7] Two interpretations to this rescaling are possible, depending on how we treat the coordinates. If we let $t \to t$, $\varphi \to \beta\varphi$, then $ds^2 \to \beta ds^2$, which corresponds to a uniform conformal transformation of the spacetime. On the other hand, if $t \to \beta^{-1/2}t$, $\varphi \to \beta^{1/2}\varphi$, we produce instead a Lorentz rotation in the (t, φ) plane. When G, H are being considered independently, the first interpretation is the more natural. The second will be used in the context of the combined group K.

[8] J. Ehlers, in *Les Theories Relativistes de la Gravitation* (CNRS, Paris, 1959).

[9] H is a group of rotations in an internal Minkowski 3-space. In addition, we may consider the reflections. There are two discrete symmetries corresponding to "space" and "time" reflection, $f \to -f$ and $\psi \to -\psi$. Other symmetries may be constructed by combining rotations and reflections, but should not be regarded as independent. For example, one of this type is $\mathcal{E} \to \mathcal{E}^{-1}$. Its electromagnetic generalization is given in Eq. (7.18).

[10] A. Barut and C. Fronsdal, Proc. R. Soc. London **A287**, 532 (1965).

[11] R. Geroch, J. Math. Phys. **12**, 918 (1971).

[12] R. Geroch, J. Math. Phys. **13**, 394 (1972).

[13] F.J. Ernst, Phys. Rev. **168**, 1415 (1968).

[14] B.K. Harrison, J. Math. Phys. **9**, 1744 (1968).

[15] W. Kinnersley, J. Math. Phys. **14**, 651 (1973).

[16] Without χ, the three given expressions for F_i would fail to transform like a 3-spinor. One can show that χ must be gage invariant, and that under infinitesimal Ehlers and Harrison transformations

$$i\chi \to i\chi + \tfrac{1}{2}i\gamma(\mathcal{G} + \mathcal{G}^*),$$

$$i\chi \to i\chi - c^*\Phi - c\Phi^*.$$

No combination of the basic field variables behaves in this manner.

Symmetries of the stationary Einstein–Maxwell field equations. II*

William Kinnersley and D. M. Chitre

Department of Physics, Montana State University, Bozeman, Montana 59715

(Received 7 February 1977)

From Einstein–Maxwell fields which are stationary and axially symmetric, we show how to construct an infinite hierarchy of potentials. The potentials form a representation of K′, the infinite-parameter symmetry group of the Einstein–Maxwell equations. For flat space, the hierarchy is calculated explicitly.

1. INTRODUCTION

In a previous paper[1] (referred to as I), we investigated the Einstein–Maxwell field equations for spacetimes which are stationary and axially symmetric. Since we could not solve the equations, we attempted to learn as much as possible about them through their symmetry group.

Considering first only the timelike Killing vector, we found[2] a symmetry group $\mathbf{H}' \cong SU(2,1)$. $\mathbf{H}'$ is generated by Ehlers transformations[3] and Harrison transformations[4] along with several gauge transformations. Considering both Killing vectors simultaneously, we found that $\mathbf{H}'$ was the beginning of an infinite parameter symmetry group $\mathbf{K}'$. The vacuum subgroup $\mathbf{K} \subset \mathbf{K}'$ had been discovered previously by Geroch.[5]

In the present paper we will examine in detail the structure of $\mathbf{K}'$. Using the electromagnetic and gravitational fields, we show how to construct an infinite hierarchy of potentials. The action of $\mathbf{K}'$ on these potentials is then analyzed. In the last section we calculate the potentials explicitly for the simplest case possible: flat space.

First we must summarize the notation and results of I. We assume the metric

$$\begin{aligned} &ds^2 = f_{AB}\, dx^A\, dx^B - e^{2\Gamma}\delta_{MN}\, dx^M\, dx^N, \\ &A,\ B = 1, 2, \quad M, N = 3, 4, \end{aligned} \tag{1.1}$$

where f_{AB}, Γ are functions of x^3, x^4.

We have[6]

$$f^{AX} f_{XB} = -\rho^2 \delta^A c, \tag{1.2}$$

where indices are raised and lowered using $\epsilon_{AB} = \pm 1$. We introduce the two-dimensional gradient operators

$$\nabla = (\partial_3, \partial_4), \quad \tilde{\nabla} = (\partial_4, -\partial_3). \tag{1.3}$$

For any 2-vector $\mathbf{V}$, the equation $\nabla \cdot \mathbf{V} = 0$ implies the existence of a scalar potential U:

$$\mathbf{V} = \tilde{\nabla} U. \tag{1.4}$$

The Maxwell equations are written

$$\nabla \cdot (\rho^{-1} f^{AX} \nabla A_X) = 0, \tag{1.5}$$

where A_A are components of the usual electromagnetic 4-potential. Using Eq. (1.4), we define another potential B_A:

$$\begin{aligned} &\tilde{\nabla} B_A = \rho^{-1} f_A{}^X \nabla A_X \\ &\Longleftrightarrow \tilde{\nabla} A_A = -\rho^{-1} f_A{}^X \nabla B_X. \end{aligned} \tag{1.6}$$

The complex combination

$$\varphi_A = A_A + iB_A \tag{1.7}$$

enables us to write the Maxwell equations in the simple form

$$\nabla \varphi_A = -i\rho^{-1} f_A{}^X \tilde{\nabla} \varphi_X. \tag{1.8}$$

Our next objective is to establish a remarkable fact: The Einstein–Maxwell equations

$$R^A{}_B = -8\pi T^A{}_B \tag{1.9}$$

can be written in a form which is virtually identical to Eq. (1.8), differing only in the presence of an extra index. The equations are

$$\begin{aligned} &\nabla \cdot [\rho^{-1}(f^{AX} \nabla f_{XB} - 2f^{AX} A_B \nabla A_X \\ &\quad - 2f_B{}^X A^A \nabla A_X)] = 0. \end{aligned} \tag{1.10}$$

Using Eq. (1.4), we define ψ_{AB} such that $\tilde{\nabla}\psi_{AB}$ is equal to the bracket above. We also define

$$\Omega_{AB} = \psi_{AB} + 2A_A B_B. \tag{1.11}$$

Then

$$\begin{aligned} &\nabla f_{AB} = \rho^{-1} f_A{}^X (\tilde{\nabla}\Omega_{XB} + 2A_B \tilde{\nabla} B_X - 2B_B \tilde{\nabla} A_X), \\ &\nabla \Omega_{AB} = -\rho^{-1} f_A{}^X (\nabla f_{XB} - 2A_B \tilde{\nabla} A_X - 2B_B \tilde{\nabla} B_X). \end{aligned} \tag{1.12}$$

The complex combination

$$\mathcal{E}_{AB} = f_{AB} + i\Omega_{AB} \tag{1.13}$$

satisfies

$$\nabla \mathcal{E}_{AB} = -i\rho^{-1} f_A{}^X (\tilde{\nabla} \mathcal{E}_{XB} - 2\varphi_B \tilde{\nabla} \varphi_X{}^*) \tag{1.14}$$

while

$$\mathcal{G}_{AB} = \mathcal{E}_{AB} - \varphi_A^* \varphi_B \tag{1.15}$$

satisfies

$$(\nabla \mathcal{G}_{AB} + \varphi_A{}^* \nabla \varphi_B) = -i\rho^{-1} f_A{}^X (\tilde{\nabla} \mathcal{G}_{XB} + \varphi_X{}^* \tilde{\nabla} \varphi_B). \tag{1.16}$$

At this point we make use of a quadratic potential K, defined by

$$\nabla K = \varphi_X^* \nabla \varphi^X. \tag{1.17}$$

To show that Eq. (1.17) does define a scalar K, we need to verify the integrability condition

$$\nabla\cdot(\varphi_X{}^*\tilde{\nabla}\varphi^X)=0 \iff \nabla\varphi_X{}^*\cdot\tilde{\nabla}\varphi^X=0.$$

The lhs is manifestly real; but using Eq. (1.8) and the symmetry of f_{AB}, we can show that it must be imaginary. Hence it vanishes. The quantity

$$\mathcal{H}_{AB}=\mathcal{G}_{AB}+\epsilon_{AB}K \tag{1.18}$$

casts the Einstein–Maxwell equations into the desired form:

$$\nabla\mathcal{H}_{AB}=-i\rho^{-1}f_A{}^X\tilde{\nabla}\mathcal{H}_{XB}. \tag{1.19}$$

We take φ_A, $\mathcal{H}_{AB}$ as the basic potentials for the electromagnetic and gravitational fields, and Eqs. (1.8), (1.19), as our basic field equations.

2. THE HIERARCHY

We now notice that there are actually four expressions quadratic in the field variables, which lead to new potentials:

$$\nabla K=\varphi_X{}^*\nabla\varphi^X, \tag{2.1}$$

$$\nabla L_B=\varphi_X{}^*\nabla\mathcal{H}^X{}_B, \tag{2.2}$$

$$\nabla M_A=\mathcal{H}^*{}_{XA}\nabla\varphi^X, \tag{2.3}$$

$$\nabla N_{AB}=\mathcal{H}^*{}_{XA}\nabla\mathcal{H}^X{}_B. \tag{2.4}$$

The integrability conditions necessary for the existence of these potentials may be easily verified. The proofs follow the pattern we used for K; they hinge only on the symmetry of f_{AB} and the fact that φ_A, $\mathcal{H}_{AB}$ satisfy Eqs. (1.8), (1.19).[7]

The next step is to show how these "potentials" can be used to construct more "fields," i.e., solutions of Eqs. (1.8), (1.19). Consider the following combination:

$$R_A=M_A+2K\varphi_A+\mathcal{H}_{AX}\varphi^X, \tag{2.5}$$

$$\begin{aligned}\Rightarrow \nabla R_A&=(\mathcal{H}_{AX}+\mathcal{H}^*_{XA}+2\varphi_A\varphi_X{}^*)\nabla\varphi^X\\&\quad+2K\nabla\varphi_A+\varphi^X\nabla\mathcal{H}_{AX}\\&=2f_{AX}\nabla\varphi^X+(3K-K^*+i\Omega^X{}_X)\nabla\varphi_A\\&\quad+\varphi^X\nabla\mathcal{H}_{AX}.\end{aligned}$$

Multiplying by f^{AB}, and using Eq. (1.2), we find that R_A does satisfy Eq. (1.8). Similarly

$$P_{AB}=N_{AB}+2\varphi_A L_B+\mathcal{H}_{AX}\mathcal{H}^X{}_B \tag{2.6}$$

may be shown to satisfy Eq. (1.19).

Now *these* fields may also be used in any combination to construct *cubic* potentials, e.g.,

$$\nabla S=R_X{}^*\nabla\varphi^X,\quad \nabla T=\varphi_X{}^*\nabla R^X,\quad \text{etc.}$$

In fact it is now abundantly clear that an infinite hierarchy of fields and potentials is involved, and we turn to the complete description.

Suppose we have an infinite sequence of fields $\overset{n}{\varphi}_A$, $\overset{n}{\mathcal{H}}_{AB}$, $n=1,2,\ldots$, with

$$\overset{1}{\varphi}_A=\varphi_A,\quad \overset{1}{\mathcal{H}}_{AB}=\mathcal{H}_{AB}, \tag{2.7}$$

all obeying the same field equations:

$$\nabla\overset{n}{\varphi}_A=-i\rho^{-1}f_A{}^X\tilde{\nabla}\overset{n}{\varphi}_X, \tag{2.8}$$

$$\nabla\overset{n}{\mathcal{H}}_{AB}=-i\rho^{-1}f_A{}^X\tilde{\nabla}\overset{n}{\mathcal{H}}_{XB}. \tag{2.9}$$

From the fields we define four families of potentials:

$$\nabla\overset{mn}{K}=\overset{m}{\varphi}_X{}^*\nabla\overset{n}{\varphi}{}^X, \tag{2.10}$$

$$\nabla\overset{mn}{L}_B=\overset{m}{\varphi}_X{}^*\nabla\overset{n}{\mathcal{H}}{}^X{}_B, \tag{2.11}$$

$$\nabla\overset{mn}{M}_A=\overset{m}{\mathcal{H}}{}^*_{XA}\nabla\overset{n}{\varphi}{}^X, \tag{2.12}$$

$$\nabla\overset{mn}{N}_{AB}=\overset{m}{\mathcal{H}}{}^*_{XA}\nabla\overset{n}{\mathcal{H}}{}^X{}_B. \tag{2.13}$$

Conversely, from the potentials we construct solutions of the field equations,

$$\overset{n+1}{\varphi}_A=i(\overset{1n}{M}_A+2\varphi_A\overset{1n}{K}+\mathcal{H}_{AX}\overset{n}{\varphi}{}^X), \tag{2.14}$$

$$\overset{n+1}{\mathcal{H}}_{AB}=i(\overset{1n}{N}_{AB}+2\varphi_A\overset{1n}{L}_B+\mathcal{H}_{AX}\overset{n}{\mathcal{H}}{}^X{}_B), \tag{2.15}$$

thereby providing a recursive definition of the original fields.

Several properties of the potentials follow quite easily from their definition. Integration by parts on Eqs. (2.10), (2.13) leads to

$$\overset{mn}{K}-\overset{nm}{K}{}^*=\overset{m}{\varphi}_X{}^*\overset{n}{\varphi}{}^X, \tag{2.16}$$

$$\overset{mn}{L}_B-\overset{nm}{M}{}^*_A=\overset{m}{\varphi}{}^*_X\overset{n}{\mathcal{H}}{}^X{}_B, \tag{2.17}$$

$$\overset{mn}{N}_{AB}-\overset{nm}{N}{}^*_{BA}=\overset{m}{\mathcal{H}}{}^*_{XA}\overset{n}{\mathcal{H}}{}^X{}_B. \tag{2.18}$$

Also, we can obtain recursion relations between adjacent potentials by inserting Eqs. (2.14), (2.15) into Eqs. (2.10)–(2.13):

$$\overset{m,n+1}{K}-\overset{m+1,n}{K}=2i\overset{m1}{K}\overset{1n}{K}+i\overset{m1}{L}_X\overset{n}{\varphi}{}^X, \tag{2.19}$$

$$\overset{m,n+1}{L}_B-\overset{m+1,n}{L}_B=2i\overset{m1}{K}\overset{1n}{L}_B+i\overset{m1}{L}_X\overset{n}{\mathcal{H}}{}^X{}_B, \tag{2.20}$$

$$\overset{m,n+1}{M}_A-\overset{m+1,n}{M}_A=2i\overset{m1}{M}_A\overset{1n}{K}+i\overset{m1}{N}_{AX}\overset{n}{\varphi}{}^X, \tag{2.21}$$

$$\overset{m,n+1}{N}_{AB}-\overset{m+1,n}{N}_{AB}=2i\overset{m1}{M}_A\overset{1n}{L}_B+i\overset{m1}{N}_{AX}\overset{n}{\mathcal{H}}{}^X{}_B. \tag{2.22}$$

To this point, we have, of course, been using the restriction $m,n\geqslant 1$. However, for some purposes it is useful and quite natural to define an extended set of potentials, in which the integers m,n may also be zero or negative.

Equations (2.14), (2.15) will hold for $n=0$ provided we define

$$\overset{10}{K}=-\tfrac{1}{2}i, \tag{2.23}$$

$$\overset{0}{\mathcal{H}}_{AB}=i\epsilon_{AB}. \tag{2.24}$$

But, then, from Eqs. (2.12), (2.13) we may identify

$$\overset{0n}{M}_A = -i\overset{n}{\varphi}_A, \tag{2.25}$$

$$\overset{0n}{N}_{AB} = -i\overset{n}{\mathcal{H}}_{AB}. \tag{2.26}$$

Thus both fields and potentials become incorporated into the same extended hierarchy, leading to a considerable simplification. Equations (2.14), (2.15) themselves are now nothing more than special cases of Eqs. (2.21), (2.22). Equations (2.16)–(2.22) will hold for *all* values of m,n, provided we define

$$\overset{1-p,p}{K} = -\overset{p,1-p}{K} = \tfrac{1}{2}i, \tag{2.27}$$

$$\overset{p,-p}{N}_{AB} = -\overset{-p,p}{N}_{AB} = \epsilon_{AB}, \tag{2.28}$$

where $p \geqslant 1$. All other quantities with $m,n \leqslant 0$ are assumed to vanish.

If desired, Eqs. (2.19)–(2.22) may now be written in an even simpler form, in which only quadratic terms appear. For example, Eq. (2.19) is equivalent to

$$0 = \sum_s (2i\overset{ms}{K}\,\overset{2-s,n}{K} - \overset{ms}{L}_X\,\overset{1-s,n}{M}{}^X),$$

where s is summed over all integer values, positive and negative.

3. THE GROUP K′

The infinitesimal transformations γ_{AB} and c_A discussed in I are also the beginning of an infinite family of transformations comprising the symmetry group **K′**. The other ones may be constructed by repeatedly forming all possible commutators, until closure is obtained. It turns out in this way that there are *three* classes of transformations in **K′**: $\overset{k}{\gamma}_{AB}$, $\overset{k}{c}_A$, and $\overset{k}{\sigma}$, where k is any integer, $-\infty < k < +\infty$. The infinitesimal parameters $\overset{k}{\gamma}_{AB}$ are real and symmetric, $\overset{k}{c}_A$ are complex, and $\overset{k}{\sigma}$ are real.[8] The corresponding action they produce on the hierarchy is as follows:

$$\begin{aligned}
\overset{k}{\gamma}_{AB}: \overset{mn}{K} &\to \overset{mn}{K} + \overset{k}{\gamma}{}^{XY}(\sum_s \overset{ms}{L}_X\,\overset{k-s,n}{M}_Y),\\
\overset{mn}{L}_B &\to \overset{mn}{L}_B + \overset{k}{\gamma}_{XB}\,\overset{m,n+k}{L}{}^X + \overset{k}{\gamma}{}^{XY}(\sum_s \overset{ms}{L}_X\,\overset{k-s,n}{N}_{YB}),\\
\overset{mn}{M}_A &\to \overset{mn}{M}_A + \overset{k}{\gamma}_{AX}\,\overset{m+k,n}{M}{}^X + \overset{k}{\gamma}{}^{XY}(\sum_s \overset{ms}{N}_{AX}\,\overset{k-s,n}{M}_Y),\\
\overset{mn}{N}_{AB} &\to \overset{mn}{N}_{AB} + \overset{k}{\gamma}_{AX}\,\overset{m+k,n}{N}{}^X{}_B + \overset{k}{\gamma}_{XB}\,\overset{m,n+k}{N}{}_A{}^X\\
&\quad + \overset{k}{\gamma}{}^{XY}(\sum_s \overset{ms}{N}_{AX}\,\overset{k-s,n}{N}_{YB});
\end{aligned} \tag{3.1}$$

$$\begin{aligned}
\overset{k}{c}_A: \overset{mn}{K} &\to \overset{mn}{K} + \overset{k}{c}{}^{*X}[\overset{m+k-1,n}{M}_X + 2i(\sum_s \overset{ms}{K}\,\overset{k-s,n}{M}_X)]\\
&\quad + \overset{k}{c}{}^X[\overset{m,n+k-1}{L}_X - 2i(\sum_s \overset{ms}{L}_X\,\overset{k-s,n}{K})],\\
\overset{mn}{L}_B &\to \overset{mn}{L}_B + \overset{k}{c}{}^{*X}[\overset{m+k-1,n}{N}_{XB} + 2i(\sum_s \overset{ms}{K}\,\overset{k-s,n}{N}_{XB})]\\
&\quad - 2i\overset{k}{c}{}^*{}_B\,\overset{m,n+k}{K} + \overset{k}{c}{}^X[-2i(\sum_s \overset{ms}{L}_X\,\overset{k-s,n}{L}_B)],\\
\overset{mn}{M}_A &\to \overset{mn}{M}_A + \overset{k}{c}{}^{*X}[2i(\sum_s \overset{ms}{M}_A\,\overset{k-s,n}{M}_X)] + 2i\overset{k}{c}_A\,\overset{m+k,n}{K}\\
&\quad + \overset{k}{c}{}^X[\overset{m,n+k-1}{N}_{AX} - 2i(\sum_s \overset{ms}{N}_{AX}\,\overset{k-s,n}{K})],\\
\overset{mn}{N}_{AB} &\to \overset{mn}{N}_{AB} + \overset{k}{c}{}^{*X}[2i(\sum_s \overset{ms}{M}_A\,\overset{k-s,n}{N}_{XB})] - 2i\overset{k}{c}{}^*{}_B\,\overset{m,n+k}{M}_A\\
&\quad + \overset{k}{c}{}^X[-2i(\sum_s \overset{ms}{N}_{AX}\,\overset{k-s,n}{L}_b)] + 2i\overset{k}{c}_A\,\overset{m+k,n}{L}_B;
\end{aligned} \tag{3.2}$$

$$\begin{aligned}
\overset{k}{\sigma}: \overset{mn}{K} &\to \overset{mn}{K} + i\overset{k}{\sigma}\,\overset{m+k,n}{K} - i\overset{k}{\sigma}\,\overset{m,n+k}{K} - 2\overset{k}{\sigma}(\sum_s \overset{ms}{K}\,\overset{k-s+1,n}{K}),\\
\overset{mn}{L}_B &\to \overset{mn}{L}_B + i\overset{k}{\sigma}\,\overset{m+k,n}{L}_B - 2\overset{k}{\sigma}(\sum_s \overset{ms}{K}\,\overset{k-s+1,n}{L}_B),\\
\overset{mn}{M}_A &\to \overset{mn}{M}_A - i\overset{k}{\sigma}\,\overset{m,n+k}{M}_A - 2\overset{k}{\sigma}(\sum_s \overset{ms}{M}_A\,\overset{k-s+1,n}{K}),\\
\overset{mn}{N}_{AB} &\to \overset{mn}{N}_{AB} - 2\overset{k}{\sigma}(\sum_s \overset{ms}{M}_A\,\overset{k-s+1,n}{L}_B).
\end{aligned} \tag{3.3}$$

The action produced on the fields follows automatically from Eqs. (2.25), (2.26). All of the above summations are assumed to run only over positive values of s. This has been done in order to display explicitly those terms which are linear. If negative values of s are included, the linear terms may be absorbed into the sums, making the entire transformation formally quadratic. On the other hand, for $k \leqslant 0$, all of the terms which are genuinely quadratic vanish, leaving linear transformations. What we have to deal with is a *nonlinear realization of an infinite parameter Lie algebra.*[9]

The commutators are

$$\begin{aligned}
&[\overset{k}{\sigma},\overset{l}{\sigma}] = 0, \quad [\overset{k}{\sigma},\overset{l}{\gamma}] = 0,\\
&[\overset{k}{\sigma},\overset{l}{c}] = \overset{k+l}{c}, \qquad \overset{k+l}{c}_A = 3i\overset{k}{\sigma}\,\overset{l}{c}_A,\\
&[\overset{k}{\gamma},\overset{l}{\gamma}] = \overset{k+l}{\gamma}, \qquad \overset{k+1}{\gamma}_{AB} = 2\overset{k}{\gamma}{}^X{}_{(A}\overset{l}{\gamma}_{B)X},\\
&[\overset{k}{c},\overset{l}{\gamma}] = \overset{k+l}{c}, \qquad \overset{k+l}{c}_A = \overset{k}{c}{}^X\overset{l}{\gamma}_{AX},\\
&[\overset{k}{c},\overset{l}{c}] = \overset{k+l-1}{\gamma} + \overset{k+l-1}{\sigma}, \qquad \overset{k+l-1}{\gamma}_{AB} = 2i(\overset{k}{c}_{(A}\overset{l}{c}{}^*_{B)} - \overset{k}{c}{}^*_{(A}\overset{l}{c}_{B)}),\\
&\qquad\qquad\qquad\qquad\qquad\qquad \overset{k+l-1}{\sigma} = \overset{k}{c}{}^X\overset{l}{c}{}^*{}_X + \overset{k}{c}{}^{*X}\overset{l}{c}_X.
\end{aligned} \tag{3.4}$$

A root diagram for **K′** is given in Fig. 1. It is seen to be a natural enlargement of the SU(2,1) diagram for **H′** given in I. In fact, the transformations of **K′** immediately surrounding the origin are precisely those special cases we have met before: $\overset{1}{c}_{A_0}$ and $\overset{1}{\gamma}_{AB}$ are the Harrison and Ehlers transformations; c_A and $\overset{-1}{\gamma}_{AB}$ are the electromagnetic and gravitational gauge transformations; $\overset{0}{\sigma}$ is the electromagnetic duality rotations; and $\overset{0}{\gamma}_{AB}$ is an infinitesimal rotation in the (x^1,x^2) plane.[10]

For many purposes we would be interested primarily in vacuum fields. The potentials $\overset{mn}{K}$, $\overset{mn}{L}_B$, $\overset{mn}{M}_A$ as well as the fields $\overset{n}{\varphi}_A$ may then be discarded, leaving only $\overset{mn}{N}_{AB}$ and $\overset{n}{\mathcal{H}}_{AB}$. The subgroup $\mathbf{K} \subset \mathbf{K}'$ which preserves vacuum consists of the $\overset{k}{\gamma}_{AB}$ alone.

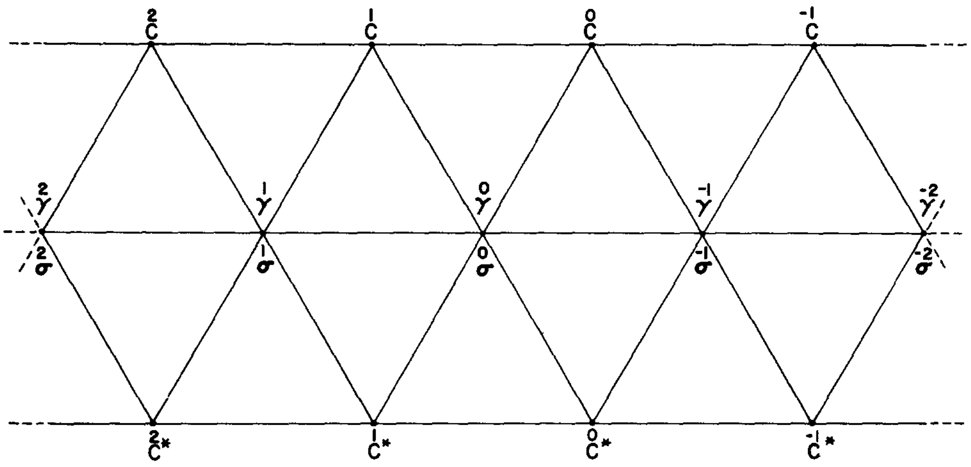


FIG. 1. Root diagram for **K′**. The figure extends indefinitely to the left and right.

Our inclusion of electromagnetism throughout this work has been an enormous help rather than a hindrance. It has revealed a striking interrelationship between electromagnetic and gravitational fields that could not possibly have been anticipated. Moreover, the Maxwell equations and Harrison transformations, being simpler than their gravitational couterparts, have served as a useful guide.

4. FLAT SPACE

As an illustration, and as an essential first step in understanding the potentials, we would like to construct $\overset{mn}{N}_{AB}$ explicitly for the simple case of flat space, in cylindrical coordinates. They will be functions of the coordinates ρ and z.

The nonvanishing quantities $\overset{n}{\mathcal{H}}_{AB}$ and $\overset{mn}{N}_{AB}$ are related by

$$\nabla\overset{mn}{N}_{AB}=\overset{m}{\mathcal{H}}{}^*{}_{XA}\nabla\overset{n}{\mathcal{H}}{}^X{}_B, \tag{4.1}$$

$$\overset{n+1}{\mathcal{H}}{}_{AB}=i(\overset{1n}{N}_{AB}+\mathcal{H}_{AX}\overset{n}{\mathcal{H}}{}^X{}_B). \tag{4.2}$$

In differential form

$$\begin{aligned}\nabla\overset{n+1}{\mathcal{H}}{}_{AB}&=i(\mathcal{H}_{AX}+\mathcal{H}^*{}_{XA})\nabla\overset{n}{\mathcal{H}}{}^X{}_B+i\overset{n}{\mathcal{H}}{}^X{}_B\nabla\mathcal{H}_{AX}.\\&=-2if_A{}^X\,\nabla\overset{n}{\mathcal{H}}_{AX}+\psi^X{}_X\nabla\overset{n}{\mathcal{H}}_{AB}+i\overset{n}{\mathcal{H}}{}^X{}_B\nabla\mathcal{H}_{AX}.\end{aligned} \tag{4.3}$$

For flat space,

$$f_{AB}=\begin{pmatrix}1&0\\0&-\rho^2\end{pmatrix},\quad \psi_{AB}=\begin{pmatrix}0&2z\\0&0\end{pmatrix}. \tag{4.4}$$

Inserting these into Eq. (4.3) with $A=1$ leads to an equation which can be immediately integrated, giving

$$\overset{n+1}{\mathcal{H}}{}_{1B}=2i\overset{n}{\mathcal{H}}_{2B}+2z\overset{n}{\mathcal{H}}_{1B}. \tag{4.5}$$

For $A=2$, we get

$$\nabla\overset{n+1}{\mathcal{H}}{}_{2B}=2i\rho^2\nabla\overset{n}{\mathcal{H}}_{1B}+2z\nabla\overset{n}{\mathcal{H}}_{2B}+2i\rho\overset{n}{\mathcal{H}}_{1B}\nabla\rho. \tag{4.6}$$

We try to satisfy Eqs. (4.5), (4.6) by means of the following ansatz:

$$\begin{aligned}&\overset{n}{\mathcal{H}}_{11}=\rho^{n-1}a_{n-1}(x),\quad \overset{n}{\mathcal{H}}_{21}=i\rho^n b_n(x),\\&\overset{n}{\mathcal{H}}_{12}=i\rho^n a_n(x),\qquad \overset{n}{\mathcal{H}}_{22}=-\rho^{n+1}b_{n+1}(x),\end{aligned} \tag{4.7}$$

where

$$x=z/\rho. \tag{4.8}$$

From Eqs. (4.5), (4.6) we obtain necessary recursion relations among a_n, b_n:

$$a_n=-2b_n+2xa_{n-1}, \tag{4.9}$$

$$(n+1)\,b_{n+1}=2na_{n-1}+2nxb_n, \tag{4.10}$$

$$b'_{n+1}=2a'_{n-1}+2xb'_n. \tag{4.11}$$

Further manipulation leads to more recursion relations and a differential equation:

$$a'_n=2na_{n-1}, \tag{4.12}$$

$$b'_n=2(n-1)\,b_{n-1}, \tag{4.13}$$

$$2n(1+x^2)\,b_{n+1}=(n+1)xb_{n+1}-na_n, \tag{4.14}$$

$$(1+x^2)a''_n-(2n-1)xa'_n+n^2a_n=0. \tag{4.15}$$

The latter serves to identify a_n, b_n as "modified Gegenbauer polynomials"[11]

$$a_n(x)=i^nC_n^{(-n)}(ix), \tag{4.16}$$

$$b_n(x)=-i^nC_n^{(-n+1)}(ix). \tag{4.17}$$

Despite the contrary appearance, a_n and b_n are both real. They are polynomials of degree n and $n-2$ respectively.

The next step is to calculate $\overset{mn}{N}_{AB}$ from Eq. (4.1). Our ansatz has made $\overset{n}{\mathcal{H}}_{11}$, $\overset{n}{\mathcal{H}}_{22}$ real and $\overset{n}{\mathcal{H}}_{12}$, $\overset{n}{\mathcal{H}}_{21}$ imaginary, and imposed the relations

$$\overset{n}{\mathcal{H}}_{A2}=i\overset{n+1}{\mathcal{H}}{}_{A1}. \tag{4.18}$$

These immediately lead to conditions on $\overset{mn}{N}_{AB}$:

$$\overset{mn}{N}_{21}=-i\overset{m+1,n}{N}_{11},\quad \overset{mn}{N}_{12}=i\overset{m,n+1}{N}_{11},\quad \overset{mn}{N}_{22}=\overset{m+1,n+1}{N}_{11}, \tag{4.19}$$

so that it is sufficient to calculate just $\overset{mn}{N}_{11}$:

$$\begin{aligned}\nabla\overset{mn}{N}_{11}&=\overset{m}{\mathcal{H}}_{11}\nabla\overset{n}{\mathcal{H}}_{21}+\overset{m}{\mathcal{H}}_{21}\nabla\overset{n}{\mathcal{H}}_{11}\\&=i\rho^{m+n-2}[na_{m-1}b_n+(n-1)\,b_ma_{n-1}]\nabla\rho\\&\quad+i\rho^{m+n-1}[a_{m-1}b'_n+b_ma'_{n-1}]\nabla x.\end{aligned}$$

The necessary integrability condition must be satisfied identically; and it is, by virtue of the recursion relations (4.12)—(4.14). Thus,

$$\overset{mn}{N}_{11} = i\rho^{m+n-1} \left[\frac{n\, a_{m-1} b_n + (n-1)\, b_m a_{n-1}}{m+n-1}\right] . \qquad (4.20)$$

A most important question to resolve is which, if any, of the transformations of $\mathbf{K}'$ preserve asymptotic flatness. For finite $\overset{k}{\gamma}_{AB}$, Eq. (3.1) is symbolically

$$\overset{1}{\mathcal{H}} \rightarrow \overset{1}{\mathcal{H}} + \overset{k}{\gamma}\, \overset{k+1}{\mathcal{H}} + (\overset{k}{\gamma})^2\, \overset{2k+1}{\mathcal{H}} + \cdots . \qquad (4.21)$$

Since the potentials for Minkowski space itself are polynomials in ρ and z of ever increasing degree, one might suppose that all of the $\overset{k}{\gamma}$, $k \geqslant 1$, would necessarily violate asymptotic flatness. However, this conclusion is not justified. Equation (4.21) is effectively a power series in $(\overset{k}{\gamma}\rho^k)$ for *small* $\overset{k}{\gamma}$. We cannot draw *any* conclusions from it regarding ρ^k *large* until the series has been explicitly summed.

*Supported by National Science Foundation Grant PHY76-12246.

[1] W. Kinnersley, J. Math. Phys. 18, 1529 (1977).

[2] W. Kinnersley, J. Math. Phys. 14, 651 (1973).

[3] J. Ehlers, in *Les théories relativistes de la gravitation* (CNRS, Paris, 1959).

[4] B. K. Harrison, J. Math. Phys. 9, 1744 (1968).

[5] R. Geroch, J. Math. Phys. 13, 394 (1972).

[6] To make the equations easier to read, we always denote contracted indices by $X, Y = 1, 2$.

[7] We can relate certain components of these potentials to certain members of the octet Ψ_i introduced in I:

$$K = -(\tfrac{1}{3}\Psi_8 + \varphi_1\Psi_5), \quad M_1 = \Psi_6 + i\varphi_1\Psi_2 - \tfrac{1}{3}\varphi_1\Psi_8,$$
$$L_1 = \Psi_7 - \mathcal{G}_{11}\Psi_5, \quad N_{11} = \Psi_3 + i\,\mathcal{G}_{11}\Psi_2 - \tfrac{1}{3}\,\mathcal{G}_{11}\Psi_8.$$

[8] We should distinguish between a parameter and the operator that produces its corresponding transformation, e.g., $\overset{k}{\sigma}$ and $\mathcal{T}_{\overset{k}{\sigma}}$, but to simplify the notation we have not done so.

[9] There are other gauge transformations which could also be included in $\mathbf{K}'$ if desired. For example, to each of the potentials defined in Eqs. (2.10)—(2.13), we may add an arbitrary constant. However, these are completely inessential, and we have no desire to make $\mathbf{K}'$ any bigger!

[10] See I, Footnote 7.

[11] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (NBS, Washington, D.C., 1964), Chap. 22.

Stability of solutions of the compressible Navier–Stokes equations

Paul Gordon

Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104

(Received 22 September 1975; revised manuscript received 25 October 1976)

The purpose of the paper is to consider the effect of the viscous terms of the compressible Navier–Stokes equations in situations where significant variations of the flow variables occur only over many mean-free paths. Under these conditions it is shown that the magnitude of the viscous terms is comparable to the addition in the inviscid equations of a forcing term which is small relative to appropriately normalized initial data. The appropriate normalization is discussed. Stability, in terms of a stability parameter, is defined relative to such perturbations. A large value of the stability parameter implies that the solution is ill-conditioned. Application of the results proceeds in the following manner. Let $\tilde{V}(t,x,y)$ be a given solution to the Navier–Stokes equations. Let $V(t,x,y)$ be the corresponding solution to the inviscid equations. Let M be the maximum deviation that occurs between $\tilde{V}$ and V. From $\tilde{V}$, one obtains an ϵ which represents an upper bound on the magnitude of the viscous terms. If M is not significant, then it is clear, without any further analysis, that the viscous terms are not significant. If M is significant, one proceeds to obtain a lower bound on the stability parameter λ, namely $\lambda \geq M/\epsilon$. If ϵ is small, it is now possible to conclude that the solution $\tilde{V}$ is ill-conditioned. (If ϵ is large, the analysis remains valid, but no useful information is obtained.) Two particular applications are made. The first considers the known solutions to shock wave structure equations (particularly the case of the weak shock). The second considers solutions to the incompressible equations. It is shown that in many situations the use of the time-dependent incompressible Navier–Stokes equations is unjustified.

I. INTRODUCTION

The Navier–Stokes equations, as a model for fluid flow phenomena, are based on the hypothesis that the flow is a continuum. When significant variations of the phenomena under consideration occur over distances of a few mean-free paths, this hypothesis becomes questionable. Thus, it is generally agreed that the use of the Navier–Stokes equations can be justified on physical grounds only if significant variations do not occur over a distance of only a few mean-free paths.[1–3]

The purpose of this paper is to consider the effect of the viscous terms in situations where significant variations occur only over many mean-free paths. The Navier–Stokes equations can be obtained by introducing second order terms into the inviscid (Euler) equations of hydrodynamics. These second order terms affect the solution in two ways:

(1) The required boundary conditions, in order to determine a unique solution, may be changed.

(2) The flow profiles internal to the flow may be changed simply because the equations are modified.

It is not the purpose of the present paper to discuss the first point (this is not to say it is unimportant). Rather, the intent is to show that, in an internal region of the flow where significant variations occur only over many mean-free paths, the viscous terms provide only a small perturbation on the forces described by the inviscid equations. In Sec. III, using a simple order of magnitude analysis, the following is shown: If a particular solution of the compressible Navier–Stokes equations is such that significant variations of the flow variables occur only over distances of many mean-free paths, then the magnitude of the viscous terms is comparable to the addition in the inviscid equations of a forcing term which is small relative to appropriately normalized initial data.

The appropriate normalized variables are discussed in Sec. II. Also definitions of stability and a stability parameter, the latter being essentially a condition number, are introduced in Sec. II.

To apply the result to a given solution of the compressible Navier–Stokes equations, one proceeds as follows. Let $\tilde{V}(t,x,y)$ be the given solution to the Navier–Stokes equations, and let $V(t,x,y)$ be the solution obtained with the inviscid equations. (Note that the analysis requires both $\tilde{V}$ and V to be known.) Let M be the maximum deviation that occurs between $\tilde{V}$ and V; that is, $M = \max_{t,x,y} \| \tilde{V} - V \|$, where the norm is a usual vector norm. From $\tilde{V}$, using the estimate of Sec. III, one obtains an ϵ which represents an upper bound on the magnitude of the viscous terms. There are now several possibilities. The first possibility is that M is insignificant; in this case, without further analysis, one concludes that the viscous terms are insignificant. A second possibility is that the solution $\tilde{V}$ is not stable relative to ϵ perturbations, where ϵ is the value obtained above; if ϵ is small, then of course little physical significance can be attached to $\tilde{V}$. In the third case, that M is significant and $\tilde{V}$ is stable, the results of Sec. II produce a lower bound on the stability parameter λ, namely $\lambda \geqslant M/\epsilon$. If ϵ is small, one now concludes that the solution $\tilde{V}$ is ill-conditioned. (It should be noted that in some situations, such as a strong shock, ϵ will be large. In these cases the analysis remains valid, but no useful information is obtained regarding the solution $\tilde{V}$.)

In Sec. IV two applications are considered. First, known solutions for shock wave structure equations (particularly for the case of weak shocks) are examined in order to see that such solutions are compatible with the analysis of the present paper. The second application is to solutions of the incompressible equations (particularly the Poiseuille solution), under the assumption that such solutions constitute close approximations to solutions of the compressible equations. Assuming a typical flow situation, it is shown that these solutions give rise to a very large stability parameter. Consequently, one must judge these solutions to be highly ill-conditioned.

Obtaining an estimate for the magnitude of the viscous terms is an important part of the analysis. The estimate of Sec. III, as noted above, puts the viscous terms in the form of a forcing function which is then added directly to the inviscid equations. Intuitively, it is more informative to express the second order viscous terms as first order terms. In such a form the viscous terms can be compared directly to terms of the inviscid system; in particular, if one has a solution (either analytic or numeric), the estimates of the present paper can be checked by simple numerical calculations. A procedure for expressing the viscous terms as first order terms is described in the Appendix. This procedure was not adopted in the main body of the paper because the remainder of the analysis becomes considerably more complicated.

II. STABILITY AND NORMALIZED VARIABLES

Hyperbolic differential equations of the following form will be considered first.

$$W_t = A W_x + F, \quad W(0,x) = \Phi(x), \tag{1}$$

where

$$F = F(t,x,W), \quad A = A(t,x,W) = (a_{ij}) \text{ is an } n \times n \text{ matrix},$$

$$\Phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix}, \quad W = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad \text{and} \quad F = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}.$$

By hyperbolic one means the following: There exists a region $R \supset (0, x_0, \phi_1(x_0), \ldots, \phi_n(x_0))$ of $2+n$ space such that in R, A has n real eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and there exists a real matrix M for which $MAM^{-1} = D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$, where $\lambda_i = \lambda_i(t,x,W)$ and $M = M(t,x,W)$. It is also assumed that the a_{ij} are twice differentiable in all variables, D is continuous, and M can be chosen to be continuous.

The following norms will be used:

$$\| W \| = \max_j |w_j|, \tag{2a}$$

$$\| A \| = \max_i \left[\sum_{j=1}^{n} |a_{ij}| \right], \tag{2b}$$

$$\| W \|_{I(t)} = \max_{x \in I(t)} \| W(t,x) \|. \tag{2c}$$

Let $V(t,x)$ be the solution to Eq. (1) corresponding to initial data $\phi(x)$ and let $\tilde{V}(t,x)$ be the solution to Eq. (1) with initial data $\tilde{\phi}(x)$. Fundamental existence theorems[4,5] establish stability relative to perturbations of initial data. Thus, there exist positive constants C, δ, and t_1 such that for $\| \tilde{\phi} - \phi \|_{I(0)} < \delta$, $\| \tilde{V} - V \|_{I(t)} \le C\delta$ for $0 < t \le t_1$, where the intervals $I(t)$ depend continuously on t. These results do not give information concerning the magnitude of C. The value of C is important, for, if C is large, it may be that Eq. (1) is well-posed in the usual mathematical sense, but ill-conditioned in the usual sense of linear algebra. (This "degree of amplification" is also discussed by Homsy.[6])

A primary difficulty in estimating C is that its value depends on the scaling of the variables. This can be illustrated by the following example.[7]

Example 1: Let $A = \begin{pmatrix} 0 & -c^2 \\ -1 & 0 \end{pmatrix}$, where c is a positive constant and $W = \begin{pmatrix} u \\ p \end{pmatrix}$. Assuming initial data $p(0,x)$ and $u(0,x)$ the solution can be written as,

$$\begin{aligned} u(t,x) &= \tfrac{1}{2}[u(0,x-ct) + u(0,x+ct)] \\ &\quad + (c/2)[p(0,x-ct) - p(0,x+ct)], \\ p(t,x) &= (1/2c)[u(0,x-ct) - u(0,x+ct)] \\ &\quad + \tfrac{1}{2}[p(0,x-ct) + p(0,x+ct)]. \end{aligned}$$

Although continuously dependent on initial data, the solution can be sensitive to perturbation if c is either large or small. The difficulty here is that perturbations of u and p do not have effects of comparable magnitude (if c is not near 1). Thus, u and p are not "consistent" quantities in terms of perturbation. In this example, the difficulty is removed by scaling the variables, $\tilde{u} = u/c$ and $\tilde{p} = p$. This gives $\tilde{W}_t = \tilde{A}\tilde{W}_x$, where $\tilde{A} = \begin{pmatrix} 0 & -c \\ -c & 0 \end{pmatrix}$ and $\tilde{W} = \begin{pmatrix} \tilde{u} \\ \tilde{p} \end{pmatrix}$. The solution is,

$$\begin{aligned} \tilde{u}(t,x) &= \tfrac{1}{2}[\tilde{u}(0,x-ct) + \tilde{u}(0,x+ct)] \\ &\quad + \tfrac{1}{2}[\tilde{p}(0,x-ct) - \tilde{p}(0,x+ct)], \\ \tilde{p}(t,x) &= \tfrac{1}{2}[\tilde{u}(0,x-ct) - \tilde{u}(0,x+ct)] \\ &\quad + \tfrac{1}{2}[\tilde{p}(0,x-ct) + \tilde{p}(0,x+ct)]. \end{aligned}$$

This scaling problem was considered in Ref. 7. One result obtained there is as follows: Suppose in Eq. (1), that A is constant and $F \equiv 0$; suppose A is irreducible (that is, independent permutations of rows and columns cannot produce a form $\begin{pmatrix} M_1 & M_2 \\ 0 & M_3 \end{pmatrix}$, where M_1 and M_3 are square matrices and 0 is a zero matrix); suppose also that A has distinct real eigenvalues; then, a smallest value of the constant C (called the condition number K) could be obtained for a best normalization of W (say, $\tilde{W} = \Gamma W$ where Γ is diagonal); if $\tilde{A} = \Gamma A \Gamma^{-1}$, $K = \| M \| \circ \| M^{-1} \|$ for some M which satisfies $M\tilde{A}M^{-1} = D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$. The scaling of W is such that equal perturbations of the components of the initial data have roughly equal effects (as indicated in the previous example).

For purposes of the present paper, one requires stability relative to perturbations of the matrix A and the vector F. The results provide an estimate for the corresponding constant C in the case of global stability, where perturbations are measured relative to normalized variables. The following definitions and Theorem 1 establish local stability and show that, in the nonlinear case, the local condition number of A and the local normalized variables are still the pertinent quantities. In the following, $V(t,x)$ will represent a solution to Eq. (1), and $\tilde{V}(t,x)$ will represent the solution to Eq. (1) where $\{\tilde{A}, \tilde{\phi}, \tilde{F}\}$ replaces $\{A, \phi, F\}$. Also, it is assumed that V is in normal form at the point $(0, x_0, \phi(x_0))$.

Definition 1: At any point of R, let $MAM^{-1} = D = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ and let $\tilde{M}\tilde{A}\tilde{M}^{-1} = \tilde{D} = \text{diag}\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n\}$. Then $\{\tilde{A}, \tilde{\Phi}, \tilde{F}\}$ is an ϵ perturbation of Eq. (1) if,

(1) $\tilde{a}_{ij}$ are twice differentiable,

(2) $\|\tilde{D} - D\| < \epsilon$,

(3) $\tilde{M}$ can be chosen so that $\|\tilde{M} - M\| \leqslant \epsilon$,

(4) $\|\tilde{\Phi} - \Phi\| \leqslant \epsilon$,

(5) $\|\tilde{F} - F\| \leqslant \epsilon$.

Definition 2: Eq. (1) is locally ϵ stable at $(0, x_0)$ with stability parameter λ, if for some $t_1 > 0$ there are intervals $I(t)$ continuous in t for $0 \leqslant t \leqslant t_1$ and with $x_0 \in I(0)$, such that for all ϵ perturbations of Eq. (1), $\|\tilde{V}(t,x) - V(t,x)\|_{I(t)} \leqslant \epsilon\lambda$ for $0 \leqslant t \leqslant t_1$ and $x \in I(t)$.

Theorem 1: Let ϵ be such that for any ϵ perturbation of Eq. (1) the following holds: For some $t_1 > 0$, there exists open intervals $I(t)$, continuous in t for $0 \leqslant t \leqslant t_1$ and with $x_0 \in I(0)$, such that $\tilde{V}(t,x)$ exists for $0 \leqslant t \leqslant t_1$ and $x \in I(t)$. Let $A_0 = A(0, x_0, \Phi(x_0))$ and let A_0 have condition number K. Let $M_0 A_0 M_0^{-1} = D_0 = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ where $\|M_0\| = \|M_0^{-1}\| = \sqrt{K}$. Then, Eq. (1) is locally ϵ stable at $(0, x_0)$ with local stability parameter λ, where $\lambda = K + R_1(\epsilon) + tR_2(\epsilon) + tR_3(t, \epsilon)$, where $\lim_{\epsilon \to 0} R_1(\epsilon) = 0$, $\lim_{t \to 0} R_3(t, \epsilon) = 0$, R_2 arises from $F(0, x_0, \Phi(x_0))$, R_1 arises from nonlinearities in the matrix M, and R_3 accounts for other nonlinearities.

Proof: Let $V_\epsilon(t,x) = \tilde{V}(t,x) - V(t,x)$. Then $(V_\epsilon)_t = \tilde{V}_t - V_t = \tilde{A}(t,x,\tilde{V})\tilde{V}_x - A(t,x,V)V_x + \tilde{F}(t,x,\tilde{V}) - F(t,x,V) = \tilde{A}(t,x,\tilde{V})(V_\epsilon)_x + [\tilde{A}(t,x,\tilde{V}) - A(t,x,V)]V_x + (\tilde{F} - F)$. Considering V as known, one obtains,

$$\cdot(V_\epsilon)_t = A^*(t,x,V_\epsilon)(V_\epsilon)_x + F^*(t,x,V_\epsilon), \tag{3a}$$

where $\lim_{\epsilon=0} \|F^*\| = 0$. Let $\tilde{A}_0 = A^*(0, x_0, V_\epsilon(0, x_0)) = \tilde{A}(0, x_0, \tilde{\Phi}(x_0))$ and let $\tilde{M}_0 \tilde{A}_0 \tilde{M}_0^{-1} = \tilde{D}_0 = \text{diag}\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n\}$. Let

$$\tilde{W} = \tilde{M}_0 V_\epsilon = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad \tilde{G}_0 = \tilde{M}_0 F^*(0, x_0, V_\epsilon(0, x_0)) = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix},$$

$$\psi = \tilde{M}_0 V_\epsilon(0, x) = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_n \end{pmatrix}.$$

Then the linearized form of Eq. (3a) can be written as $\tilde{W}_t = \tilde{D}_0 \tilde{W}_x + \tilde{G}_0$ and has the solution $w_i(t,x) = \psi_i(x + \tilde{\lambda}_i t) + g_i t$. The solution to Eq. (3a) can now be written as,

$$V_\epsilon(t,x) = \tilde{M}_0^{-1} W(t,x) + H(t,x), \tag{3b}$$

where the solution is valid in $I(t)$ given by the choice of ϵ, and $H(t,x)$ is of order greater than 1 in t and of order at least 1 in ϵ. Thus,

$$\|V_\epsilon(t,x)\| \leqslant \|\tilde{M}_0^{-1}\| [\|\psi\| + t\|\tilde{G}_0\|] + \|H(t,x)\|$$
$$\leqslant \|\tilde{M}_0^{-1}\| [\|\tilde{M}_0\| \circ \|V_\epsilon(0,x)\| + t\|\tilde{G}_0\|] + \|H(t,x)\|,$$

or,

$$\|V_\epsilon(t,x)\|_{I(t)} \leqslant \|\tilde{M}_0^{-1}\| \circ \|\tilde{M}_0\| [\|\tilde{V}(0,x) - V(0,x)\|_{I(0)}$$
$$+ t\|\tilde{F}(0,x,\tilde{\Phi}) - F(0,x,\Phi)\|]_{I(0)} + \|H(t,x)\|_{I(t)}.$$

Thus,

$$\|V_\epsilon(t,x)\|_{I(t)} \leqslant [\|\tilde{M}_0^{-1}\| \circ \|\tilde{M}_0\| + tR_2(\epsilon) + tR_3(t,\epsilon)]\epsilon. \tag{3c}$$

It remains to estimate $\|\tilde{M}_0^{-1}\| \circ \|\tilde{M}_0\|$. Let

$$K_1 = \max \left\{ \frac{\|M(0, x_0, \Phi(x_0)) - M(0, x_0, \tilde{\Phi}(x_0))\|}{\|\Phi(x_0) - \tilde{\Phi}(x_0)\|} \right\}$$

over all Φ and $\tilde{\Phi}$ satisfying $\|\Phi(x_0) - \tilde{\Phi}(x_0)\| \leqslant \epsilon$. It can be assumed that $\|\tilde{M}_0\| = \|\tilde{M}_0^{-1}\|$.

$$M_0 - \tilde{M}_0 = M(0, x_0, \Phi(x_0)) - \tilde{M}(0, x_0, \tilde{\Phi}(x_0))$$
$$= M(0, x_0, \Phi(x_0)) - M(0, x_0, \tilde{\Phi}(x_0))$$
$$+ M(0, x_0, \tilde{\Phi}(x_0)) - \tilde{M}(0, x_0, \tilde{\Phi}(x_0))$$

or

$$\|M_0 - \tilde{M}_0\| \leqslant K_1 \|\Phi(x_0) - \tilde{\Phi}(x_0)\| + \|M(0, x_0, \tilde{\Phi}(x_0))$$
$$- \tilde{M}(0, x_0, \tilde{\Phi}(x_0))\| \leqslant (K_1 + 1)\epsilon.$$

Thus,

$$\|\tilde{M}_0\| = \|M_0 - (M_0 - \tilde{M}_0)\| \leqslant \|M_0\| + \|M_0 - \tilde{M}_0\| \leqslant \sqrt{K} + \epsilon(1 + K_1).$$

This gives

$$\|\tilde{M}_0^{-1}\| \circ \|\tilde{M}_0\| \leqslant [\sqrt{K} + \epsilon(1 + K_1)]^2 = K + \epsilon(2\sqrt{K})(1 + K_1)$$
$$+ \epsilon^2(1 + K_1)^2 = K + R_1(\epsilon).$$

Substituting into Eq. (3c) one obtains $\lambda = K + R_1(\epsilon) + tR_2(\epsilon) + tR_3(t,\epsilon)$, with the properties given by the theorem.

Remark: As noted earlier, Theorem 1 implies that the local condition number is essentially unchanged by the nonlinearities of the equation. Consequently, even for the nonlinear case, the local (linear) normalization is still appropriate.

Next, consider the more general equation,

$$W_t = AW_x + BW_y + F, \tag{4}$$

where (x, y) is in a closed bounded set S in 2-space and $t \geqslant 0$, where $W(0, x, y) = \phi(x, y)$, where $F = F(t, x, y, W, W_{xx}, W_{yy}, W_{xy})$, and where boundary conditions are as required for the particular problem. In regard to Eq. (4), the effect of perturbations of A, B, and F on steady-state solutions are to be considered.

Definition 3: Let τ be the domain $t \geqslant 0$ and let $S^* = S \times \tau$. Let $V(t, x, y)$ be the solution to Eq. (4) in S^*. Let $\tilde{V}(t, x, y)$ be the solution to Eq. (4) in S^* where $\{\tilde{A}, \tilde{B}, \tilde{F}\}$ replaces $\{A, B, F\}$. Then, $V(t, x, y)$ is an asymptotically ϵ stable solution to Eq. (4), with parameter λ, if $\|\tilde{V} - V\| \leqslant \epsilon\lambda$ in S^* for all $\{\tilde{A}, \tilde{B}, \tilde{F}\}$ which are ϵ perturbations of Eq. (4).

Remark: ϵ perturbation is in the same sense as in Definition 1. Also, if a lower bound for λ is large, then $V(t, x, y)$ will be said to be an ill-conditioned solution [although Eq. (4) might still be well-posed in the usual mathematical sense].

Example 2: Let $u_t = cu_x$ in R, where c is a positive constant. Let $u(t,1) = \psi(t)$ where $\lim_{t\to\infty}\psi(t) = \alpha = \text{const}$. The solution is constant on the lines $x + tc = \text{const}$, or $\lim_{t\to\infty} u(t,x) = \alpha$ for all x. If $\tilde{c} > 0$ and $|\tilde{\psi}(t) - \psi(t)| \leq \epsilon$, then $|\tilde{u}(t,x) - u(t,x)| \leq \epsilon\lambda(t)$, where $\lim_{t\to\infty}\lambda(t) = 1$. Thus, the solution is asymptotically ϵ stable for $|\epsilon| \leq \epsilon_1 < c$. The value of λ depends on ϵ_1.

Example 3: Let A be as in Example 1. Each component of W, u and p, individually satisfies a wave equation, or the system may have solutions which are periodic in time. Such solutions will not be asymptotically stable (given any $\lambda > 0$ and ϵ arbitrarily small, $\max_{0\leq x\leq 1} \| V(t,x) - \tilde{V}(t,x)\| \geq \epsilon\lambda$ for t sufficiently large). However, suppose the equation is transformed to $\tilde{W}_t = D\tilde{W}_x$, where $\tilde{W} = MW$, $W = \binom{u/c}{p}$, $\tilde{W} = \binom{w_1}{w_2}$, and $D = \begin{pmatrix} -c & 0 \\ 0 & c\end{pmatrix}$. Suppose the following boundary conditions are given: $w_1(t,0) = \alpha_1 = \text{const}$ and $w_2(t,1) = \alpha_2 = \text{const}$. Then, as in Example 2, the solution is asymptotically stable.

Example 4: $u_t = u_{xx} + 2\beta u_x + \epsilon$, $0 < x < 1$, and $t > 0$. ϵ and β are constants, $u(0,x) = f(x)$ and $u(t,0) = u(t,1) = 0$. The solution can be written as $u(t,x) = -U(x) + \sum_{n=1}^{\infty} c_n \times \exp(-\beta x)(\sin n\pi x)\exp(-\beta^2 t)\exp(-n^2\pi^2 t)$, where $c_n = 2\int_0^1 [f(x) + U(x)]\exp(\beta x)\sin n\pi x\,dx$ and

$$U(x) = \frac{\epsilon}{2\beta}\left[x - \frac{1 - \exp(-2\beta x)}{1 - \exp(-2\beta)}\right].$$

[Note that at $\beta = 0$, $U(x)$ reduces to $(\epsilon/2)x(x-1)$, which is the correct value.] The solution is locally stable and asymptotically stable with respect to both ϵ and β.

Example 5: Suppose $\tilde{V}(t,x)$ is a solution to Eq. (1). Suppose in a neighborhood of $(0,x_0)$, $F(t,x,\tilde{V}) = A_\epsilon\tilde{V}_x$. Thus, $\tilde{V}(t,x)$ is a solution to $W_t = \tilde{A}W_x$ where $\tilde{A} = A + A_\epsilon$. Let $V(t,x)$ be the solution to $W_t = AW_x$, where $\tilde{V}(0,x) = V(0,x)$. Then if Eq. (1) is locally ϵ stable, then by Theorem 1 $\|\tilde{V} - V\| \leq \epsilon\lambda$ near $(0,x_0)$.

Example 6: Suppose $V(t,x)$ is an ϵ asymptotically stable solution to Eq. (1) such that for all (t,x) $F = A_\epsilon V$ where $\|A_\epsilon V\| \leq \epsilon$. Then, by Definition 3, Eq. (1) with $F \equiv 0$ must have a solution $\tilde{V}(t,x)$, satisfying the same boundary conditions and initial conditions as $V(t,x)$, such that $\|\tilde{V} - V\| \leq \epsilon\lambda$ for all x and t.

In applications to the Navier–Stokes equations, the situation will be as in Examples 5 and 6. The given information will then be used to obtain a lower bound on the stability parameter λ. This is the thrust of the following theorem.

Theorem 2: Let $V(t,x,y)$ be an asymptotically ϵ stable solution to Eq. (4) such that in S^* at least one of the following two estimates hold:

(i) $F = A_\epsilon V_x + B_\epsilon V_y$ where A_ϵ and B_ϵ constitute ϵ perturbations of A and B,

(ii) $\|F\| \leq \epsilon$.

Let $\tilde{V}(t,x,y)$ be the solution of Eq. (4) with $F \equiv 0$. Then, $\lambda \geq (1/\epsilon)\max_{S^*}\|V - \tilde{V}\|$.

Proof: By Definition 3 $\tilde{V}$ exists and $\|V - \tilde{V}\| \leq \epsilon\lambda$ in S^*.

Remark: If the lower bound for λ, produced by Theorem 2, is large, then the solution $V(t,x,y)$ will be said to be ill-conditioned. If the bound is small, then the analysis remains valid, but no useful information is obtained.

III. AN ESTIMATE FOR THE VISCOUS TERMS OF THE COMPRESSIBLE NAVIER–STOKES EQUATIONS

An important part of the analysis is of course involved in obtaining an estimate for the viscous terms of the compressible Navier–Stokes equations. It is significant that little precision is required in this aspect of the problem. That is, the following lemma, which is used to estimate the viscous terms, is a simple result.

Lemma 1: Let $f(x)$ be differentiable in an interval I of the x axis. For any $\Delta > 0$ associate $\epsilon(\Delta) > 0$ such that the following conditions hold whenever $|\Delta x| < \Delta$ and $x + \Delta x$ in I:

(a) $|f(x+\Delta x) - f(x)| \leq \epsilon$,

(b) $\left|\dfrac{f(x+\Delta x) - f(x)}{\Delta x} - \dfrac{df}{dx}(x)\right| \leq \epsilon$.

Then, $(df/dx)(x) = \epsilon_1(x)/\Delta$ where $|\epsilon_1(x)| \leq \epsilon(1+\Delta)$.

Proof: From (b),

$$\frac{df}{dx}(x) = \tilde{\epsilon}(x,\Delta) + \frac{f(x+\Delta x) - f(x)}{\Delta x},$$

where $|\tilde{\epsilon}| < \epsilon$. From (a),

$$\frac{df}{dx}(x) = \tilde{\epsilon} + \frac{\tilde{\tilde{\epsilon}}(x,\Delta)}{\Delta},$$

where $|\tilde{\tilde{\epsilon}}| < \epsilon$, or

$$\frac{df}{dx}(x) = \frac{\epsilon_1(x)}{\Delta}$$

where $|\epsilon_1| \leq |\tilde{\tilde{\epsilon}}| + \Delta|\tilde{\epsilon}| \leq \epsilon(1+\Delta)$.

Using Lemma 1, the viscous terms will now be expressed as a perturbation, in the form of a forcing term, of the inviscid system. It is perhaps intuitively more satisfying to see the viscous terms expressed as a perturbation of the matrix which defines the inviscid system. A procedure for accomplishing this is given in the Appendix.

The one-dimensional equations can be written as follows:

$$\frac{\partial W}{\partial t} = -A\frac{\partial W}{\partial x} + F, \tag{5}$$

where

$$W = \begin{pmatrix} \rho \\ u \\ T \end{pmatrix}, \quad A = \begin{pmatrix} u & \rho & 0 \\ \dfrac{1}{\rho}\dfrac{\partial p}{\partial \rho} & u & \dfrac{1}{\rho}\dfrac{\partial p}{\partial T} \\ 0 & \dfrac{p}{\rho c_v} & u \end{pmatrix},$$

$$F = \begin{pmatrix} 0 \\ \left(\dfrac{2\mu + \lambda}{\rho}\right)\dfrac{\partial^2 u}{\partial x^2} \\ \left(\dfrac{2\mu+\lambda}{\rho c_v}\right)\left(\dfrac{\partial u}{\partial x}\right)^2 + \left(\dfrac{k}{\rho c_v}\right)\dfrac{\partial^2 T}{\partial x^2} \end{pmatrix},$$

$p =$ pressure $= p(\rho, T)$,

$\rho =$ density,

$u = x$ component of velocity,

$T =$ temperature,

$t =$ time coordinate,

$x =$ distance coordinate,

$c_v =$ specific heat,

$\mu =$ coefficient of shear viscosity $=$ const,

$\lambda =$ coefficient of bulk viscosity $=$ const,

$k =$ coefficient of thermal conductivity $=$ const.

The following quantities will also be used:

$$\tau = \text{the mean-free path} = \frac{2\mu + \lambda}{\rho c_s},$$

$$P_r^* = \frac{c_v}{k}(2\mu + \lambda) = \frac{c_v \rho \tau c_s}{k},$$

$$b = \frac{(2\mu + \lambda)}{p} \, \frac{c_s}{\tau} = \frac{(2\mu + \lambda) c_s \rho c_s}{p(2\mu + \lambda)} = \frac{\rho c_s^2}{p},$$

$$c_s^2 = \frac{\partial p}{\partial \rho} + \frac{p}{\rho^2 c_v} \, \frac{\partial p}{\partial T}.$$

Remark: The definition of τ corresponds generally to that used in fluid mechanics.[8,9] P_r^* is essentially a Prandtl number. However, the Prandtl number is defined normally in terms of c_p, rather than c_v.

Suppose a twice differentiable solution $V(t, x)$ to Eq. (5) is given about any base point (t_0, x_0). Let $\Delta = \tau$ and let I be an interval of the x axis containing x_0. Applying Lemma 1 successively to the case where $f = \partial^2 u/\partial x^2$, $\partial u/\partial x$, and $\partial^2 T/\partial x^2$, one obtains

$$\frac{\partial^2 u}{\partial x^2} = \frac{\epsilon_1}{\tau},$$
$$\left(\frac{\partial u}{\partial x}\right)^2 = \frac{\epsilon_2}{\tau} \, \frac{\partial u}{\partial x}, \tag{6}$$
$$\frac{\partial^2 T}{\partial x^2} = \frac{\epsilon_3}{\tau},$$

where ϵ_1 is the variation of $\partial u/\partial x$, ϵ_2 is the variation of u, and ϵ_3 is the variation of $\partial T/\partial x$. F can now be written as follows:

$$F = \begin{pmatrix} 0 \\ (2\mu + \lambda)\dfrac{\epsilon_1}{\rho\tau} \\ \dfrac{(2\mu + \lambda)}{\rho c_v}\left(\dfrac{\epsilon_2}{\tau}\right) \dfrac{\partial u}{\partial x} + \left(\dfrac{k}{\rho c_v}\right) \dfrac{\epsilon_3}{\tau} \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ c_s \epsilon_1 \\ \epsilon_2 \dfrac{p}{\rho c_v}\left(\dfrac{b}{c_s} \dfrac{\partial u}{\partial x}\right) + \dfrac{c_s \epsilon_3}{P_r^*} \end{pmatrix} = F_\epsilon.$$

Thus, the solution $V(t, x)$ of Eq. (5) also solves the following equation in a neighborhood of (t_0, x_0),

$$\frac{\partial W}{\partial t} = -A \frac{\partial W}{\partial x} + F_\epsilon(t, x). \tag{7}$$

F_ϵ does not depend on W because F was evaluated in terms of the known solution V.

Remark: It has been assumed for simplicity that the viscous coefficients are constant. For example, in the variable case the second component of F could be estimated as follows:

$$\frac{\partial}{\partial x}\left((2\mu + \lambda)\frac{\partial u}{\partial x}\right) = \frac{\epsilon}{\tau}(2\mu + \lambda),$$

where $\epsilon = \tilde{\epsilon}_1/(2\mu + \lambda)$ and $\tilde{\epsilon}_1$ is the variation of $(2\mu + \lambda) \times (\partial u/\partial x)$. (See the Appendix.)

Theorem 3: Let ϵ_1, ϵ_2, ϵ_3 be as defined in Eq. (6). Let

$$\epsilon^* = \max\left[\frac{|\epsilon_1|}{c_s}, \frac{|\epsilon_2|}{c_s}, |\epsilon_3| \frac{\rho c_v}{p}\right].$$

Let

$$\epsilon = \epsilon^* \max\left[1, \frac{1}{P_r^*} + \frac{b}{c_s}\left|\frac{\partial u}{\partial x}\right|\right].$$

Then, in a neighborhood of (t_0, x_0), the viscous terms constitute an ϵ perturbation of the equation $\partial W/\partial t = -A(\partial W/\partial x)$.

Proof: The result follows from Definition 1, after it is shown that ϵ^* is in terms of normalized variables. The following was established in Ref. 7. The normalized variables for $\partial W/\partial t = -A(\partial W/\partial x)$, where A is given by Eq. (5), are $\tilde{W} = \Gamma W$, where $\Gamma = \operatorname{diag}\{1/\rho\beta_1, 1/c_s, \rho c_v/p\beta_2\}$ is evaluated at the base point (t_0, x_0). The β_i and the condition number are defined as follows:

$$\beta_1 = 1 + \frac{2\alpha}{(1+\alpha)(K-1)}, \quad \beta_2 = 1 + \frac{2}{(1+\alpha)(K-1)},$$

$$\alpha p_\rho = \frac{p p_T}{\rho^2 c_v}, \quad \tilde{\alpha} = \frac{4\alpha}{(1+\alpha)^2},$$

and $K =$ condition number $= \frac{1}{2}[3 + (1 + 4\tilde{\alpha})^{1/2}]$. (The factors β_i are such that $1 \leq \beta_i \leq 3$.)

Now one needs only to note that Eq. (7) can be written as,

$$\frac{1}{c_s}\Gamma \frac{\partial W}{\partial t} = -A^* \Gamma \frac{\partial W}{\partial x} + \frac{1}{c_s}\Gamma F_\epsilon,$$

where

$$A^* = \frac{1}{c_s}\Gamma A \Gamma^{-1} = \begin{pmatrix} \dfrac{u}{c_s} & 1 & 0 \\ \dfrac{1}{c_s^2} \dfrac{\partial p}{\partial \rho} & \dfrac{u}{c_s} & \dfrac{p p_T}{\rho^2 c_v c_s^2} \\ 0 & 1 & \dfrac{u}{c_s} \end{pmatrix}$$

and

$$\frac{\Gamma}{c_s} F_\epsilon = \begin{pmatrix} 0 \\ \dfrac{\epsilon_1}{c_s} \\ \dfrac{\epsilon_2 b}{\beta_2 c_s}\left(\dfrac{1}{c_s} \dfrac{\partial u}{\partial x}\right) + \dfrac{\epsilon_3 \rho c_v}{P_r^* p} \end{pmatrix}.$$

Remark: $2 \leq K \leq (3+\sqrt{5})/2 \sim 2.62$, or the inviscid system is well conditioned in the sense that it has a small condition number.

Remark: The ϵ_i represent, for the corresponding quantities, the relative deviation that exists over a distance τ. In general, the ϵ_i need not be small. For example, in the case of a strong shock, the ϵ_i would be quite large.

Consider next the two-dimensional equations,

$$\frac{\partial W}{\partial t} = -A\frac{\partial W}{\partial x} - B\frac{\partial W}{\partial y} + F, \tag{8}$$

where

$$A = \begin{pmatrix} u & p & 0 & 0 \\ \dfrac{1}{\rho}\dfrac{\partial p}{\partial \rho} & u & 0 & \dfrac{1}{\rho}\dfrac{\partial p}{\partial T} \\ 0 & 0 & u & 0 \\ 0 & \dfrac{p}{\rho c_v} & 0 & u \end{pmatrix},$$

$$B = \begin{pmatrix} v & 0 & \rho & 0 \\ 0 & v & 0 & 0 \\ \dfrac{1}{\rho}\dfrac{\partial p}{\partial \rho} & 0 & v & \dfrac{1}{\rho}\dfrac{\partial p}{\partial T} \\ 0 & 0 & \dfrac{p}{\rho c_v} & v \end{pmatrix},$$

$$V = \begin{pmatrix} \rho \\ u \\ v \\ T \end{pmatrix}, \quad F = \begin{pmatrix} 0 \\ F_2 \\ F_3 \\ F_4 \end{pmatrix}, \quad v = y \text{ component of velocity},$$

$$F_2 = \left(\frac{2\mu+\lambda}{\rho}\right)\frac{\partial^2 u}{\partial x^2} + \left(\frac{\lambda+\mu}{\rho}\right)\frac{\partial^2 v}{\partial x\,\partial y} + \frac{\mu}{\rho}\frac{\partial^2 u}{\partial y^2},$$

$$F_3 = \frac{\mu}{\rho}\frac{\partial^2 v}{\partial x^2} + \left(\frac{\lambda+\mu}{\rho}\right)\frac{\partial^2 u}{\partial x\,\partial y} + \left(\frac{2\mu+\lambda}{\rho}\right)\frac{\partial^2 v}{\partial y^2},$$

$$F_4 = \frac{k}{\rho c_v}\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) + \frac{\lambda}{\rho c_v}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)^2 + \frac{\mu}{\rho c_v}\left[2\left(\frac{\partial u}{\partial x}\right)^2 + 2\left(\frac{\partial v}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2\right].$$

Let

$$\tau = \frac{1}{\rho c_s}\max[(2\mu+\lambda),\ |\mu+\lambda|,\ \mu,\ |\lambda|], \qquad P_r^* = \frac{c_v\rho\tau c_s}{k}, \quad b = \frac{\rho c_s^2}{p}. \tag{9}$$

Letting $V(t,x,y)$ be given solution to Eq. (8), Lemma 1 will now be applied to the various terms of F.

$$\left(\frac{2\mu+\lambda}{\rho}\right)\frac{\partial^2 u}{\partial x^2} = \frac{2\mu+\lambda}{\rho\tau}\tilde{\epsilon}_1 = \epsilon_1 c_s,$$

where from Eq. (9)

$$|\epsilon_1| \leq |\tilde{\epsilon}_1|, \quad \left(\frac{\lambda+\mu}{\rho}\right)\frac{\partial^2 v}{\partial x\,\partial y} = \frac{\lambda+\mu}{\rho\tau}\tilde{\epsilon}_2 = \epsilon_2 c_s,$$

$$\frac{\mu}{\rho}\frac{\partial^2 u}{\partial y^2} = \frac{\mu}{\rho\tau}\tilde{\epsilon}_3 = \epsilon_3 c_s, \quad \frac{\mu}{\rho}\frac{\partial^2 v}{\partial x^2} = \epsilon_4 c_s,$$

$$\frac{\lambda+\mu}{\rho}\frac{\partial^2 u}{\partial x\,\partial y} = \epsilon_5 c_s, \quad \left(\frac{2\mu+\lambda}{\rho}\right)\frac{\partial^2 v}{\partial y^2} = \epsilon_6 c_s,$$

$$\left(\frac{k}{\rho c_v}\right)\frac{\partial^2 T}{\partial x^2} = \frac{k\epsilon_7}{\rho c_v \tau} = \frac{\epsilon_7 c_s}{P_r^*}, \quad \left(\frac{k}{\rho c_v}\right)\frac{\partial^2 T}{\partial y^2} = \frac{\epsilon_8 c_s}{P_r^*},$$

$$\begin{aligned}
\frac{2\mu+\lambda}{\rho c_v}\left(\frac{\partial u}{\partial x}\right)^2 &= \left(\frac{2\mu+\lambda}{\rho c_v \tau}\right)\tilde{\epsilon}_9\frac{\partial u}{\partial x} \\
&= \left(\frac{p}{\rho c_v}\right)\left(\frac{\rho c_s}{p}\right)\left(\frac{2\mu+\lambda}{\rho c_s \tau}\right)\tilde{\epsilon}_9\frac{\partial u}{\partial x} \\
&= \frac{p}{\rho c_v}\left(\frac{b}{c_s}\right)\epsilon_9\frac{\partial u}{\partial x} \\
&= \left(\epsilon_9\frac{p}{\rho c_v}\right)b\left(\frac{1}{c_s}\frac{\partial u}{\partial x}\right),
\end{aligned}$$

$$\frac{2\mu+\lambda}{\rho c_v}\left(\frac{\partial v}{\partial y}\right)^2 = \left(\epsilon_{10}\frac{p}{\rho c_v}\right)b\left(\frac{1}{c_s}\frac{\partial v}{\partial y}\right),$$

$$\frac{\mu}{\rho c_v}\left(\frac{\partial v}{\partial x}\right)^2 = \left(\epsilon_{11}\frac{p}{\rho c_v}\right)b\left(\frac{1}{c_s}\frac{\partial v}{\partial x}\right),$$

$$\frac{\mu}{\rho c_v}\left(\frac{\partial u}{\partial y}\right)^2 = \left(\epsilon_{12}\frac{p}{\rho c_v}\right)b\left(\frac{1}{c_s}\frac{\partial u}{\partial y}\right),$$

$$\frac{2\lambda}{\rho c_v}\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} = \frac{p}{\rho c_v}\left[\frac{\epsilon_{13}b}{c_s}\frac{\partial u}{\partial x} + \frac{\epsilon_{14}b}{c_s}\frac{\partial v}{\partial y}\right],$$

$$\frac{2\mu}{\rho c_v}\frac{\partial v}{\partial x}\frac{\partial u}{\partial y} = \frac{p}{\rho c_v}\left[\frac{\epsilon_{15}b}{c_s}\frac{\partial v}{\partial x} + \frac{\epsilon_{16}b}{c_s}\frac{\partial u}{\partial y}\right].$$

Thus, the solution $V(t,x,y)$ of Eq. (8) also solves the following equation in a neighborhood of (t_0, x_0, y_0),

$$\frac{\partial W}{\partial t} = -A\frac{\partial W}{\partial x} - B\frac{\partial W}{\partial y} + F_\epsilon(t,x,y). \tag{10}$$

As in Eq. (7), F_ϵ does not depend on W because F was evaluated in terms of the known solution V. The following theorem can now be stated.

Theorem 4: Let $\{\epsilon_i\}$ be as defined above. Let

$$\epsilon^* = \max\left[\frac{|\epsilon_i|}{c_s} : i \neq 7, 8;\ \frac{\rho c_v}{p}|\epsilon_i| : i = 7, 8\right],$$

and let

$$\epsilon = \epsilon^* \max\left[3,\ \frac{2}{P_r^*} + 8b\delta\right],$$

where

$$\delta = \frac{1}{c_s}\max\left[\left|\frac{\partial u}{\partial x}\right|,\ \left|\frac{\partial v}{\partial x}\right|,\ \left|\frac{\partial u}{\partial y}\right|,\ \left|\frac{\partial v}{\partial y}\right|\right].$$

Then, in a neighborhood of (t_0, x_0, y_0) the viscous terms constitute an ϵ perturbation of the equation $\partial W/\partial t = -A(\partial W/\partial x) - B(\partial W/\partial y)$.

Proof: One needs only to note that the normalized variables can be taken as $\tilde{W} = \Gamma W$, where

$$\Gamma = \operatorname{diag}\left[\frac{1}{\rho\beta_1},\ \frac{1}{c_s},\ \frac{1}{c_s},\ \frac{\rho c_v}{p\beta_2}\right].$$

IV. APPLICATIONS

The previous analysis provides a technique for estimating the stability parameter of an asymptotically stable solution to the compressible Navier–Stokes equations. In order to apply the results, one needs to examine the given solution over distances of one mean-free path. This examination produces an ϵ, such that a linear approximation to the flow variables and their derivatives is accurate to within this ϵ. By Theorem 4, the viscous terms roughly provide an ϵ perturbation of the inviscid equations. By Theorem 2, it is then possible to obtain a lower bound on the stability parameter.

It is perhaps worthwhile emphasizing that by measuring over distances of a mean-free path, the equations simplify in form in a rather remarkable fashion. That is, Eq. (10) is exact in the neighborhood of the base point and indicates that if the ϵ_i are small, then the viscous terms are small.

There are two areas where versions of the compressible Navier–Stokes equations have been studied extensively, namely shock wave structure and boundary layers. In general, these solutions exhibit steep gradients (significant variation occur over small distances) and also exhibit pronounced viscous effects. Such solutions are in agreement with the results of the present paper. The ϵ_i of Eq. (10) will be large. However, there are also solutions in which viscous effects are important, but in which significant variations occur only over large distances. One such case, that of weak shocks, will be analyzed in detail in order to see how such a situation is compatible with our results. As an application of the analysis, solutions obtained from the incompressible equations are considered as possible candidates for approximations to asymptotically stable solutions of the compressible equations.

A. Shock wave structure equations

It is well known that the thickness of strong shock waves is on the order of a few mean-free paths (see, for example, Ref. 10). Before considering a full set of equations, it is of interest to examine the case of a single equation in order to see directly how the analysis proceeds. The pertinent equation is the following[11]:

$$\rho u \frac{du}{dx} - \mu \frac{d^2u}{dx^2} + \frac{dp}{dx} = 0,$$

where, from the continuity equation, $\rho u = \text{const}$.

After an integration, one obtains

$$\rho \left[u^2 + \frac{1}{\rho} p - \frac{\mu}{\rho} \frac{du}{dx} \right] = \text{const}.$$

Suppose now that, in the vicinity of the shock, u and du/dx can be approximated to within ϵ over distances less than or equal to τ, where $\tau =$ the mean-free path $= \mu/\rho c_s$, i.e., in terms of Lemma 1,

$$|u(x+\tau) - u(x)| \leqslant \epsilon,$$

$$\left| \frac{u(x+\tau) - u(x)}{\tau} - \frac{du}{dx}(x) \right| \leqslant \epsilon.$$

Then, from Lemma 1 $du/dx = \epsilon_1(x)/\tau$, where $|\epsilon_1| \leqslant \epsilon(1+\tau)$. Thus,

$$\frac{\mu}{\rho} \frac{du}{dx} = \epsilon_1 c_s. \tag{11}$$

Note that the above analysis can always be made. However, interesting conclusions, regarding the relative importance of the viscous terms, can only be obtained in the case that ϵ_1 is small. If ϵ_1 is large, as would be the case for a strong shock, Eq. (11) produces no useful information.

Consider next the one-dimensional steady-state shock structure equations for the case of a Prandtl number $= \frac{3}{4}$. The equations are,[10]

$$\rho v = m = \text{const},$$

$$p + mv - (2\mu + \lambda) v' = \text{const} = c_1,$$

$$mc_v T - kT' + c_1 v - mv^2/2 = \text{const},$$

where $(\ )' = (d/d\xi)(\ )$, $\xi = u_w t - x$, $u_w =$ wave velocity $= \text{const}$, and $v = u_w - u$.

If the Prandtl number $= \frac{3}{4}[(2\mu + \lambda)/k] c_p = \frac{3}{4}$, then the solution can be written in the form,

$$\begin{aligned} (2\mu + \lambda) v' &= \frac{m(\gamma+1)}{2\gamma v} (v - v_{-\infty})(v - v_{+\infty}), \\ kT' &= -2mc_p^2 \left(\frac{\gamma+1}{\gamma}\right) \left(\frac{T_{-\infty} - T}{v + v_{-\infty}}\right) \left(\frac{T_{+\infty} - T}{v + v_{+\infty}}\right), \end{aligned} \tag{12}$$

where $\pm\infty$ denotes the value at $x = \pm\infty$. One can now demonstrate analytically, without using Lemma 1, that estimates of the form of Eq. (6) are valid.

Assume now that the solution of Eq. (12) is the steady-state limit of an asymptotically ϵ stable solution to Eq. (5), where ϵ is given in Theorem 3. Then, Eq. (5) has a solution $\tilde{V}(t,x)$ valid for $0 \leqslant t \leqslant \infty$ and $x(t) \leqslant x \leqslant \infty$. Suppose this solution to be generated by boundary conditions simulating a piston moving into a stationary fluid with velocity u_0 [12]; thus, $x(t) = x(0) + u_0 t$. Also, $\lim_{t\to\infty} \tilde{V}(t,x) =$ solution of Eq. (12) in some neighborhood of the moving shock wave. Next consider the inviscid equations with the same boundary conditions and with initial data $= \tilde{V}(t,x)$ for any t. Under these conditions it is known that the inviscid equations will have a solution, $V(t,x)$, which in a relatively short period of time will develop a discontinuity. Thus, for t large, $\max_x \|\tilde{V}(t,x) - V(t,x)\| \geqslant |u_0/c_s - 0| \geqslant 1$.

From Theorem 2 one can now calculate a lower bound for the stability parameter λ, namely $\lambda \geqslant 1/\epsilon$. In the case of a weak shock, ϵ is small or λ is large. This indicates that the solution is ill-conditioned: such a conclusion is in agreement with the physical behavior of a weak shock. In the case of a strong shock, where ϵ is large, one can only conclude that λ is larger than a small number. Thus, although the analysis remains valid, it produces no useful information.

B. Solutions of the incompressible equations

It should be noted that the results of this paper are not valid for the incompressible equations. The estimate of the viscous terms according to Lemma 1 can still be accomplished. However, the inviscid time-dependent incompressible equations are not hyperbolic and thus the normalized variables and Theorem 1 would not be applicable. There exist many calculations, with the incompressible Navier–Stokes equations, that exhibit viscous effects in regions where significant variations occur over many mean-free paths. Results of the present paper indicate that for the compressible case such solutions could be asymptotically stable only if the stability parameter were very large. This would indicate that the solution is ill-conditioned. (In fact, many of these solutions appear to be such that small perturbations of initial data produce large variations at later times.)

As a specific example, consider the case of water flow. In the study of Ref. 13, the following properties were chosen:

$p = p_0 + c_s^2(\rho - \rho_0)$: equation of state,

$c_s = 4820$ ft/sec: speed of sound,

$$\mu = 7 \times 10^{-4}\,\frac{\text{lb. mass}}{\text{ft. sec}},$$

$$\lambda = -(2/3)\,\mu,$$

$$\rho_0 = 62.4\,\frac{\text{lb. mass}}{\text{ft.}^3}.$$

With these assumptions, the energy equation is uncoupled from the system. The mean-free path is calculated as,

$$\tau = \frac{\frac{4}{3}\mu}{\rho c_s} \cong 3.1 \times 10^{-9}\ \text{ft.}$$

In Ref. 13, the problem was that of flow through a tube of radius $R = 0.02$ ft. The intention was to produce relatively smooth profiles, and it is immediately clear that significant variations would occur only over thousands of mean-free paths.

The incompressible equations have been extensively used to model water flow in a tube. The classical Poiseuille flow gives a parabolic velocity distribution with at least the qualitative behavior (namely, curved) that is expected

Suppose now that Poiseuille flow is a good approximation to an asymptotically stable solution of the compressible Navier–Stokes equations. In particular, the following is assumed:

There exists a solution of Eq. (8), $\tilde{V}(t, r, z)$, valid for $0 \leqslant t \leqslant \infty$, $0 \leqslant r \leqslant R$, $0 \leqslant z \leqslant \infty$. The solution is determined by inflow conditions at $z = 0$, a nonslip boundary condition at the wall, and $\tilde{V}(0, r, z) = 0.$[13] For $t \geqslant t_0$ let $\epsilon(t_0)$ be the estimate for the viscous terms for $\tilde{V}(t, r, z)$ given by Theorem 4. Let ϵ be the estimate for the viscous terms for the Poiseuille solution given by Theorem 4. Then, for t_0 sufficiently large and $z > z^*$, for some z^*, $\epsilon(t_0) \cong \epsilon$.

Remark: Transformation to cylindrical coordinates does not affect the estimate of Theorem 2.

As noted above, ϵ will be small. Next, assume that $\tilde{V}(t, r, z)$ is asymptotically ϵ stable. Then, the inviscid equations must have a solution $V(t, r, z)$ satisfying the same boundary conditions as $\tilde{V}(t, r, z)$ and having initial conditions $V(0, r, z) = \tilde{V}(t_0, r, z)$ for any t_0. It is expected[13] that $V(t, r, z)$ will approach "plug" flow for t large. Whether or not the flow actually becomes discontinous, one can conclude that $\|\tilde{V}(t, r, z) - V(t, r, z)\| \geqslant |\delta(t) - u_{\max}|$ where $\lim_{t\to\infty} \delta(t) = 0$ and $u_{\max} =$ maximum velocity. Thus, from Theorem 2 the stability parameter λ must be such that $\lambda \geqslant |u_{\max} - \delta(t)|/\epsilon$. Because ϵ is very small, one must conclude that the solution is highly ill-conditioned.

There is experimental evidence to indicate the existence of pipe flow with a curved but relatively smooth profile (having stability parameter much less than that given above). The problem is to find a theoretical model that can predict such profiles while maintaining appropriate stability properties. The results of the present paper indicate the following alternatives:

(a) Either the inviscid compressible equations can predict these kinds of profiles (in which case the solution would not tend to plug flow, as postulated above), or

(b) A mathematical model different from the time-dependent compressible Navier–Stokes equations is required.

It should also be made clear that it is not possible to conclude, from the above analysis, that Poiseuille flow gives incorrect steady-state values. Rather, one concludes that Poiseuille flow cannot be the steady-state limit of a well-conditioned solution to the time-dependent equations.

As noted earlier, there exist many calculations, with the incompressible Navier–Stokes equations, in which the viscous effects are pronounced and yet significant variations occur only over many mean-free paths. The analysis of the present paper indicates that the use of the time-dependent incompressible Navier–Stokes equations in these situations is totally unjustified. However, it is conceivable that the solutions to the steady-state equations, perhaps fortuitously, have physical significance.

It is of interest to note that even in the capillary tubes studied by Poiseuille our results indicate that the viscous terms of the compressible equations are small. For, the smallest tube that Poiseuille considered was with radius $R = 0.0015$ cm.[14,15] or $R \sim 15{,}600$ mean-free paths. Thus, the variation of the flow variables over one mean-free path would still be quite small.

The following, in order to complete the discussion of the water flow problem, considers the possibility of viscous effects in the energy equation. P_r^* and b of Eq. (9) are needed. Let

$c_v = 1$ BTU/lb. mass °R,

$$k = 2.0 \times 10^{-4}\,\frac{\text{BTU}}{\text{ft. sec }°R},$$

Pressure = 1 atmosphere,
or $p_0 = (2116)\ (\text{g}) = (2116.0)(32.2)$.

Then,

$$P_r^* = \frac{4}{3}\,\frac{\mu c_v}{k} = \frac{4}{3}\,\frac{(7.0 \times 10^{-4})(1)}{2.0 \times 10^{-4}} \cong 4.67$$

and

$$b = \frac{\rho c_s^2}{p} \cong \frac{(62.4)(4820)^2}{(2116.0)(32.2)} \cong 2.13 \times 10^{+4}.$$

In the problem of Ref. 4, the maximum velocity was 1 ft./sec,

$$\left|\frac{u}{c_s}\, b\right| \cong \left|\frac{b}{c_s}\right| \cong \frac{2.13\times 10^{+4}}{0.482\times 10^{+4}} \cong 4.4.$$

Thus, for both the one-dimensional and two-dimensional cases, the effect of the viscous terms is again small.

APPENDIX: THE COMPRESSIBLE NAVIER-STOKES EQUATIONS WITH THE SECOND-ORDER TERMS EXPRESSED AS FIRST-ORDER TERMS

The purpose of the Appendix is to describe a procedure for expressing the second-order viscous terms as first-order terms. Such a procedure would seem to be more "natural," and, as noted in the Introduction, is intuitively more satisfactory. That is, the effect of the viscous terms is more directly displayed by Eq. (15) than Eq. (7); the viscous terms are displayed as first-order terms in Eq. (15) and as forcing terms in Eq. (7).

The second-order terms are expressed as first-order terms by applying the following lemma.

Lemma 2: Let $f(x)$ be differentiable in an interval I of the real axis.

For any $\Delta > 0$ associate $\epsilon > 0$ such that the following conditions hold whenever $|\Delta x| \leq \Delta$ and $x + |\Delta x|$ in I:

(a) $|f(x+\Delta x) - f(x)| \leq \epsilon |f(x)|$,

(b) $\left|\dfrac{f(x+\Delta x)-f(x)}{\Delta x} - \dfrac{df}{dx}(x)\right| \leq \epsilon \left|\dfrac{df}{dx}(x)\right|$.

Then,

$$\frac{df}{dx} = \frac{\epsilon_1(x)}{\Delta} f(x),$$

where $|\epsilon_1(x)| \leq \epsilon/(1-\epsilon)$.

Proof: From (b),

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} - \frac{df}{dx}(x) = \tilde{\epsilon}(x,\Delta x)\frac{df}{dx}(x),$$

where $|\tilde{\epsilon}| \leq \epsilon$, or,

$$\frac{df}{dx}(x) = \frac{f(x+\Delta x)-f(x)}{\Delta x(1+\tilde{\epsilon})}.$$

From (a),

$$\frac{df}{dx}(x) = \frac{\tilde{\tilde{\epsilon}}(x,\Delta x)}{\Delta x(1+\tilde{\epsilon})} f(x),$$

where $|\tilde{\tilde{\epsilon}}| \leq \epsilon$.

Thus,

$$\left|\frac{df}{dx}\right| \leq \frac{\epsilon}{\Delta(1-\epsilon)}|f(x)|, \quad \text{or} \quad \frac{df}{dx} = \frac{\epsilon_1}{\Delta} f(x),$$

where $|\epsilon_1| \leq \epsilon/(1-\epsilon)$.

The vector F of Eq. (5), for the case of variable viscous coefficients, becomes

$$F = \begin{pmatrix} 0 \\ \dfrac{1}{\rho}\dfrac{\partial}{\partial x}\left((2\mu+\lambda)\dfrac{\partial u}{\partial x}\right) \\ \dfrac{2\mu+\lambda}{\rho c_v}\left(\dfrac{\partial u}{\partial x}\right)^2 + \dfrac{1}{\rho c_v}\dfrac{\partial}{\partial x}\left(k\dfrac{\partial T}{\partial x}\right) \end{pmatrix}. \tag{13}$$

Suppose a twice differentiable solution, $V(t,x)$, to Eq. (5) is given. Choose a base point (t_0, x_0). Let $\tau = \Delta$ and let I be the interval $(x_0 - \tau, x_0 + \tau)$. Applying Lemma 2 successively to the case where $f = (2\mu+\lambda)(\partial u/\partial x)$, $f = k(\partial T/\partial x)$, and $f = u$, one obtains,

$$\begin{aligned} &\frac{\partial}{\partial x}\left((2\mu+\lambda)\frac{\partial u}{\partial x}\right) = \frac{\epsilon_1}{\tau}(2\mu+\lambda)\frac{\partial u}{\partial x}, \\ &\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) = \frac{\epsilon_2}{\tau} k\frac{\partial T}{\partial x}, \quad \frac{\partial u}{\partial x} = \frac{\epsilon_3}{\tau} u. \end{aligned} \tag{14}$$

Substituting these terms into the modified Eq. (5), one obtains,

$$\frac{\partial W}{\partial t} = -(A^*)\frac{\partial W}{\partial x}, \tag{15}$$

where

$$A^* = \begin{pmatrix} u & \rho & 0 \\ \dfrac{1}{\rho}\dfrac{\partial p}{\partial \rho} & c_s\left(\dfrac{u}{c_s} - \epsilon_1\right) & \dfrac{1}{\rho}\dfrac{\partial p}{\partial T} \\ 0 & \dfrac{p}{\rho c_v}\left(1 - \epsilon_3 b\dfrac{u}{c_s}\right) & c_s\left(\dfrac{u}{c_s} - \dfrac{\epsilon_2}{P_r^*}\right) \end{pmatrix}.$$

Note that Eqs. (14) are exact expressions. By continuity, it can be assumed that $\epsilon_i = \epsilon_i(t,x)$. Thus, if Eq. (15) has a solution, $\tilde{V}(t,x)$, with initial data at t_0 taken from the given solution, then in some neighborhood of (t_0, x_0), $\tilde{V}(t,x) = V(t,x)$.

Note that the matrix A^* of Eq. (15) is written in terms of normalized variables. It remains to show that this matrix is an ϵ perturbation, for some ϵ of the same magnitude as the $\{\epsilon_i\}$, of the matrix A^*. This can be established in a straightforward fashion, although the algebra becomes tedious (bounds can be obtained on the perturbed eigenvalues and the perturbed diagonalizing matrix).

For the two-dimensional case, vector F of Eq. (8) becomes,

$$\begin{aligned} F_2 &= \frac{1}{\rho}\frac{\partial}{\partial x}\left((2\mu+\lambda)\frac{\partial u}{\partial x}\right) + \frac{1}{\rho}\frac{\partial}{\partial x}\left(\lambda\frac{\partial v}{\partial y}\right) \\ &\quad + \frac{1}{\rho}\frac{\partial}{\partial y}\left[\mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)\right], \\ F_3 &= \frac{1}{\rho}\frac{\partial}{\partial x}\left[\mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)\right] + \frac{1}{\rho}\frac{\partial}{\partial y}\left(\lambda\frac{\partial u}{\partial x}\right) \\ &\quad + \frac{1}{\rho}\frac{\partial}{\partial y}\left((2\mu+\lambda)\frac{\partial v}{\partial y}\right), \\ F_4 &= \frac{1}{\rho c_v}\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) + \frac{1}{\rho c_v}\frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right) \\ &\quad + \frac{\lambda}{\rho c_v}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)^2 + \frac{\mu}{\rho c_v}\left[2\left(\frac{\partial u}{\partial x}\right)^2 \right. \\ &\quad \left. + 2\left(\frac{\partial v}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2\right]. \end{aligned} \tag{16}$$

The same argument can be carried out, except for the terms which arise from the tangential shear term in F_4 [that is, $(\partial v/\partial x)^2 + (\partial u/\partial y)^2$]. It turns out, because of the two equal eigenvalues in A and B of Eq. (8), that with these terms not identically zero, the resulting matrices cannot in general be diagonalized. Thus, with the possible exception of the tangential shear term in F_4, the viscous terms provide an ϵ perturbation of the matrices A and B.

Finally, one other difficulty arises in the use of Lemma 2, namely the use of a relative error. In regions where $f(x)$ or df/dx is small, this condition will not be meaningful.

[1] J.O. Hirschfelder, C.F. Curtiss, and R.B. Bird, *Molecular Theory of Gases and Liquids* (Wiley, New York, 1954), pp. 19, 695.

[2] A.S. Iberall, "Contributions Toward Solutions of the Equations of Hydrodynamics; Part A: The Continuum Limitations of Fluid Dynamics," prepared for Office of Naval Research, Washington, D.C. Contract No. NONR 34-5 (00), task NR 062-264, p. 265.

[3] F.S. Sherman, Annu. Rev. Fluid Mech. **1**, 319 (1969).

[4] R. Courant, Commun. Pure Appl. Math. **XIV**, 257—265 (1961).

[5] R. Courant and P. Lax, Commun. Pure Appl. Math. **II**, 255—274 (1949).

[6] George M. Homsy, J. Fluid Mech. **62**, 387—403 (1974).

[7] P. Gordon, SIAM J. Appl. Math. **30**, 391—401 (1976).

[8] Sydney Goldstein, *Lectures on Fluid Mechanics* (Interscience, New York, 1960), p. 95.

[9] Ref. 2, p. 8.

[10] M. Morduchow and A.P. Libby, J. Aeronaut. Sci. **16**, 674—84 (1949).

[11] P.A. Thompson, *Compressible Fluid Dynamics* (McGraw-Hill, New York, 1972), p. 362.

[12] S.M. Scala and P. Gordon, Phys. Fluids **9**, 1158—66 (1966).

[13] P. Gordon and S.M. Scala, "Nonlinear Theory of Pulsatile Blood Flow Through Viscoelastic Blood Vessels," Proceedings of the AGARD Specialists' Meeting of Fluid Dynamics of Blood Circulation and Respiratory Flow, May 1970, Naples, Italy, Technical Editing and Reproduction Ltd., London, 1970.

[14] Horace Lamb, *Hydrodynamics* (Cambridge U.P., Cambridge, 1930), Fifth ed.

[15] Jean Poiseuille, Comptes Rendus **11**, 961 (1840); **12**, 112 (1841).

The Fermi method of quantizing the electromagnetic field as a model for quantum field theory

A. L. Carey and C. A. Hurst

Department of Mathematical Physics, University of Adelaide, S. Australia 5001

(Received 20 November 1976)

In a previous paper we demonstrated that the Fermi method for quantizing the electromagnetic field had a rigorous C^*-algebra version. Here we investigate some properties of our formalism and show (a) that the Fermi method provides an example of spontaneous symmetry breaking in a quantum field theory, (b) that it raises some interesting questions about automorphisms of Weyl systems, and (c) that it provides a prototype for higher-spin zero-mass field theories.

INTRODUCTION

Recent work has demonstrated that the three procedures of heuristic field theory, Gupta–Bleuler, radiation gauge and the Fermi method, all have rigorous versions,[1–3] each of which from the viewpoint of Wightman field theory,[4] suffers some defect. The most highly developed of these, the indefinite metric formalism,[1] has a clearly stated axiomatic formulation[5] which makes evident the departure from the usual field theory axioms.[4] However Segal has argued[2] that his radiation gauge proposal (which appears not to have received the attention it deserves) represents from a physical viewpoint the minimal necessary modification of massive field theory. Most of the difficulties of zero mass theories can in fact be traced to the insistence on a field which transforms in a manifestly covariant way. Our modern understanding of relativistic invariance,[6,7] suggests that this demand stems from historical prejudice (albeit a prejudice with material benefits). Segal's scheme retains Poincaré invariance while dropping the requirement of manifest covariance. In a previous paper[3] we revived the Fermi method (to which we refer the reader for a detailed discussion, see also however Refs. 8 and 9) and showed that, in a sense, it is the natural generalization of the quantization methods which have been used for the scalar fields (as distinct from the ad hoc character of Gupta–Bleuler). Furthermore, it is important to emphasize that in Ref. 3 we were able to show that the radiation gauge theory of Segal could be deduced from the Fermi method.

Since the Fermi method attempts to retain manifest covariance it deviates from a Wightman theory[4] in other respects. These peculiar features of the Fermi method were discussed in part in Ref. 3 and it is the purpose of this paper to explore them in more detail. We begin in Sec. 1 with a summary of Ref. 3, noting in particular that the Fock representation of the CCR algebra for the vector potential (which provides the rigorous version of the Fermi method) contains a noninvariant vacuum. To overcome this we introduced in Ref. 3 a covariant representation for the Fermi method and in Sec. 2 of this paper we show to what extent this covariant representation is a model of spontaneously broken symmetry.

An important question in the C^*-algebra approach to field theory is whether certain automorphisms of the algebra of observables (in particular the time evolution) are in some sense inner. Since the algebra of observables for the electromagnetic field is "smaller" than the Fermi CCR algebra (which is the algebra for the full four-component vector potential) we need to determine how the time evolution "restricts" to the observable algebra. This problem is tackled in Sec. 3 and the analysis extended to all automorphisms of the Fermi CCR algebra induced by the Poincaré group. This makes contact with more general questions on automorphisms of the Segal/Manuceau[10] CCR algebra. Part of this problem is to determine the behavior of the radiation gauge states (for the one-particle theory) under Lorentz transformations, a question previously investigated by Bender.[11] Our approach, being adapted specifically to the Weyl algebra formalism, differs from his in that we do not write down the transformation law for radiation gauge explicitly. Rather, we write it in a form which enables us to demonstrate that the Poincaré transformations for radiation gauge actually define automorphisms of the whole Fermi CCR algebra.

Finally in Sec. 4, we prove a theorem which constrains the Weyl algebra formalism for zero mass theories, showing in particular that there is a Fermi method only for certain higher spin tensor fields.

1. THE FERMI METHOD

We begin with the following space of real valued solutions of the wave equation. Let $\mathcal{S}(R^4, R^4)$ be the Schwartz space of R^4 valued C^∞ functions F of fast decrease on Minkowski space. Let D denote the Green's function of the wave operator, that is,

$$\Box D = 0,$$

$$D(\mathbf{x}, 0) = 0, \quad \frac{\partial}{\partial t} D(\mathbf{x}, t)\Big|_{t=0} = -\delta(\mathbf{x}).$$

Define the space M_0 to consist of all functions ϕ on Minkowski space taking values in R^4, of the form $F * D$ where $F \in \mathcal{S}(R^4, R^4)$. (Here convolution is defined componentwise $\phi_\mu = F_\mu * D$, $\mu = 0, 1, 2, 3.$) Then M_0 is a space of smooth real valued solutions of the wave equation. We can define a symplectic form on M_0 by

$$B(\phi, \phi') = -\int d^3\mathbf{x}[\phi^\mu(x)\dot{\phi}'_\mu(x) - \dot{\phi}^\mu(x)\phi'_\mu(x)]. \qquad (1.1)$$

The Poincaré group acts on M_0 by

$$(a, \Lambda)\phi(x) = \Lambda\phi(\Lambda^{-1}(x - a)), \qquad (1.2)$$

where a is a translation and Λ a Lorentz transformation. The symplectic form B is well known to be invariant under these transformations. Further, M_0 is also invariant.

In Ref. 3 a rigorous, C^*-algebra version of the Fermi method for quantizing the electromagnetic vector potential was presented. We summarize the essential features of that paper.

Every element of M_0 may be written as

$$\phi_\mu(\mathbf{x}, t) = \frac{1}{(2\pi)^{3/2}} \int \frac{d^3\mathbf{k}}{2k} \{\hat{\phi}_\mu(\mathbf{k}) \exp[-i(kt - \mathbf{k}\cdot\mathbf{x})] + \hat{\phi}_\mu(\mathbf{k})^* \exp[i(kt - \mathbf{k}\cdot\mathbf{x})]\}, \tag{1.3}$$

where $k = |\mathbf{k}|$ and $\hat{\phi}_\mu$ is a C-valued function on

$$X_0^+ = \{k \in R^4 \,|\, k_0 > 0,\ k_0^2 = \mathbf{k}^2\}.$$

The C^4-valued functions $\hat{\phi}$ form a complex pre-Hilbert space with inner product

$$\langle \hat{\phi}, \hat{\phi}' \rangle_F = \int \frac{d^3\mathbf{k}}{k} [\hat{\boldsymbol{\phi}}(\mathbf{k})^* \hat{\boldsymbol{\phi}}'(\mathbf{k}) + \hat{\phi}_0(\mathbf{k}) \hat{\phi}_0'(\mathbf{k})^*]. \tag{1.4}$$

(For the zeroth component the complex conjugation appears on the second variable because the form is sesquilinear with respect to J_F.) This inner product arises by modifying the complex structure on $\hat{M}_0$ from the Lorentz invariant one of multiplication by i to the structure J_F given by

$$(J_F \phi)(k) = -ig\phi(k) \tag{1.5}$$

[$g = \mathrm{diag}(-1, -1, -1, 1)$ is the metric tensor and we write

$$\phi(\mathbf{k}) = (\boldsymbol{\phi}(\mathbf{k}), \phi_0(\mathbf{k}))].$$

Now J_F does not commute with Lorentz transformations and so the inner product (1.4) is not invariant. Nevertheless we have, after transferring B across to the momentum space picture,

$$\langle \hat{\phi}, \hat{\phi}' \rangle_F = B(\hat{\phi}, J_F \hat{\phi}') + iB(\hat{\phi}, \hat{\phi}'). \tag{1.6}$$

We complete $\hat{M}_0$ in the norm defined by (1.4) to give a Hilbert space M, say. Henceforth we will drop the $\hat{\phi}$ notation and write ϕ for the elements of M.

By (1.6), B is the imaginary part of the scalar product (1.4) and so we are in a position to apply Manuceau's variant[10] of Segal's Weyl algebra formalism.[12] We denote by $\Delta_c(M)$ the C^*-algebra of the canonical commutation relations over (M, B) and refer the reader to Refs. 3, 10, or 13 for its precise definition. A dense subspace of $\Delta_c(M)$ is spanned by the functions δ_ϕ on M where

$\delta_\phi(\phi') = 1$ if $\phi = \phi'$ and zero otherwise.

Multiplication of the functions δ_ϕ is given by

$$\delta_\phi \delta_{\phi'} = \delta_{\phi+\phi'} \exp[-(i/2)B(\phi, \phi')].$$

In the usual way,[12,3] we consider generating functionals $\rho: M \to C$ such that

(i) $\rho(0) = 1$,

(ii) $\rho(\lambda\phi + \phi')$ is continuous in $\lambda \in R$ for all $\phi, \phi' \in M$,

(iii) for every finite sequence $\{c_j\}$ in C and $\{\phi_j\}$ in M,

$$\sum_{jk} \bar{c}_j c_k \rho(\phi_j - \phi_k) \exp[-(i/2)B(\phi_j, \phi_k)] \geq 0.$$

Then define states E_ρ on $\Delta_c(M)$ by $\sum_i \lambda_i \delta_{\phi_i} \to \rho(\sum_i \lambda_i \phi_i)$ to give via the GNS construction, a cyclic representation π_ρ of $\Delta_c(M)$ with cyclic vector Ω such that

$$\rho(A) = \langle \Omega, \pi_\rho(A)\Omega \rangle$$

for all $A \in \Delta_c(M)$.

The Fock representation of $\Delta_c(M)$ is given by the generating functional $\rho(\phi) = \exp(-\frac{1}{4}\|\phi\|_F^2)$, which is clearly not invariant under pure Lorentz transformations (i.e., boosts) because of the noninvariance of $\|\ \|_F$. Thus the Fock representation contains a noninvariant vacuum (the cyclic vector Ω given by the GNS construction). Furthermore, one can show[3] that only the subgroup H of the Poincaré group consisting of translations and spatial rotations is unitarily implementable in the Fock representation.

To overcome this difficulty we introduced in Ref. 3 a new covariant representation of $\Delta_c(M)$. Before describing it we digress on a technical point. It will be convenient to use both the C^*-algebra $\Delta_c(M)$ introduced by Manuceau[10] and that of Segal,[12] in what follows. The connection between them is found by noting that if $\mathcal{J}$ is the family of all complex linear finite dimensional subspaces of M on which B is nondegenerate, then $\Delta_c(M)$ is the inductive limit of the C^*-algebras $\Delta_c(F)$, $F \in \mathcal{J}$ where F is equipped with the symplectic form obtained by restriction of B.[10] In Segal's formulation the C^*-algebra of the CCR's $\mathcal{W}(M, B)$ is defined as the C^*-inductive limit of the algebras $\mathcal{W}(F, B)$, $F \in \mathcal{J}$, where $\mathcal{W}(F, B)$ is the weak closure of $\Delta_c(F)$ in the weak topology of the Schrödinger representation. Clearly therefore, there is a continuous injection of $\Delta_c(M)$ into $\mathcal{W}(M, B)$ and, using the characterization of physical representations by generating functionals, there is a 1—1 correspondence between the physical representations of each algebra.

If (a, Λ) is a Poincaré transformation then we denote by $\theta(a, \Lambda)$ the corresponding automorphism of $\Delta_c(M)$ defined via

$$\theta(a, \Lambda)\delta_\phi = \delta_{(a,\Lambda)\phi}.$$

If we inject $\Delta_c(M)$ into $\mathcal{W}(M, B)$ then we will denote the elements in $\mathcal{W}(M, B)$ corresponding to the δ_ϕ by $W(\phi)$. Thus $\theta(a, \Lambda)$ is the automorphism of $\mathcal{W}(M, B)$ given by

$$\theta(a, \Lambda)W(\phi) = W((a, \Lambda)\phi).$$

The norm given by (1.4) is invariant under the action of H and these transformations are unitarily implemented in the Fock representation, however the Lorentz boosts cannot be unitarily implemented. We therefore introduce the homogeneous space,

$$X = \{p \in R^4 \,|\, p_0^2 - \mathbf{p}^2 = 1,\ p_0 > 0\}$$

of the Lorentz group. Let $p \to \Lambda(p)$ be a map which assigns to each $p \in X$ Lorentz transformation $\Lambda(p)$ mapping $\hat{p} = (0, 0, 0, 1)$ to p. [Note that we write four vectors as $(\mathbf{p}, p_0)$.] We choose $p \to \Lambda(p)$ to be measurable with respect to the Lorentz invariant measure $d\mu$ on X.

Let ρ_p be the generating functional

$$\rho_p(\phi)=\exp(-\tfrac{1}{4}\|\Lambda(p)^{-1}\phi\|_F^2).$$

[ρ_p satisfies (i) and (ii) trivially while (iii) follows from the invariance of B under Poincaré transformations.] Then using the GNS construction there is a Hilbert space K_p and a representation π_p of $\Delta_c(M)$ and $\mathcal{W}(M, B)$ in K_p with cyclic vector Ω_p such that

$$\rho_p(A)=\langle\Omega_p, \pi_p(A)\Omega_p\rangle$$

for all $A\in\Delta_c(M)$ or $\mathcal{W}(M, B)$. We will write $W_p(\phi)$ for $\pi_p(\delta_\phi)$. We note that the representations π_p are all inequivalent [this follows from the fact that the automorphisms $\theta(0,\Lambda(p))$ are not implementable in the Fock representation, see Shale[14]].

We form the direct integral

$$K=\int_X K_p\, d\mu(p)$$

of these Hilbert spaces. We note that in the Hilbert space K_p the translation subgroup of the Poincare group can be unitarily implemented, while a Lorentz transformation is unitarily implemented if and only if $\Lambda(p)^{-1}\Lambda\Lambda(p)$ is a spatial rotation. [This is because $\rho_p(\Lambda\phi)=\rho_p(\phi)$ if and only if

$$\|\Lambda(p)^{-1}\Lambda\phi\|=\|\Lambda(p)^{-1}\phi\|.]$$

The elements of K are equivalence classes of μ-measurable functions F on X such that $F(p)\in K_p$ and satisfying

$$\int_X\|F(p)\|_p^2\, d\mu(p)<\infty$$

where $\|\ \|_p$ denotes the norm in K_p. Following Shale,[14] we define a map $Y_{a,\Lambda}: K_p\to K_{(a,\Lambda)p}\equiv K_{\Lambda p}$ for each Poincaré transformation (a,Λ) by

$$Y_{a,\Lambda}\,\pi_p(A)\Omega_p=\pi_{\Lambda p}(\theta(a,\Lambda)A)\Omega_{\Lambda p} \tag{1.7}$$

[noting that the vectors $\pi_p(A)\Omega_p$, $A\in\mathcal{W}(M,B)$ are dense in K_p.] It follows by Theorem 6.1(b) of Shale[14] that $Y_{a,\Lambda}$ is unitary for each (a,Λ) and that

$$Y_{a,\Lambda}W_p(\phi)Y_{a,\Lambda}^{-1}=W_{\Lambda p}((a,\Lambda)\phi). \tag{1.8}$$

Now, a representation π of $\Delta_c(M)$ [and $\mathcal{W}(M,B)$] is defined in K by

$$\pi(A)F(p)=\pi_p(A)F(p) \tag{1.9}$$

while a representation U of the Poincaré group is defined by

$$(U_{a,\Lambda}F)(p)=Y_{a,\Lambda}F(\Lambda^{-1}p). \tag{1.10}$$

It was observed in Ref. 3 that there are no subspaces of K invariant under both these representations. Furthermore we have the relation

$$U_{a,\Lambda}\pi(W(\phi))U_{a,\Lambda}^{-1}=\pi(W(\theta(a,\Lambda)\phi)).$$

Thus the Poincaré transformations are unitarily implemented in this representation of $\mathcal{W}(M,B)$.

We note finally that

$$\phi,\phi'\to\langle\Lambda(p)^{-1}\phi, \Lambda(p)^{-1}\phi'\rangle_F$$

is an inner product on M which has B as its imaginary part. Thus the representations ρ_p are "Fock-like." We may think of them as being associated with the complex structure

$$\Lambda(p)J_F\Lambda(p)^{-1}.$$

2. SPONTANEOUS SYMMETRY BREAKING

The covariant representation (1.9) of the CCR algebra exhibits all the features of a spontaneously broken symmetry in a field theory—but in this case the symmetry broken is Lorentz invariance. That is, the decomposition of (1.9) into irreducibles, each with a unique vacuum, destroys the Lorentz covariance.

The representation U of the Poincaré group acting in K can be thought of as an induced representation. This is because the properties of $Y_{a,\Lambda}$ endow K with the structure of a Hilbert Poincaré bundle (in the terminology of Varadarajan[6]). That is,

$$Y_{a,\Lambda}: K_p\to K_{(a,\Lambda)p},$$

and in the fiber $K_{\hat p}$, $(a,R)\to Y_{a,R}$ is a unitary representation of the subgroup H. [This is because $\theta(a,R)$ is unitarily implemented in $K_{\hat p}$.] U is therefore the representation of the Poincaré group induced by the representation $a,R\to Y_{a,R}$ of H in $K_{\hat p}$.

The behavior of this covariant quantization of the electromagnetic potential as a spontaneously broken symmetry may be seen more clearly by analyzing the "vacuum" and "one-particle spaces". We take $\mathcal{H}_0$ to be the subspace of K defined from functions F of the form $F(p)=c(p)\Omega_p$ where $c(p)$ is a complex number for each $p\in X$,

$$\int_X|c(p)|^2\, d\mu(p)<\infty,$$

and the subspace $\mathcal{H}_1$ is taken to consist of the direct integral of the one-particle spaces in each K_p. $\mathcal{H}_1$ can be constructed directly as follows. For each $\phi\in M$ we write

$$\pi(W(\phi))=\exp iR(\phi), \tag{2.1}$$

where $R(\phi)$ is self-adjoint and may be regarded as the second quantized version of the potential ϕ. The annihilation and creation operators are defined by

$$\begin{aligned} a(\phi)&=\frac{1}{\sqrt{2}}[R(\phi)+iR(J_F\phi)],\\ a(\phi)^*&=\frac{1}{\sqrt{2}}[R(\phi)-iR(J_F\phi)]. \end{aligned} \tag{2.2}$$

The domains of these operators are the direct integrals of the domains of the corresponding operators $a_p(\phi)$, $a_p(\phi)^*$, and $R_p(\phi)$ in each K_p. Explicitly we have

$$(R(\phi)F)(p)=R_p(\phi)F(p),$$

where $\exp iR_p(\phi)=\pi_p(W(\phi))$ and $F(p)$ lies in the domain of $R_p(\phi)$ for μ-almost all $p\in X$. Similar expressions hold for $a(\phi)$ and $a(\phi)^*$.

If F is a function of the form $F(p)=c(p)\Omega_p$, $c(p)\in C$, then F defines an element of K in the domain of $R(\phi)$. The one-particle space $\mathcal{H}_1$ consists of the linear span of $\{R(\phi)F\mid F\in\mathcal{H}_0,\ \phi\in M\}$.

The various "vacuum states" Ω_c where $\Omega_c(p)=c(p)\Omega_p$ each satisfy $a(\phi)\Omega_c=0$ for all $\phi\in M$. If $c(p)$ is nonzero for μ-almost all ρ then Ω_c is a cyclic vector for the representation π. (If not, we could consider the subspace K' of K generated by $\{\pi(W(\phi))\Omega_c\mid\phi\in M\}$. Then in the central decomposition of π, K' decomposes as

$$K' = \int_X K'_p \, d\mu(p)$$

with $K'_p \subseteq K_p$. The subspaces K'_p are invariant for μ-almost all $p \in X$ under $\{W_p(\phi) \mid \phi \in M\}$ and hence are either zero or equal to K_p. Since c is nonzero μ-almost everywhere, then $K'_p = K_p$ for μ-almost all p. Thus $K' = K$ and Ω_c is cyclic.)

We note that the vacuum expectation values have the expected form for a free theory

$$\langle \Omega_c, R(\phi)\Omega_c \rangle = 0$$

and

$$\langle \Omega_c, R(\phi_1)R(\phi_2)\Omega_c \rangle = \tfrac{1}{2} \int |c(p)|^2 \times \langle \Lambda(p)^{-1}\phi_1, \Lambda(p)^{-1}\phi_2 \rangle \, d\mu(p).$$

Next we observe that the one-particle space is isomorphic to a direct integral over X of Hilbert spaces M_p where M_p is the completion of M_0 in the norm

$$\|\phi\|^2_{M_p} = B(\phi, \Lambda(p)J_F\Lambda(p)^{-1}\phi)$$

for $\phi \in M_0$. Let

$$\mathcal{M} = \int_X M_p \, d\mu(p).$$

The map between $\mathcal{M}$ and $\mathcal{H}_1$ is given by

$$R(\phi)\Omega_c \to (1/\sqrt{2})f,$$

where $f(p) = c(p)\phi$ and $\phi \in M_0$. Note that

$$\|(1/\sqrt{2})f\|^2 = \tfrac{1}{2} \int |c(p)|^2 \|\phi\|^2_{M_p} \, d\mu(p) = \|R(\phi)\Omega_c\|^2$$

so this map extends to an isometry of $\mathcal{M}$ with $\mathcal{H}_1$.

Let $X_0^+ = \{k \in R^4 \mid k^2 = 0, \ k_0 > 0\}$. We may now regard the elements of $\mathcal{H}_1$ as equivalence classes of C^4-valued functions F defined on $X \times X_0^+$ and such that

$$\int F(p,k)^* \Lambda(p)^{-1*} \Lambda(p)^{-1} F(p,k) \, d\mu(p) \, d^3\mathbf{k}/2k < \infty. \qquad (2.3)$$

The representation U leaves the spaces $\mathcal{H}_0$, $\mathcal{H}_1$ invariant for:

$$U(a,\Lambda)R(\phi)\Omega_c = R(a,\Lambda)\phi)\Omega_{(a,\Lambda)c} \in \mathcal{H}_1,$$

where $[(a,\Lambda)c](p) = c(\Lambda^{-1}p)$. Thus on the functions $p, k \to F(p,k)$ we have

$$[U(a,\Lambda)F](p,k) = \exp(ia^\mu k_\mu)\Lambda F(\Lambda^{-1}p, \Lambda^{-1}k). \qquad (2.4)$$

That this representation is unitary on $\mathcal{H}_1$ can be verified directly using (2.3). Restricting this representation to the translation subgroup we observe that the energy—momentum spectrum lies in the cone

$$V^+ = \{k \in R^4 \mid k_0 \geq 0, \ k^2 \geq 0\},$$

from which we deduce that the covariant representation π of the CCR algebra satisfies the energy—momentum spectral conditions. Finally we remark that (2.4) can be decomposed into irreducibles although the argument is tedious. One obtains a direct sum, over the four helicities of the vector potential, of a direct integral of zero mass, continuous spin representations of the Poincaré group. It is not clear what the physical meaning of this could be.

We summarize the observations of this section as a proposition.

Proposition: (i) The field operators $R(\phi)$, $\phi \in M$ are densely defined in K and are self-adjoint.

(ii) The Poincaré group is unitarily implemented, i.e.,

$$U(a,\Lambda)R(\phi)U(a,\Lambda)^{-1} = R((a,\Lambda)\phi)$$

and the spectral condition is satisfied.

(iii) The representation π of the Weyl algebra of the CCR on K is cyclic. There is an invariant "vacuum subspace" $\mathcal{H}_0$ whose elements Ω_c satisfy

$$a(\phi)\Omega_c = 0 \quad \text{for all } \phi \in M.$$

3. AUTOMORPHISMS

A general question which arises in C^*-algebra approaches is whether the automorphisms of the algebra of observables (in particular the time evolution) are in some sense "inner." In the Fermi quantization of the electromagnetic field one starts with an algebra $\Delta_c(M)$ which treats each component of the vector potential as an independent object. The physical algebra of observables is constructed as follows. Take N to be the closed subspace of M defined by the functions satisfying

$$k^\mu \phi_\mu(k) = 0$$

and T to be the closed subspace defined by the functions of the form

$$\phi_\mu(k) = k_\mu \psi(k),$$

where ψ is a C-valued function on X_0^+. We note that, with the complex structure J_F on M, N and T are *real* (but not complex) linear subspaces. Form the algebra $\Delta_c(T)$ [resp. $\Delta_c(N)$] as the C^*-algebra of $\Delta_c(M)$ generated by the elements δ_ϕ such that $\phi \in T$ (resp. $\phi \in N$). The "physical photons," being those with transverse components, form the subspace S of M defined by the conditions

$$\phi_0 = 0, \quad \mathbf{k} \cdot \boldsymbol{\phi}(\mathbf{k}) = 0.$$

S is a complex linear subspace and we associate with it the algebra $\Delta_c(S)$ in the same way as with $\Delta_c(N)$ and $\Delta_c(T)$. Clearly S is not Poincaré invariant and consequently $\Delta_c(S)$ is not invariant under the corresponding automorphisms of $\Delta_c(M)$. Since $N = S \oplus T$ (Hilbert space direct sum), the map $\sum_i \lambda_i \delta_{\phi_i} \to \sum \lambda_i \delta_{\phi'_i}$ where $\phi'_i = S$-component of $\phi'_i \in N$ is well defined. In Ref. 3 it was shown that this map extends to a C^*-algebra homomorphism of $\Delta_c(N)$ onto $\Delta_c(S)$, thus defining an algebra $\Delta_c(N)/I$ isomorphic to $\Delta_c(S)$. We interpret $\Delta_c(N)/I$ as the algebra of physical observables. The ideal I is generated by those elements $\sum_i \lambda_i \delta_{\phi_i}$ $(\phi_i \in T)$ such that $\sum \lambda_i = 0$. Any automorphism of $\Delta_c(N)$ which leaves these elements invariant defines in the obvious way an automorphism of $\Delta_c(N)/I$. The automorphisms $\theta(a,\Lambda)$ are examples of such implying that $\Delta_c(N)/I$ is Poincaré invariant.

Now the automorphisms $\theta(a,\Lambda)$ of $\Delta_c(M)$ cannot be expected to be "inner" for the physical algebra $\Delta_c(N)/I$. However we know that there exists a representation of the Poincaré group by automorphisms of $\Delta_c(N)/I$. What we shall do below is show how these automorphisms can be extended to all of $\Delta_c(M)$.

We need to distinguish the Lorentz invariant complex structure on M_0 from that which gives rise to the inner product on M. This is because it is the Lorentz invariant complex structure which arises in the analysis of automorphisms. For example, time translations are given by

$$(T_t\phi)(\mathbf{k}) = \exp(itk)\phi(\mathbf{k}).$$

Thus we define $J_s\phi = i\phi$. We have the representation (1.2) of the Poincaré group by operators which are bounded and real linear but do not commute with J_F. Thus we need to write (1.2) as

$$(a, \Lambda)\phi(k) = \exp(J_s k^\mu a_\mu)\Lambda\phi(\Lambda^{-1}k). \qquad (3.1)$$

At this point we digress to make contact with the heuristic formulation of the Fermi method. We note that in the Fock representation the generators of time translations, as defined via (3.1), is the formal operator

$$H_s = \int \frac{d^3\mathbf{k}}{2k}[a_0(\mathbf{k})^* a_0(\mathbf{k}) + \sum_j a_j(\mathbf{k})^* a_j(\mathbf{k})],$$

where the "operators" $a_\mu(k)$ are not related to the annihilation and creation operators (2.2) by simply smearing with test functions. Their precise definition is a technical matter and we will not need to go into it in this paper (see however Ref. 8). Note that H_s can nevertheless be given meaning as a quadratic form on Fock space just as in the case of the scalar field.[15]] Now H_s is not the energy of the electromagnetic field. This is given formally by

$$H_F = \int \frac{d^3k}{2k}[- a_0(\mathbf{k})^* a_0(\mathbf{k}) + \sum_j a_j(k)^* a_j(\mathbf{k})].$$

It is interesting to note that H_F is the formal generator of the second quantized version of the one-parameter group on M,

$$(T_t^F\phi)(\mathbf{k}) = \exp(J_F tk)\phi(\mathbf{k}).$$

Thus we see a further reason for introducing the Fermi complex structure J_F—it gives rise to the second quantized version of the classical expression $\mathbf{E}^2 + \mathbf{B}^2$ for the energy of the electromagnetic field.

In the Fermi norm M is a tensor product

$$C^4 \otimes L(X_0^+, d^3\mathbf{k}/k)$$

of Hilbert spaces, where C^4 has the complex structure given by the operator $-ig$. The representation (3.1) acts on this tensor product by

$$(a, \Lambda)v \otimes \psi = \Lambda v \otimes V(a, \Lambda)\psi, \quad v \in C^4$$

with $\psi \in L^2(X_0^+, d^3\mathbf{k}/k)$ and

$$[V(a, \Lambda)\psi](\mathbf{k}) = \exp(ik^\mu a_\mu)\psi(\Lambda^{-1}k).$$

Now V is a continuous unitary representation of the Poincaré group on the complex Hilbert space $L^2(X_0^+, d^3\mathbf{k}/k)$. Hence by a theorem of Nelson[16] there is a dense subspace $\mathcal{D}$ of $L^2(X_0^+, d^3\mathbf{k}/k)$ consisting of analytic vectors for the Lie algebra of the Poincaré group. If X is a generator of the Poincaré group in the representation (3.1) then we have

$$X = A_X + \tilde{X},$$

where A_X is a 4×4 matrix (in the Lie algebra of the Lorentz group) and $\tilde{X}$ is a differential operator (we include operators of multiplication by complex valued functions of $\mathbf{k}$ in $\tilde{X}$). For $\psi \in \mathcal{D}$,

$$\begin{aligned}\sum_n \frac{t^n}{n!}\|X^n(v\otimes\psi)\| &= \sum_n \frac{t^n}{n!}\|(A_X \otimes 1 + 1 \otimes \tilde{X})^n v \otimes \psi\| \\ &= \sum_n \frac{t^n}{n!}\|\sum_n \binom{n}{k} A_x^k v \otimes \tilde{X}^{n-k}\psi\| \\ &\leqslant \sum_n \frac{t^n}{n!}\sum_n \binom{n}{k}\|A_X^k v\|\,\|\tilde{X}^{n-k}\psi\| \\ &= \sum_{k=0}^\infty \frac{t^k}{k!}\|A_X^k v\|\sum_{n=k}^\infty \|\tilde{X}^{n-k}\psi\|\ \frac{t^{n-k}}{(n-k)!}\,.\end{aligned}$$

This last expression is finite by the fact that ψ is analytic. We denote by $\mathcal{D}'$ the space $\mathbb{C}^4 \otimes \mathcal{D}$. This proves:

Lemma 3.1: M contains a dense subspace $\mathcal{D}'$ of vectors which are analytic for the generators of the Poincaré group.

Lemma 3.2: The projection P_T from M onto T is the real linear operator of multiplication by the matrix function of k,

$$K(k)_{ij} = \tfrac{1}{2}k_i k_j/k^2, \quad i, j = 1, 2, 3,$$

$$K(k)_{00} = 1, \quad K(k)_{0j} = K(k)_{j0} = k_j/k, \quad j = 1, 2, 3,$$

and satisfies

$$B(P_T\phi, J_F\phi') = B(\phi, J_F P_T\phi'). \qquad (3.2)$$

Proof: Direct calculation gives the first part of the lemma while (3.2) follows because P_T is self-adjoint (as an operator on M) with respect to the real linear inner product

$$\phi, \psi \to B(\phi, J_F\psi).$$

Lemma 3.3: P_T leaves the space $\mathcal{D}'$ of analytic vectors invariant.

Proof: Let ϕ be an element of $\mathcal{D}'$. Then by Lemma 3.1 it is sufficient to consider the components of $P_T\phi$, each of which has the form

$$(P_T\phi)_\mu : k \to k_\mu \sum_\nu k_\nu \phi_\nu(\mathbf{k})/k^2.$$

Thus we can define the map

$$\Gamma\phi_\nu(k) = \gamma(k)\phi_\nu(k),$$

where $\gamma(k) = k_\mu k_\nu/k^2$ and it is now sufficient to show that $\Gamma\phi_\nu$ is analytic. Let $\tilde{X}_0$ denote that part of $\tilde{X}$ arising from the Lorentz Lie algebra. Then $\tilde{X}_0$ is a differential operator and we define

$$\gamma^n(k) = (\tilde{X}_0^n\gamma)(k), \quad n = 0, 1, 2, \cdots.$$

Because $\tilde{X}_0$ is homogeneous of degree zero a brief calculation reveals that

$$|\gamma^n(k)| \leqslant 1 \quad \text{for all } k \in X_0^+ \text{ and } n = 0, 1, 2, 3, \cdots.$$

Thus it follows that

$$\tilde{X}^n\Gamma\phi_\nu = \sum_{l=0}^n \binom{n}{l}\Gamma^{(n)}\tilde{X}^{n-l}\psi,$$

where $\Gamma^{(n)}$ denotes the operator of multiplication by $\gamma^n(k)$. Now,

$$\sum_{n=0}^\infty \frac{t^n}{n!}\|\tilde{X}^n\Gamma\phi_\nu\| \leqslant \sum_{l=0}^\infty \frac{t^l}{l!}\sum_{n=l}^\infty \frac{t^{n-l}}{(n-l)!}\|\tilde{X}^{n-l}\psi\|$$

which is finite. Hence $\Gamma\phi_\nu$ is analytic for $\tilde{X}$ completing the proof.

Remarks 3.4: The projection P_N and P_S onto the subspaces N and S respectively also leave $\mathcal{D}'$ invariant as they are given by multiplication by a matrix valued function of **k** with the same properties as $K(k)$.

Now let X be an arbitrary element of the Poincaré Lie algebra in the representation given by (3.1). Consider the operator

$$Z_X = (1 - P_T)XP_N. \tag{3.3}$$

By Lemma 3.3 and Remark 3.4 this operator has a dense set $\mathcal{D}'$ of analytic vectors. Before discussing its properties we disgress to consider the simpler case where X is a generator of the translation subgroup. In this case X is just multiplication by a scalar valued function of **k**, say $\omega(\mathbf{k})$, and so commutes with $1 - P_T$ and P_N. Hence (3.3) reduces to $Z_X = XP_S = P_S X$ as $(1 - P_T)P_N = P_S$. Now XP_S is self-adjoint as an operator on the complex Hilbert space M and so we can write, using the spectral theorem,

$$v_t\phi = \exp(j_S tXP_S)\phi. \tag{3.4}$$

Similarly we define

$$w_t\phi = \exp[J_S tX(1 - P_S)]\phi.$$

Thus we have factorized the operator $\exp(J_S tX)$ into the product $v_t w_t = w_t v_t$. This factorization corresponds to the decomposition of M into the subspace S of physical photons (one-particle) states, and its orthogonal complement.

We now establish a relation analogous to (3.4) for general X.

Lemma 3.5: The operator Z_X defined by (3.3) on $\mathcal{D}'$ is essentially self-adjoint.

Proof: We note firstly that we can write

$$(1 - P_T)XP_N = (1 - P_T)J_F(J_F X J_F)(1 - P_T)J_F$$

because

$$J_F(1 - P_T)J_F = -g(1 - P_T)g = -P_N \quad \text{and} \quad J_F^2 = -1.$$

But now by (3.2),

$$B((1 - P_T)J_F\phi, \psi) = -B(\phi, (1 - P_T)J_F\psi)$$

for all $\phi, \psi \in \mathcal{D}'$. Further, as $\exp(tJ_S X)$ is a one-parameter group of symplectic automorphisms, it follows that

$$B(J_S X\phi, \psi) = -B(\phi, J_S X\psi)$$

and hence

$$B(J_F X J_F \phi, \psi) = B(\phi, J_F X J_F \psi),$$

for all $\phi, \psi \in \mathcal{D}'$. Consequently

$$B((1 - P_T)XP_N\phi, \psi) = B(\phi, (1 - P_T)XP_N\psi).$$

for all $\phi, \psi \in \mathcal{D}'$. Hence in order to prove that Z_X is symmetric on $\mathcal{D}'$ it remains only to check that Z_X commutes with J_F. Now we recall that X leaves N and T invariant while J_F consists of multiplication by $-ig$. Thus if we write each $\phi \in \mathcal{D}'$ as

$$\phi = \phi_{T'} + \phi_S + \phi_T,$$

where ϕ_S, ϕ_T and $\phi_{T'}$ are the components of ϕ in S, T, and the orthogonal complement of N, respectively, and we use the relation $gP_T g = P_{T'}$, it is easy to check that

$$Z_X g\phi = (1 - P_T)X\phi_S$$

(note that g acts as the identity on S). Further,

$$gZ_X\phi = g(1 - P_T)X\phi_S = (1 - P_T)X\phi_S$$

as $(1 - P_T)X\phi_S \in S$. Thus Z_X and J_F commute on $\mathcal{D}'$ and Z_X is therefore symmetric. But $\mathcal{D}'$ is a dense subspace of M consisting of vectors analytic for Z_X. Hence by a theorem of Nelson[16] Z_X is essentially self-adjoint.

Corollary 3.6: The map $t \to \exp[J_S t(1 - P_T)XP_N]$ is a strongly continuous one-parameter unitary group of operators on M.

Theorem 3.7: For each one-parameter subgroup of the Poincaré group which, in the representation (3.1), has generator X, we can define a strongly continuous one-parameter unitary group $t \to v_t$ of operators on M, with generator $(1 - P_T)XP_N$ such that

(a) v_t coincides on N/T with the action of $\exp(tJ_S X)$.

(b) v_t acts trivially on T and hence defines an automorphism of the quotient algebra $\Delta_c(N)/1$.

Proof: (a) For $\phi + T \in N/T$ we have

$$(1 - P_T)XP_N\phi + T = (1 - P_T)X\phi + T = X\phi + T,$$

from which (a) now follows.

(b) If $\phi \in T$ then $XP_N\phi = X\phi \in T$. Hence

$$(1 - P_T)XP_N\phi = 0$$

and the assertion follows from previous remarks.

In accord with the notation of Sec. 1, π_0 denotes the Fock representation of $\Delta_c(M)$ on K_0. Let $\mathcal{N}$, $\mathcal{T}$, $\mathcal{S}$ and $\mathcal{S}^\perp$ denote the weak closures of $\pi_0(\Delta_c(N))$, $\pi_0(\Delta_c(T))$, $\pi_0(\Delta_c(S))$ and $\pi_0(\Delta_c(S^\perp))$ respectively (here $S^\perp$ is the orthogonal complement in M of S).

Lemma 3.8: The commutant of $\mathcal{N}$ is $\mathcal{T}$.

Proof: From Manuceau (Ref. 10, 3.4.1) there is a unitary operator U from K_0 to $\mathcal{H} \otimes \mathcal{H}_\perp$ where $\mathcal{H}$ (resp. $\mathcal{H}_\perp$) is the Hilbert space carrying the Fock representation of $\Delta_c(S)$ [resp. $\Delta_c(S^\perp)$], which sets up a spatial isomorphism of $\mathcal{S}$ with $B(\mathcal{H}) \otimes \mathcal{D}_1$ and $\mathcal{S}^\perp$ with $\mathbb{C}_1 \otimes B(\mathcal{H}_\perp)$, respectively, $B(\mathcal{H})$, etc., denote the bounded operators on $\mathcal{H}$, etc. Thus $\mathcal{S}' \triangleq (B(\mathcal{H}) \otimes \mathbb{C}_1)' = \mathbb{C}_1 \otimes B(\mathcal{H}_\perp) \triangleq \mathcal{S}^\perp$. As $\mathcal{S}^\perp \subset \mathcal{S}'$ this means $\mathcal{S}' = \mathcal{S}^\perp$. Thus $\mathcal{N}' = \mathcal{S}' \cap \mathcal{T}' = \mathcal{S}^\perp \cap \mathcal{T}'$. To prove that $\mathcal{N}' = \mathcal{T}$ it remains only to show that $\mathcal{T}$ is maximal Abelian in $\mathcal{S}^\perp$. Clearly, we need only show therefore that the image $\mathcal{T}_0$ of $\mathcal{T}$ in $B(\mathcal{H}_\perp)$ under the above spatial isomorphism is maximal Abelian in $B(\mathcal{H}_\perp)$.

Recall the notation of Lemma 3.5 where we introduced the subspace T' of $S^\perp$ such that $T \oplus T' = S^\perp$. Now as $J_F T' = T$ we have $S^\perp = T' \oplus J_F T'$. Thus T' and T play the role of the real and imaginary subspaces of $S^\perp$ with respect to the Fermi complex structure J_F. It then follows by a fairly well known procedure[12] that the Fock representation of $\Delta_c(S^\perp)$ may be constructed on $L^2(T', d\nu)$, where $d\nu$ is the canonical Gaussian cylinder measure on T' (as a real Hilbert space), by defining:

$$(V(\phi)F)(\psi) = \exp[iB(\phi, \psi)]F(\psi), \quad \phi \in T, \ \psi \in T',$$

$$(U(\phi_1)F(\psi)) = \exp[-B(\phi_1, J_F\psi) - \tfrac{1}{2}B(\phi_1, J_F\phi_1)]F(\psi + \phi_1),$$

$$\phi_1 \in T', \quad \psi \in T',$$

where $F \in L^2(T', d\nu)$. The vacuum, in this realization, is the constant function and is cyclic for $\mathcal{T}_0$ as the linear span of the functions $\psi \to \exp iB(\phi, J\psi)$, $(\psi \in T', \phi \in T)$ is dense in $L^2(T', d\nu)$. Hence $\mathcal{T}_0$ is maximal Abelian as required. (More on this may be found in Ref. 17.)

Corollary 3.9: $\mathcal{T}$ is the center of $\mathcal{N}$.

Theorem 3.10: Let $t \to \beta_t$ and $t \to \alpha_t$ be the one-parameter groups of automorphisms of $\Delta_c(M)$ corresponding to $t \to v_t$ and $t \to \exp(tJ_sX)$, respectively. Then α_t and β_t induce the same automorphism of $\Delta_c(N)/I$. In any cyclic representation of $\Delta_c(M)$ arising from a generating functional on M, $t \to \beta_t$ is implemented by a one-parameter unitary group which in the Fock representation, lies in $\mathcal{N}$.

Proof: It remains only to verify the assertion in the last sentence, the rest following from previous results. Now as v_t is unitary for all $t \in R$, $t \to \beta_t$ is implemented by a one-parameter unitary group in any cyclic representation of $\Delta_c(M)$.[11] Thus in the Fock representation π_0, this group is implemented by $t \to V_t$, say. But v_t acts trivially on T, so β_t acts as the identity on $\Delta_c(T)$ and hence V_t lies in the commutant of $\pi_0(\Delta_c(T))$. That is $V_t \in \mathcal{N}$ by Lemma 3.8.

Remarks 3.11: (a) This theorem shows in what sense $t \to \beta_t$ is an "inner" automorphism. In particular, for the free time evolution $\phi \to \exp(J_s tXP_S)\phi$ (where X is multiplication by k), the corresponding automorphism group $t \to \beta_t^H$ of $\Delta_c(M)$ will be unitarily implemented in any cyclic representation π. Thus we can work with its generator, the physical Hamiltonian, as an operator on the Hilbert space in which $\pi(\Delta_c(M))$ acts. This fact has already been exploited in perturbation theory calculations based on the Fermi method.[18]

(b) The problem discussed above, of determining whether a given automorphism is in some sense inner for the algebra of observables, is of course of more general interest (cf. Kadison[19]). Note that the above analysis shows that in the direct integral decomposition of $\mathcal{N}$ produced by "diagonalizing" its center $\mathcal{T}$ (see Ref. 3), the Poincaré automorphisms are also "diagonalized.' This is important for any discussion of Poincaré transformations of the Fermi physical states.[3]

Proposition 3.12: The automorphism groups $t \to \beta_t$ of Theorem 3.11 are unitarily implemented in the covariant representation π [defined by Eq. (1.9)] by one-parameter unitary groups in $\pi(\Delta_c(N))''$.

Proof: For each representation π_p, $p \in X$, we have (Theorem 3.11) a one-parameter unitary group $t \to U_p(t)$ implementing the automorphism group $t \to \beta_t$. The operator U_t on K given by

$$(U_tF)(p) = U_p(t)F(p)$$

clearly implements $t \to \beta_t$ in the covariant representation. Now the proof of Lemma 3.8 carries over to the representations π_p [as these are Fock representations associated with the complex structures $\Delta(p)J_F\Delta(p)^{-1}$] and so $U_p(t) \in \pi_p(\Delta_c(N))''$ for all $t \in R$ and all $p \in X$. Hence $U_t \in \pi(\Delta_c(N))''$ as required.

Proposition 3.13: The map $X \to Z_X \upharpoonright \mathcal{D}'$ is a Lie algebra homomorphism.

Proof: Clearly $X \to Z_X$ is linear so it remains to prove

$$[Z_X, Z_Y] = Z_{[X,Y]}.$$

To demonstrate this it is sufficient to show that $Z_XZ_Y = Z_{XY}$. Now recall the notation of Lemma 3.5. If $\phi \in \mathcal{D}'$ we have

$$\begin{aligned} Z_XZ_Y\phi &= (1 - P_T)XP_SY\phi_S \\ &= (1 - P_T)XY\phi_S - (1 - P_T)X(1 - P_S)Y\phi_S. \end{aligned}$$

Now $(1 - P_S)Y\phi_S \in T$ and so $X(1 - P_S)Y\phi_S \in T$. Thus

$$Z_XZ_Y\phi = (1 - P_T)XY\phi_S = Z_{XY}\phi,$$

completing the proof.

The above argument actually implies more. In fact we have

$$[(1 - P_T)XP_N]^n = (1 - P_T)X^nP_N$$

so that

$$\exp(tJ_SZ_X) = (1 - P_T)\exp(tJ_SX)P_N + (1 - P_S).$$

Thus we can write

$$\exp(tJ_SZ_X) \upharpoonright S = (1 - P_T)\exp(tJ_SX) \cdot P_N \upharpoonright S. \qquad (3.5)$$

Now for each (a, Λ) in the Poincaré group let $\alpha(a, \Lambda)$ be the operator on S given by

$$\alpha(a, \Lambda)\phi = (1 - P_T)(a, \Lambda)P_N\phi \quad (\phi \in S). \qquad (3.6)$$

Thus for each one parameter subgroup $t \to (a(t), \Lambda(t))$ of the Poincaré group (3.5) shows that $\alpha(a(t), \Lambda(t))$ is a unitary operator on S. Since the Poincaré group is generated by its one parameter subgroups, we have $\alpha(a, \Lambda)$ unitary on S for all elements (a, Λ) of the Poincaré group. This proves

Theorem 3.14: The map $(a, \Lambda) \to \alpha(a, \Lambda)$ defined by (3.6) is a unitary representation of the Poincaré group on S.

This theorem (which is implicit in Ref. 3) has consequences for the radiation gauge quantization of the electromagnetic field. We note that α induces a representation of the Poincaré group by automorphisms of $\Delta_c(S)$.

We conclude this section by noting that many of the above arguments and results carry over to more general situations involving the Weyl algebra formalism. Whenever the physically observable algebra is "embedded" in a larger algebra then this question of relating the automorphisms of the two algebras will arise (for example in a theory with a gauge group). In the case of the electromagnetic field this problem really arises from the gauge freedom implied by the 4-vector potential.

4. GENERAL THEORY OF ZERO MASS FIELDS

The arguments and results of this and our preceding paper[3] are sufficiently general to be applicable to higher spin tensor fields. We leave the generalization of the preceding results [in particular the construction of $\Delta_c(N)$, $\Delta_c(S)$, and $\Delta_c(N)/I \cong \Delta_c(S)$] to the reader. The arguments of this section are intended to point out the applicability of the Weyl algebra formalism in other zero mass theories.

Consider the space $\mathcal{S}_0$ of C^∞ functions on X_0^+ taking values in some complex vector space V which carries an irreducible tensor representation of the Lorentz group (finite dimensional) and satisfying

(a) every $f \in \mathcal{S}_0$ is of fast decrease (together with all its derivatives) as $|k| \to \infty$,

(b) every $f \in \mathcal{S}_0$ vanishes at the origin.

On $\mathcal{S}_0$ we can impose a skew-symmetric form as follows. Every tensor representation of the Lorentz group preserves a matrix η such that $\eta^* = \eta$, $\eta^2 = 1$ (say $\eta = -g \otimes -g \otimes \cdots \otimes -g$). Define

$$(\phi, \eta\phi') = \int_{X_0} \phi(k)^* \eta \phi'(k)\, d\mathbf{k}/k \tag{4.1}$$

for $\phi, \phi' \in \mathcal{S}_0$. Now write

$$B(\phi, \phi') = \frac{1}{2i}[(\phi, \eta\phi') - \overline{(\phi, \eta\phi')}].$$

Then B is a skew-symmetric nongenerate form on $\mathcal{S}_0$. Consequently we can form the algebra $\Delta_c(\mathcal{S}_0)$ for the CCR's defined by B. Clearly $(\phi, \eta\phi')$ cannot be positive definite. A positive definite form analogous to the Fermi inner product may be defined by

$$\langle \phi, \phi' \rangle = B(\phi, J\phi') + iB(\phi, \phi'), \tag{4.2}$$

where $J\phi = -i\eta\phi$ is a non-Lorentz invariant complex structure on $\mathcal{S}_0$. Thus for every tensor field there exists a Fermi method, viz., form the completion M of $\mathcal{S}_0$ in the norm defined by (4.2) and construct $\Delta_c(M)$ and its representations.

In general a zero mass tensor field contains redundant components just as for the electromagnetic potential. These redundant components are eliminated by supplementary equations and gauge conditions, neither of which make sense as operator conditions on the quantized fields. To make sense of them one needs to interpret them as defining subalgebras or quotient algebras of $\Delta_c(M)$.

Now a physical zero mass particle has only two helicity states corresponding to helicities ± s say ($s = 0, \frac{1}{2}, 1, \frac{3}{2}, \cdots$). This follows by a group theoretical argument by choosing an appropriate subspace of M (or quotient space of a subspace) which carries a unitary representation of the Poincaré group equivalent to the direct sum of a zero mass helicity $+s$ and a zero mass helicity $-s$ representation. It is the purpose of wave equations and gauge conditions to choose this "physical space."

There is however a constraint placed on zero mass theories which is a consequence of the fact that, whenever the tensor representation of the Lorentz group in V is irreducible, the corresponding manifestly covariant representation of the Poincaré group on $\mathcal{S}_0$ is indecomposable.[20] This constraint follows from the

Theorem 4.1: Let U be an indecomposable representation of a group G by bounded operators on a Hilbert space $\mathcal{H}$ (real or complex). Let η be a bounded self-adjoint (resp. skew-adjoint) operator such that $\eta^2 = 1$ (resp. $\eta^2 = -1$) and the form

$$f, f' \to (f, \eta f')$$

satisfies

$$(U_g f, \eta U_g f') = (f, \eta f')$$

for all $f, f' \in \mathcal{H}$ and $g \in G$. If U is not irreducible we suppose that $\mathcal{H}'$ is a vector space direct sum (possibly infinite) of closed subspaces of $\mathcal{H}$, each of which carries an irreducible subrepresentation of U. Then

$$(f, \eta f') = 0 \quad \text{for all } f, f' \in \mathcal{H}'.$$

Proof: Note firstly that $f, f' \to (f, \eta f')$ is nondegenerate. Now let $\mathcal{H}_0$ be an irreducible (closed) subspace of $\mathcal{H}$. Let

$$\mathcal{H}_0' = \{f \in \mathcal{H} \mid \langle f, \eta f_0 \rangle = 0 \text{ for all } f_0 \in \mathcal{H}_0\}.$$

Then $\mathcal{H}_0'$ is a closed G-invariant subspace of $\mathcal{H}$ and so the irreducibility of $\mathcal{H}_0$ implies $\mathcal{H}_0 \cap \mathcal{H}_0' = (0)$ or $\mathcal{H}_0$. Now it is not difficult to show (cf. Krein[21]) that the orthogonal complement in $\mathcal{H}$ of the closure of $\mathcal{H}_0 + \mathcal{H}_0'$ is $\eta(\mathcal{H}_0 \cap \mathcal{H}_0')$. Hence if $\mathcal{H}_0 \cap \mathcal{H}_0' = (0)$ then $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_0'$ a Hilbert space direct sum of invariant subspaces of $\mathcal{H}$. Now we have assumed that $\mathcal{H}$ is not decomposable and neither $\mathcal{H}_0$ nor $\mathcal{H}_0'$ is zero, thus we must have $\mathcal{H}_0 \cap \mathcal{H}_0' = \mathcal{H}_0$.

Now let $\mathcal{H}_1$ be a second irreducible (closed) subspace of $\mathcal{H}$. The considerations of the previous paragraph also apply to $\mathcal{H}_1$. Form

$$\mathcal{H}_2 = \mathcal{H}_0 + \mathcal{H}_1$$

(a vector space direct sum) and

$$\mathcal{H}_2' = \{f' \in \mathcal{H} \mid (f, \eta f') = 0 \text{ for all } f \in \mathcal{H}_2\}.$$

By the above argument $\mathcal{H}_2 \cap \mathcal{H}_2' \neq (0)$. So let $f \in \mathcal{H}_2 \cap \mathcal{H}_2'$ and write f uniquely as the sum $f_0 + f_1$, where $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$. As

$$(f, \eta f_0') = 0 \quad \text{for all } f_0' \in \mathcal{H}_0$$

we have

$$(f_1, \eta f_0') = 0 \quad \text{for all } f_0' \in \mathcal{H}_0.$$

The set of all $f_1 \in \mathcal{H}_1$ such that $f_1 \in \mathcal{H}_0'$ is an invariant subspace of $\mathcal{H}_1$. Suppose it is the zero subspace. Then $f = f_0 \in \mathcal{H}_0$ and by the irreducibility of $\mathcal{H}_0$, $\mathcal{H}_2 \cap \mathcal{H}_2' = \mathcal{H}_0$. But if $f_1' \in \mathcal{H}_1$ then $f_1' \in \mathcal{H}_1'$ and further $(f_1', \eta f_0) = 0$ for all $f_0 \in \mathcal{H}_0$ as $\mathcal{H}_0 \subset \mathcal{H}_2'$. This implies that $\mathcal{H}_2 \cap \mathcal{H}_2'$ contains $\mathcal{H}_1$, a contradiction. Thus $\mathcal{H}_2 \cap \mathcal{H}_2'$ contains elements of $\mathcal{H}_1$ and, being G invariant, therefore contains all of $\mathcal{H}_1$. A similar argument to the above gives $\mathcal{H}_0 \subseteq \mathcal{H}_2$, $\mathcal{H}_2'$ implying that $\mathcal{H}_2 \cap \mathcal{H}_2' = \mathcal{H}_2$.

An induction argument gives the result for any finite vector space direct sum of irreducibles. One assumes that it is true for k irreducibles $\mathcal{H}_1, \ldots, \mathcal{H}_k$, and shows that the $(k+1)$th irreducible is in $\mathcal{H}_i'$ for each i by the argument of the previous paragraph. Since $\mathcal{H}_{k+1} \subset \mathcal{H}_{k+1}'$ we have the result.

Infinite direct sums are dealt with as follows. Apply Zorn's lemma to obtain a maximal null direct sum of irreducibles say $\mathcal{H}_m$ (null meaning $\mathcal{H}_m \subset \mathcal{H}_m'$). Suppose $\mathcal{H}_1$ is irreducible and is not contained in $\mathcal{H}_m$. Form $\mathcal{H}_{m+1} = \mathcal{H}_m + \mathcal{H}_1$ (vector space direct sum). $\mathcal{H}_1$ can be shown to be in $\mathcal{H}_0'$ for every irreducible $\mathcal{H}_0 \subseteq \mathcal{H}_m$. Thus $\mathcal{H}_{m+1} \cap \mathcal{H}'_{m+1} = \mathcal{H}_{m+1}$, a contradiction. Thus any maximal direct sum of irreducibles is null, completing the proof.

This result shows that if the "physical fields" of helicity ± s form an invariant subspace of M (and consequently are equivalent to the direct sum of two unitary irreducible representations of the Poincaré group) then there is no Poincaré invariant form on $\mathcal{S}_0$ which is nondegenerate on the physical subspace. Consequently there is no physically sensible Fermi Weyl algebra formalism for this case (the obvious "Weyl algebra" being Abelian on the physical subspace).

These considerations apply to the antisymmetric tensor form of Maxwell's equations,

$$k^\mu F_{\mu\nu} = 0, \quad k^\mu F^{\rho\sigma} + k^\rho F^{\sigma\mu} + k^\sigma F^{\rho\mu} = 0,$$

as the solutions to these equations carry a representation of the Poincaré group which can be shown to be equivalent to a direct sum of irreducibles. The above theorem also applies to any zero mass tensor theory which does not include gauge conditions.

Conversely, if the physical components of a tensor field span a subspace of $\mathcal{S}_0$ on which some Poincaré invariant symplectic form (derivable from an inner product on $\mathcal{S}_0$ by means of a complex structure J) is nondegenerate, these physical components cannot form a Poincaré invariant subspace of $\mathcal{S}_0$. Hence gauge conditions are essential in such a theory.

We see therefore that in a zero mass tensor theory to which a Fermi Weyl algebra formalism is applicable there are two alternative procedures:

(a) a Fermi method in which $\Delta_c(M)$ is constructed and certain subalgebras and quotient algebras specified as physical,

(b) a Segal "radiation gauge" formalism in which the symplectic space M is taken to be a quotient space of a subspace of $\mathcal{S}_0$, specified by supplementary and gauge conditions.

The considerations of Ref. 3 imply that the features of (b) can be extracted from (a).

5. CONCLUSIONS

We have observed that the Fermi method of quantizing the electromagnetic 4-vector potential provides:

(a) an example of spontaneous symmetry breaking in a field theory,

(b) a problem in the analysis of automorphisms of Weyl systems which may be of more general interest,

(c) a unitary representation of the Poincaré group for the radiation gauge formalism,

(d) a prototype for higher spin zero mass quantum field theories.

ACKNOWLEDGMENTS

One of us (A. L. C.) gratefully acknowledges the financial support of a Rothmans Fellowship.

[1] F. Strocchi and A.S. Wightman, J. Math. Phys. 15, 2198 (1974).

[2] I.E. Segal, *Lecture Notes in Mathematics*, No. 140 (Springer-Verlag, Berlin, 1970), p. 30–58.

[3] A.L. Carey, J.M. Gaffney, and C.A. Hurst, University of Adelaide preprint, 1976.

[4] R.F. Streater and A.S. Wightman, *PCT, Spin and Statistics, and All That* (Benjamin, New York, 1964).

[5] L. Garding and A.S. Wightman, Ark. Fys. 28, 129–84 (1964).

[6] V.S. Varadarajan, *Geometry of Quantum Theory* (Van Nostrand, Princeton, N.J., 1970).

[7] U. Niederer and L. O'Raifeartaigh, Fortschr. Phys. 24, 111–30, 131–58 (1974).

[8] C.A. Hurst, Nuovo Cimento 21, 274 (1961).

[9] G. Källen, *Handbuch der Physik* (Springer-Verlag, Berlin, 1958), Vol. 1.

[10] J. Manuceau, Ann. Inst. Henri Poincaré 8, 139–61 (1968).

[11] C.M. Bender, Phys. Rev. 168, 1809 (1968).

[12] I.E. Segal, *Mathematical Problems of Relativistic Physics* (Amer. Math. Soc., Providence, R.I., 1963).

[13] G.G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley, New York, 1972).

[14] D. Shale, Trans. Amer. Math. Soc. 103, 149 (1962).

[15] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II, Fourier Analysis, Self Adjointness* (Academic, New York, 1975).

[16] E. Nelson, Ann. Math. 70, 572 (1959).

[17] J.M. Gaffney, thesis, Adelaide, 1974.

[18] J. Wright, University of Adelaide preprint, 1975.

[19] R.V. Kadison, *Lectures in Modern Analysis and Applications*, Springer Lecture Notes in Mathematics, No. 140 (Springer-Verlag, Berlin, 1970).

[20] A.O. Barut and R. Raczka, Ann. Inst. Herni Poincaré 17, 111 (1972).

[21] M.G. Krein, Amer. Math. Soc. Transl., Ser. 2, 93, 15–92 (1965).

Wavepacket scattering in potential theory[a)]

T. A. Weber and C. L. Hammer

Ames Laboratory-ERDA and Department of Physics, Iowa State University, Ames, Iowa 50011

(Received 13 January 1977)

A contour integration technique is developed which enforces the initial conditions for wavepacket-potential scattering. The expansion coefficients for the exact energy eigenstate expansion are automatically expressed in terms of the plane wave expansion coefficients of the initial wavepacket thereby simplifying what is usually a tedious, mathematical process. The method is applicable regardless of the initial spatial separation of the wavepacket from the scattering center.

INTRODUCTION

In this paper we present a formalism which, although specifically designed to treat the potential scattering of a wavepacket when the packet is initially at a finite distance from the scattering center, is applicable also if the wavepacket is within the potential well. Thus decaying states are included in the analysis.

The basic idea is analogous to the Regge pole analysis in that the discrete sum over bound states is replaced by a contour integration. We show that given an initial wavepacket

$$\psi(x,0)=\int d\mathbf{p}\,a(\mathbf{p})\exp(i\mathbf{p}\cdot\mathbf{x}), \tag{1}$$

a contour $\mathcal{C}$, in the complex $|\mathbf{p}|$ plane, can be found such that

$$\psi(x,t)=\int_C d\mathbf{p}\,a(\mathbf{p})\,\varphi(\mathbf{p},\mathbf{x})\exp(-iEt) \tag{2}$$

is a solution to the time dependent Schrödinger equation if

$$\varphi(\mathbf{p},\mathbf{x})=[1+R(\mathbf{x},\mathbf{p})]\exp(i\mathbf{p}\cdot\mathbf{x}) \tag{3}$$

is a solution to the stationary state Schrödinger equation

$$H\varphi=E\varphi. \tag{4}$$

The wavefunction $\psi(\mathbf{x},t)$ is made the unique solution of the time dependent Schrödinger equation by choosing the contour so that the singularities of $R(\mathbf{x},\mathbf{p})$ are excluded, giving

$$\int_C d\mathbf{p}\,a(\mathbf{p})\,R(\mathbf{x},\mathbf{p})\exp(i\mathbf{p}\cdot\mathbf{x})=0, \tag{5}$$

thereby satisfying the initial condition, Eq. (1).

The analysis is particularly applicable for initial wavepackets that are localized to a finite region of space outside of a finite range potential V. For this case a contour $\mathcal{C}$ which satisfies the condition Eq. (5) can always be found. In the case where the wavepacket is inside the range of the potential, which includes the infinite range potential, a contour can still be found but in addition the amplitude $a(\mathbf{p})$ must be modifed from that of the plane wave coefficients.

The connection to other formalisms, in particular the Lippmann–Schwinger formalism, is made by noting that $R(\mathbf{x},\mathbf{p})$ may be written as

$$R(x,p)=(E-H+i\eta)^{-1}V. \tag{6}$$

However, in their formalism the initial wavepacket must be considered at an infinite distance from the scattering center for the solution to be exact. In this limit $\mathcal{C}$ becomes the real $|\mathbf{p}|$ axis since there are no bound state contributions to the initial wavepacket.

The usefulness of the analysis presented here lies in the fact that the initial wavepacket is expanded in terms of *plane wave coefficients* rather than the expansion coefficients for the *exact eigenfunctions* of the Hamiltonian—a procedure which is usually long and tedious. As will be shown, the connection between the plane wave coefficients and the exact expansion coefficients is made simply by "straightening" out the contour.

The formalism is first developed for the one-space-diemensional case and then generalized to three space dimensions.

ONE SPACE DIMENSION

A. Wavepacket properties

To facilitate the discussion of the scattering of wavepackets from one-dimensional potentials, it is necessary to know the analytic properties and the asymptotic behavior of the plane wave coefficients, $a(p)$. The type of wavepacket that is to be considered can be represented by the Fourier integral

$$\chi(x,0)=(2\pi)^{-1/2}\int_{-\infty}^{\infty}dp\,a(p)\exp(ipx), \tag{7}$$

where

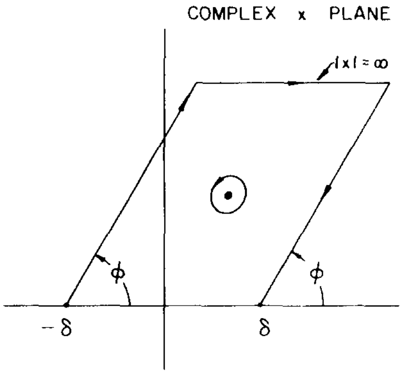


FIG. 1. Contour integration in the complex x plane.

[a)] Work performed for the U.S. Energy Research and Development Administration under Contract No. W-7405-eng-82.

$$\chi(x, 0) = 0, \quad |x| > \delta, \tag{8}$$

and $a(p)$ is presumed centered about $p = 0$. In general, the integration in Eq. (7) can be replaced by a closed contour integration in the complex p plane along the real p axis returning along an infinite arc either in the upper or lower p plane depending on the range of x. The behavior of $a(p)$ for large p, which can be determined from the reciprocal Fourier transform

$$a(p) = (2\pi)^{-1/2} \int_{-\delta}^{\delta} dx\, \chi(x, 0) \exp(-ipx), \tag{9}$$

must be known to determine this range. The integral of Eq. (9) can be replaced by the contour integration shown in Fig. 1, where the path of integration which damps most rapidly is along the lines of constant phase

$$\phi = -\operatorname{phase}(ip). \tag{10}$$

The leading terms in the asymptotic expansion of $a(p)$ then become the Laplace transforms,

$$a(p) \approx (2\pi)^{-1/2} \exp(ip\delta) \int_0^{\infty \exp(i\phi)} dz\, \chi(z - \delta, 0) \exp(-ipz)$$
$$+ (2\pi)^{-1/2} \exp(-ip\delta) \int_{\infty \exp(i\phi)}^{0} dz\, \chi(z + \delta, 0) \exp(-ipz). \tag{11}$$

Contributions from singularities such as the pole shown in Fig. 1 damp exponentially to zero relative to the leading terms for large p. The first term of the asymptotic expansion is

$$\underset{|p| \to \infty}{a(p)} \approx (2\pi)^{-1/2}(ip)^{-1}[\exp(i\delta p)\chi(-\delta, 0)$$
$$- \exp(-i\delta p)\chi(\delta, 0)], \tag{12}$$

as obtained by expanding χ in the integrands of Eq. (11) about $z = 0$. If this expression is used in the integrand of Eq. (7) for large p, it is clear that for $x > \delta$ the contour can be closed in the upper half p plane whereas for $x < \delta$, the contour can be closed in the lower half p plane. In either case, the result of the integral is zero because of the condition imposed by Eq. (8). Thus $a(p)$ has no singularities in the entire finite p plane. This conclusion can also be seen directly from Eq. (9) since the analytic properties of $a(p)$ are the same as those of $\exp(ipx)$ as long as the integral exists.

A wavepacket that is centered about the point $x = x_0$ rather than at the point $x = 0$, and whose momentum is centered about p_0 rather than zero is

$$\psi(x, 0) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} dp\, a(p - p_0) \exp[ip(x - x_0)], \tag{13}$$

as obtained from Eq. (7) by displacing both x and p.

B. The step potential

As a first example to illustrate the analysis in a simple way, consider a wavepacket, initially represented by Eq. (13) with $x_0 < -\delta$, incident upon a step potential

$$V(x) = 0, \quad x < 0,$$
$$= -V_0, \quad x > 0. \tag{14}$$

The Schrödinger equation is

$$\left(-\frac{1}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right) \varphi_a(p, x) = E\varphi_a(p, x) \tag{15}$$

with solutions

$$\varphi_a(p, x) = a(p - p_0) \exp(-ipx_0) \exp(ipx) + b(p) \exp(-ipx),$$
$$x \leq 0,$$

and

$$\varphi_a(p, x) = c(p) \exp(ip'x), \quad x \geq 0, \tag{16}$$

where

$$p' = (p^2 + 2mV_0)^{1/2}, \quad E = (p^2/2m). \tag{17}$$

Since the wavefunction and its derivative are continuous across the step,

$$c(p) = 2p(p + p')^{-1} a(p - p_0) \exp(-ipx_0),$$
$$b(p) = (p - p')(p + p')^{-1} a(p - p_0) \exp(-ipx_0), \tag{18}$$

for all p. Even for complex p, Eqs. (16) and (18) represent a solution to the Schrödinger equation which is continuous across the origin.

Then, according to Eq. (2), the wavefunction at any time t is

$$\psi(x, t) = (2\pi)^{-1/2} \int_C dp\, a(p - p_0) \exp(-ipx_0)$$
$$\times\{1 + (p - p')(p + p')^{-1} \exp(-i2px)\}$$
$$\times \exp[i(px - Et)], \quad x \leq 0$$

and

$$\psi(x, t) = (2\pi)^{-1/2} \int_C dp\, 2p(p + p')^{-1} a(p - p_0)$$
$$\times \exp(-ipx_0) \exp[i(p'x - Et)], \quad x \geq 0, \tag{19}$$

where C must be chosen so that the initial condition expressed in Eq. (13) is satisfied, that is,

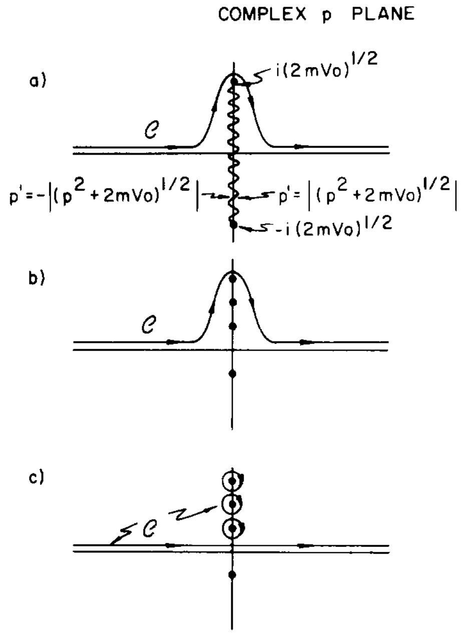


FIG. 2. Contour integration in the complex p plane.

$$\int_C dp\, a(p-p_0)(p-p')(p+p')^{-1}\exp[-ip(x+x_0)]=0, \quad x\leqslant 0,$$

$$\int_C dp\, a(p-p_0)2p(p+p')^{-1}\exp[i(p'x-px_0)]=0, \quad x\geqslant 0. \tag{20}$$

The only singularities in $c(p)$ and $b(p)$ different from those in $a(p-p_0)$ are the branch points due to p' located at

$$p=\pm i(2mV_0)^{1/2} \tag{21}$$

as shown in Fig. 2. The phases shown in the figure are chosen to make p' real and positive when p is real and positive. With this choice, Eq. (18) corresponds to the usual plane wave results for real, positive p. The initial condition is automatically satisfied if C is taken along the real p axis and over the cut as shown in Fig. 2(a) as is readily verified from Eqs. (12) and (20) by closing the contour in the upper half-plane. Thus, Eq. (19) represents a unique solution to the time dependent Schrödinger equation

$$\left(-\frac{1}{2m}\frac{\partial^2}{\partial x^2}+V(x)\right)\psi(x,t)=i\left(\frac{\partial\psi(x,t)}{\partial t}\right) \tag{22}$$

with Eq. (13) as the initial condition.

It should be noted that by choosing C to satisfy the initial conditions, $\psi(x,t)$ has automatically been expanded in terms of a complete set of bounded, stationary state solutions. For real $p>0$, the integration along C corresponds to a superposition of the usual plane wave results. Along the cut, the integration represents a superposition of damped exponential solutions for $x<0$ and a superposition of standing waves for $x>0$ corresponding to the "bound" states of a semi-infinite square well. This follows from Eq. (19) since along the cut

$$\psi(x,t)=-2i(2\pi)^{-1/2}(mV_0)^{-1}\int_0^{(2mV_0)^{1/2}}(d|p|)|pp'|$$
$$\times a(i|p|-p_0)\exp[|p|(x+x_0)]\exp[i(|p|^2/2m)t],$$
$$x\leqslant 0$$

and

$$\psi(x,t)=-2i(mV_0)^{-1}(2\pi)^{-1/2}\int_0^{(2mV_0)^{1/2}}(d|p|)|p|a(i|p|-p_0)$$
$$\times\exp(|p|x_0)\exp[i|p|^2/2m)t][|p'|\cos|p'|x$$
$$+|p|\sin|p'|x], \quad x\geqslant 0, \tag{23}$$

where $x_0\leqslant 0$ and $|p'|=(2mV_0-|p|^2)^{1/2}$.

For real $p<0$, the transmission and reflection coefficients given in Eq. (18) remain unchanged because p' also changes sign along the negative p axis. This part of the solution represents two plane waves incident upon the step potential one from the left and the other from the right, giving rise to that part of the initial wavepacket which corresponds to wavelets traveling to the left.

C. The general finite range potential

Let the finite range potential be represented by

$$V=0, \quad x\geqslant a, \quad x\leqslant 0,$$
$$V=V(x), \quad 0\leqslant x\leqslant a, \tag{24}$$

and let the wavepacket be incident from the right so that

$$\psi(x,0)=(2\pi)^{-1/2}\int_{-\infty}^{\infty}dp\, a(p-p_0)\exp[-ip(x-x_0)],$$
$$x\geqslant a+\delta. \tag{25}$$

The most convenient form for the solutions of Eq. (15) is

$$\varphi_a=a(p-p_0)\exp[-ip(x-x_0)]+b(p)\exp(ipx), \quad x\geqslant a,$$
$$\varphi_a=a(p-p_0)f_-(p,x)\exp(ipx_0)+b(p)f_+(p,x), \quad 0\leqslant x\leqslant a,$$
$$\varphi_a=c(p)\exp(-ipx), \quad x\leqslant 0, \tag{26}$$

where

$$f_\pm(p,x)=f(\pm p,x). \tag{27}$$

The usual boundary conditions at $x=a$ are satisfied if

$$f_\pm(p,a)=\exp(\pm ipa),$$
$$(df_\pm(p,a)/da)=\pm ip\exp(\pm ipa), \tag{28}$$

so that in terms of two linearly independent solutions y_1 and y_2,

$$f(p,x)=[\exp(ipa)/w(p)]\{[y_1'(p,a)-ipy_1(p,a)]y_2(p,x)$$
$$-[y_2'(p,a)-ipy_2(p,a)]y_1(p,x)\}, \tag{29}$$

where

$$w(p)=y_1'y_2-y_1y_2', \quad y'=\frac{dy}{dx}.$$

The boundary conditions at $x=0$ are satisfied for all p if in Eq. (26)

$$b(p)=-a(p-p_0)\exp(ipx_0)(ip\mathcal{L}_-+\mathcal{L}_-')(ip\mathcal{L}_++\mathcal{L}_+')^{-1},$$
$$c(p)=2ipa(p-p_0)\exp(ipx_0)(ip\mathcal{L}_++\mathcal{L}_+')^{-1}, \tag{30}$$

where

$$\mathcal{L}_\pm(p)=\lim_{x\to 0}f_\pm(p,x), \quad \mathcal{L}_\pm'(p)=\lim_{x\to 0}f_\pm'(p,x). \tag{31}$$

The function $\mathcal{L}_\pm(p)$ is analogous to the Jost function for the case of s wave scattering.[1] The solution to Eq. (22) with Eq. (25) as the initial condition is

$$\psi(x,t)=(2\pi)^{-1/2}\int_C dp\, a(p-p_0)\exp i(px_0-Et)[\exp(-ipx)$$
$$-(ip\mathcal{L}_-+\mathcal{L}_-')(ip\mathcal{L}_++\mathcal{L}_+')^{-1}\exp(ipx)], \quad x\geqslant a,$$

$$\psi(x,t)=(2\pi)^{-1/2}\int_C dp\, a(p-p_0)2ip(ip\mathcal{L}_++\mathcal{L}_+')^{-1}$$
$$\times\exp\{-i[p(x-x_0)+Et]\}, \quad x\leqslant 0,$$

and

$$\psi(x,t)=(2\pi)^{-1/2}\int_C dp\, a(p-p_0)\exp i(px_0-Et)$$
$$\times\varphi(p,x), \quad 0\leqslant x\leqslant a, \tag{32}$$

where

$$\varphi(p,x)=f_-(p,x)-(ip\mathcal{L}_-+\mathcal{L}_-')(ip\mathcal{L}_++\mathcal{L}_+')^{-1}f_+(p,x). \tag{33}$$

The contour C must be taken over the poles, which occur because of the zeros of $(ip\mathcal{L}_++\mathcal{L}_+)$, as shown in Fig. 2b) or equivalently as shown in Fig. 2c).

To see that this contour actually satisfies the initial conditions, the asymptotic properties of the integrands of Eq. (32) must be examined for large $\mathrm{Im}p > 0$. This is most easily accomplished from the integral equation

$$\varphi(p, x) = 2ip(\mathcal{L}'_+ + ip\mathcal{L}_+)^{-1} \exp(-ipx) + 2mp^{-1} \int_0^x dx' \sin[p(x-x')]V(x')\varphi(p, x'). \tag{34}$$

Following the iteration procedure used by Newton,[2] it is easy to show for $a \geqslant x \geqslant 0$, and $\nu = \mathrm{Im}p > 0$, that

$$\begin{aligned} &\varphi(p, x) \leqslant C\, 2ip(\mathcal{L}'_+ + ip\mathcal{L}_+)^{-1} \exp[\nu x + C|p|^{-1} \int_0^x |V(x')|\, dx], \\ &|\sin p(x-x')| \leqslant C \exp[\nu(x-x')], \quad x > x', \end{aligned} \tag{35}$$

where C is a positive number independent of x or p. Consequently as $|p| \to \infty$, $\mathrm{Im}p > 0$, it follows from substitution into Eq. (34) for the quantities in Eq. (35), that $\varphi(p, x)$ approaches its unperturbed value

$$\lim_{|p|\to\infty} \varphi(p, x) = 2ip(\mathcal{L}'_+ + ip\mathcal{L}_+)^{-1} \exp(-ipx). \tag{36}$$

Because of Eqs. (28) and (33), the Wronskian of $\varphi(p, x)$, as defined in Eq. (34), with f_+ and f_-, evaluated at $x = a$, gives

$$(\mathcal{L}'_- + ip\mathcal{L}_-)(\mathcal{L}'_+ + ip\mathcal{L}_+)^{-1} = -m(ip)^{-1} \int_0^a dx'\, V(x')\varphi(p, x') \exp(-ipx'), \tag{37}$$

and

$$2ip(\mathcal{L}'_+ + ip\mathcal{L}_+)^{-1} = 1 + m(ip)^{-1} \int_0^a dx'\, V(x')\varphi(p, x') \exp(ipx'). \tag{38}$$

Substitution for the asymptotic limit of φ from Eq. (36) into Eqs. (37) and (38) gives

$$\begin{aligned} &\lim_{|p|\to\infty} (\mathcal{L}'_+ + ip\mathcal{L}_+) = 2ip, \quad \mathrm{Im}p > 0, \\ &\lim_{|p|\to\infty} (\mathcal{L}'_- + ip\mathcal{L}_-)(\mathcal{L}'_+ + ip\mathcal{L}_+)^{-1} \sim \exp(-2ipa), \quad \mathrm{Im}p > 0. \end{aligned} \tag{39}$$

As a result of these asymptotic conditions, Eqs. (35) and (39), at $t = 0$ for $x_0 \geqslant a + \delta$, the contours of Eq. (32) shown in Fig. 2b) can be closed in the upper complex p plane so that the only contribution arises from the initial wavepacket as required.

As in the previous section, the choice of the contour results in the expansion of $\psi(x, t)$ in terms of a complete set of stationary state solutions. For real $p > 0$, the integration along C corresponds to a superposition of the usual plane wave results. The sum of the residues from the bound state poles is a superposition of bound state eigenfunctions, exponentially damped for $x > a$ and $x < 0$. For real $p < 0$ the interpretation is similar to that given for the step potential.

II. THREE-SPACE DIMENSIONS

A. Spherically symmetric potentials of finite range

The discussion of the previous section is readily extended to the consideration of spherically symmetric potentials

$$\begin{aligned} &V = V(r), \quad r \leqslant a, \\ &V = 0, \quad r \geqslant a. \end{aligned} \tag{40}$$

The radial part of the solution then is

$$\begin{aligned} \varphi_l(p, r) = {}& (2pr)^{-1} \exp[i(\pi/2)(l+1)] \\ & \times [f_{l_-}(p, r) - \exp(-i\pi l)S_l f_{l_+}(p, r)], \\ & r \leqslant a \end{aligned} \tag{41}$$

where

$$r^{-2} \frac{\partial}{\partial r} \frac{r^2 \partial\varphi_l}{\partial r} - l(l+1)r^{-2}\varphi_l + 2m[E - V(r)]\varphi_l = 0,$$

and S_l is the S matrix defined by the ratio of the Jost functions,[3]

$$S_l = [\mathcal{J}_{l_-}(p)/\mathcal{J}_{l_+}(p)] \exp(i\pi l), \tag{42}$$

$$\mathcal{J}_{l_\pm}(p) = (2l+1) \lim_{r\to 0} r^l f_l(\pm p, r), \tag{43}$$

and

$$\begin{aligned} \varphi_l(p, r) &= \tfrac{1}{2}[h_l^{(2)}(pr) + S_l h_l^{(1)}(pr)] \\ &= j_l(pr) + \tfrac{1}{2}(S_l - 1)h_l^{(1)}(pr), \quad r \geqslant a, \end{aligned} \tag{44}$$

where $j_l(pr)$ is the spherical Bessel function and $h_l^{(1)}$ and $h_l^{(2)}$ spherical Hankel functions.[4]

The boundary conditions at $r = a$ are satisfied if

$$\begin{aligned} &(pa)^{-1} f_{l_+}(p, a) = h_l^{(1)}(pa) \exp[i(\pi/2)(l+1)], \\ &(pa)^{-1} f_{l_-}(p, a) = h_l^{(2)}(pa) \exp[-i(\pi/2)(l+1)]. \end{aligned} \tag{45}$$

Therefore,

$$\lim_{a\to\infty} f_{l_\pm}(p, a) = \exp(\pm ipa) \tag{46}$$

in parallel to the usual asymptotic boundary conditions for potentials of infinite range.

As shown in the next section in Eqs. (67) and (68) the incident wavepacket

$$\psi(\mathbf{x}, 0) = \int d\mathbf{p}\, a(\mathbf{p} - \mathbf{p}_0) \exp[i\mathbf{p} \cdot (\mathbf{x} - \mathbf{x}_0)] \tag{47}$$

can be expanded in the form

$$\psi(\mathbf{x}, 0) = \sum_{l,m} \int_{-\infty}^{\infty} p^2\, dp\, j_l(pr) a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0) Y_{lm}(\hat{x}), \tag{48}$$

$$a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0) = (p\pi)^{-1} \int_0^{\infty} r^2\, dr\, h_l^{(1)}(pr) \times \int d\Omega\, Y^*_{lm}(\hat{x}) \psi(\mathbf{x}, 0), \tag{49}$$

where $r = |\mathbf{x}|$. For a finite wavepacket $\psi(\mathbf{x}, 0)$ there is a δ such that

$$\psi(\mathbf{x}, 0) = 0, \quad |\mathbf{x} - \mathbf{x}_0| \geqslant \delta. \tag{50}$$

Therefore, the asymptotic properties of $a_{l,m}$ can be obtained from

$$a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0) \rightarrow (p\pi)^{-1} \int_{|x_0|-\delta}^{|x_0|+\delta} r^2\, dr\, h_l^{(1)}(pr) \int d\Omega\, Y^*_{lm}(\hat{x}) \psi(\mathbf{x}, 0), \tag{51}$$

in parallel to the arguments leading to Eq. (12). The exponential behavior for large p then is

$$\lim_{p\to\infty} a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0) \to p^{-3} \exp ip(|x_0| \pm \delta). \tag{52}$$

The wavefunction for all time t is therefore

$$\psi(x, t) = (2r)^{-1} \sum_{l,m} \exp[i(\pi/2)(l+1)] Y_{l,m}(\hat{x}) \times \int_C p\, dp\, a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0)[f_{l-}(pr) - \exp(-i\pi l) S_l f_{l+}(pr)] \exp(-iEt), \quad r \leqslant a \tag{53}$$

and

$$\psi(x, t) = \sum_{l,m} Y_{lm}(\hat{x}) [\int_{-\infty}^{\infty} p^2\, dp\, j_l(pr) \times a_{l,m}(p, \mathbf{p}_0, \mathbf{x}_0) \exp(-iEt) + \tfrac{1}{2} \int_C p^2\, dp\, a_{l,m}(p, \mathbf{p}_0, \mathbf{x}_0)(S_l - 1) \times h_l^{(1)}(pr) \exp(-iEt), \quad r \geqslant a, \tag{54}$$

where the contour C is the same as that shown in Figs. 2b) and 2c).

The asymptotic behavior of S_l and φ_l have been investigated extensively by Newton[5] and found to be essentially the same as the corresponding quantities given in the previous section for the one-dimensional case. Consequently at $t = 0$, since $|\mathbf{x}_0| \gg a + \delta$, the contour in Eqs. (53) and (54) can be closed in the upper half p plane so that the only contribution arises from the initial wavepacket.

Again the choice of the contour results in the expansion $\psi(x, t)$ in terms of a complete set of stationary state solutions, the bound state contributions arising from the poles due to the zero's of the Jost functions $\mathcal{J}_{l+}(p)$.

B. The general finite range potential

This result is readily generalized to the general three-dimensional potential of finite range if it is assumed that the corresponding S matrix has the appropriate asymptotic properties for large $\mathrm{Im}p > 0$. The general scattering solution outside the range of the potential can then be taken to be

$$\psi(\mathbf{x}, t) = \int_C d\mathbf{p}\, a(\mathbf{p} - \mathbf{p}_0) \exp[-i(\mathbf{p}\cdot\mathbf{x}_0 + Et)] \times \{1 + [E - H]^{-1} V\} \exp(i\mathbf{p}\cdot\mathbf{x}). \tag{55}$$

If $\exp(i\mathbf{p}\cdot\mathbf{x})$ is expanded in terms of spherical Bessel functions and Legendre polynomials in the usual fashion

$$\psi(\mathbf{x}, t) = \sum_{l,m} Y_{lm}(\hat{x}) \{\int_{-\infty}^{\infty} p^2\, dp\, j_l(pr) a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0) \times \exp(-iEt) + \int_C p^2\, dp\, [i\eta + E - H]^{-1} V j_l(pr) \times a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0) \exp(-iEt)\}, \tag{56}$$

where C again excludes the singularities in the upper p plane and where the sign of η is chosen so that the contour can be closed in the upper half of the p plane thereby enabling the initial conditions to be satisfied.

C. Potentials with infinite range and completeness relationships

The result expressed in Eq. (56) could just as well apply for potentials of infinite range since the term in brackets is a general stationary state solution. In particular, if the potential goes to zero strongly enough for large r, such as a superposition of Yakawa potentials, the S matrix goes to unity along the infinite arc in the upper half of the p plane,[6]

$$\lim_{|p|\to\infty} S_l(p) = 1. \tag{57}$$

It then follows from the symptotic properties of $a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0)$ and Green's function that the contour can be completed in the upper half p plane thereby eliminating the second term of Eq. (56) as required to satisfy the initial conditions.

For potentials that do not go to zero sufficiently rapidly for large r, the S matrix is not well defined off the real p axis or it may take on values such that the contour cannot be closed in the upper half of the p plane for all values of r. In this event, or, if the wavepacket is inside the potential well, it is necessary to expand the initial wavepacket directly in terms of a complete set of states.

However, the techniques described earlier can still be used to obtain this expansion. For example, the completeness relationship can also be written as a contour integral.

$$\int_C p^2\, dp\, \varphi_l(p, r) \varphi_l(-p, r') = \pi(rr')^{-1}[\delta(r - r') - (-1)^l \delta(r + r')], \tag{58}$$

where C is a particular choice of contour which is described below. The function $\varphi_l(p, r)$, defined in Eq. (41), is now assumed to apply for all r, and

$$\varphi_l(-p, r) = (2pr)^{-1} \exp[-i(\pi/2)(l+1)] \times [f_{l+}(p, r) - (\mathcal{J}_{l+}/\mathcal{J}_{l-}) f_{l-}(pr)]. \tag{59}$$

The proof of Eq. (58) can be seen more directly after substitution for $\varphi_l(\pm p, r)$,

$$\int_C p^2\, dp\, \varphi_l(p, r) \varphi_l(-p, r') = (4rr')^{-1} \int_C dp [f_{l-}(p, r) - (\mathcal{J}_{l-}/\mathcal{J}_{l+}) f_{l+}(p, r)] \times [f_{l+}(p, r') - (\mathcal{J}_{l+}/\mathcal{J}_{l-}) f_{l-}(p, r')] = (4rr')^{-1} \{\int_{C_1} dp\, f_{l+}(p, r')[f_{l-}(p, r) - (\mathcal{J}_{l-}/\mathcal{J}_{l+}) f_{l+}(p, r)] + \int_{C_2} dp\, f_{l-}(p, r')[f_{l+}(p, r) - (\mathcal{J}_{l+}/\mathcal{J}_{l-}) f_{l-}(p, r)]\}. \tag{60}$$

It is well known that $f_{l\pm}(p, r)$ is analytic everywhere in the upper (lower) half p plane[7] and that $\varphi_l(\pm p, r)$ is analytic everywhere[8] in the upper (lower) half p plane except for poles of branch points on the positve (negative) imaginary p axis. Consequently, C_1 is chosen as shown in Fig. 2b) and C_2 is chosen so that it goes under the singularities on the negative imaginary p axis. Thus by changing p to $-p$ in the second term of Eq. (60),

$$\int_C p^2\, dp\, \varphi_l(p, r)\varphi_l(-p, r')$$
$$= (2rr')^{-1} \int_{C_1} dp\, f_{l_+}(p, r')[f_{l_-}(p, r) - (\mathcal{F}_{l_-}/\mathcal{F}_{l_+}) f_{l_+}(p, r)]. \tag{61}$$

For large $|p|$, $\text{Im}p > 0$,[7]

$$\lim_{|p|\to\infty} [f_{l_+}(p, r)/\mathcal{F}_{l_+}(p)]$$
$$= [(2l+1)!!]^{-1} p^l \exp i[pr - (\pi/2)l],$$

$$\lim_{|p|\to\infty} [\mathcal{F}_{l_+}(p) f_{l_-}(p, r) - \mathcal{F}_{l_-}(p) f_{l_+}(p, r)]$$
$$= -2ip^{-l}(2l+1)!!\sin[pr - (\pi/2)l]. \tag{62}$$

If $r' > r$, these asymptotic properties allow C_1 to be closed in the upper half of the p plane, excluding all singularities thereby causing the right-hand side of Eq. (61) to vanish. The same result is obtained if $r > r'$, since the original integral is symmetric to an interchange of r' and r as can be seen from the first equality in Eq. (60). That the integral is the delta function given in Eq. (58) then follows directly from the asymptotic properties given by Eq. (62). Furthermore as a result of Eq. (61),

$$\int_{C_1} dp\, f_{l_+}(p, r')[f_{l_-}(p, r) - (\mathcal{F}_{l_-}/\mathcal{F}_{l_+}) f_{l_+}(p, r)]$$
$$= 2\pi[\delta(r - r') - (-1)^l \delta(r + r')]. \tag{63}$$

This expression can now be used to expand an arbitrary wavepacket in terms of the stationary state solution $\varphi_l(p, r)$. Thus from Eqs. (41), (48), and (63),

$$\begin{aligned}\psi_{l,m}(r, 0) &\equiv \int_{-\infty}^{\infty} p^2\, dp\, j_l(pr)\, a_{lm}(p, \mathbf{p}_0, \mathbf{x}_0)\\ &= \int_{-\infty}^{\infty} p^2\, dp\, a_{l,m}(p, \mathbf{p}_0, \mathbf{x}_0)\\ &\quad\times \int_0^{\infty} (r'\, dr'/r)\delta(r - r') j_l(pr')\\ &= \int_{C_1} k^2\, dk\, A_{l,m}(k)\varphi_l(k, r)\end{aligned} \tag{64}$$

with

$$\begin{aligned}A_{l,m} &= (\pi k)^{-1} \exp[-i(\pi/2)(l+1)] \int_{-\infty}^{\infty} p^2\, dp\\ &\quad a_{l,m}(p, \mathbf{p}_0, \mathbf{x}_0) \int_0^{\infty} r'\, dr'\, j_l(pr') f_{l_+}(kr')\\ &= (\pi k)^{-1} \exp[-i(\pi/2)(l+1)]\\ &\quad\times \int_0^{\infty} r'\, dr'\, \psi_{l,m}(r', 0) f_{l_+}(k, r'),\end{aligned} \tag{65}$$

where it is assumed that $\psi(\mathbf{x}, 0)$ is zero for $\mathbf{x} = 0$ so that the term containing $\delta(r + r')$ in Eq. (63) can be ignored for all values of r'. Since $\varphi_l(k, r)$ is the stationary state solution, regular at the origin, it follows that the solution as a function of time is

$$\begin{aligned}\psi(\mathbf{x}, t) &= \sum_{l,m} Y_{l,m}(\hat{x}) \int_{C_1} k^2\, dk\, A_{l,m}(k)\\ &\quad\times \varphi_l(k, r)\exp(-iEt).\end{aligned} \tag{66}$$

This expansion theorem has a simple form for the free particle case. The function $\varphi_l(kr)$ then reduces to $j_l(kr)$ and $f_{l_+}(k, r)$ is proportional to $h^{(1)}(kr)$ as defined in Eq. (45). Thus the initial wavepacket $\psi(\mathbf{x}, 0)$ can be represented by

$$\psi(\mathbf{x}, 0) = \sum_{l,m} Y_{lm}(\hat{x}) \int_{-\infty}^{\infty} k^2\, dk\, a_{lm}(k) j_l(kr) \tag{67}$$

with

$$\begin{aligned}a_{lm}(k) &= (k\pi)^{-1} \int_0^{\infty} r^2\, dr\, h_l^{(1)}(kr)\\ &\quad\times \int d\Omega\, Y_{lm}^*(\hat{x})\psi(\mathbf{x}, 0).\end{aligned} \tag{68}$$

CONCLUSIONS

The formalism presented here removes the necessity for considering the initial wavepacket at an infinite distance from the scattering center at an initial time $t = -\infty$. As a result, final states can be obtained for times large compared to the interaction time but not necessarily infinite. Therefore, decaying states can be readily examined using this formalism rather than the time dependent formalism previously developed[9] which depended explicitly upon an initial interaction time t_0. There seems to be no conceptual difficulty in extending the concepts developed here to relativistic theories and to quantized field theories since the restrictions on the S matrix which are normally assumed are sufficient to ensure that the contour can be closed in the upper half of the p plane.

[1] Roger G. Newton, *Scattering Theory of Waves and Particles* (McGraw-Hill, New York, 1966), Chap. 12, Sec. 12.1.2, p. 340.

[2] See Ref. 1, p. 331.

[3] See Ref. 1, p. 373.

[4] Leonard I. Schiff, *Quantum Mechanics* (McGraw-Hill, New York, 1955), 2nd ed., Chap. 4, pp. 77–9.

[5] See Ref. 1, Chap. 12.

[6] See Ref. 1, p. 347.

[7] See Ref. 1, p. 373.

[8] Marvin L. Goldberger and Kenneth M. Watson, *Collision Theory* (Wiley, New York, 1964), Chap. 8.

[9] C. L. Hammer and T. A. Weber, J. Math. Phys. 8, 494 (1966).

Causally symmetric spacetimes

Frank J. Tipler

Mathematics Department, University of California at Berkeley, Berkeley, California 94720

(Received 11 November 1976)

Causally symmetric spacetimes are spacetimes with $J^+(S)$ isometric to $J^-(S)$ for some set S. We discuss certain properties of these spacetimes, showing for example that if S is a maximal Cauchy surface with matter everywhere on S, then the spacetime has singularities in both $J^+(S)$ and $J^-(S)$. We also consider totally vicious spacetimes, a class of causally symmetric spacetimes for which $I^+(p)=I^-(p)=M$ for any point p in M. Two different notions of stability in general relativity are discussed, using various types of causally symmetric spacetimes as starting points for perturbations.

1. INTRODUCTION

The concept of symmetry is basic to physics. Symmetry in general relativity is usually based on a local one-parameter group of isometries generated by a vector field: a Killing vector field. In this paper I shall develop a notion of symmetry which is based on the global causal structure of spacetime. The reasons for analyzing spacetimes with causal symmetries are in part the same as the reasons for considering spacetimes with Killing symmetries: First, the spacetimes having such symmetries mimic important known features of the actual universe while simplifying the problem of solving the field equations; and second, the mathematical simplicity of such spacetimes allows them to be used as easily understood examples of exotic spacetime structures—structures which may form a part of the actual universe.[1]

I shall discuss both applications of symmetries in this paper. After setting down the definitions and basic relations between the various types of causal symmetries in Sec. 2, I shall show in Sec. 3 that all time symmetric universes which contain matter everywhere are singularity symmetric. That is, these universes have singularities both to the past and to the future of the spacelike hypersurface about which the universe is time symmetric. Many writers believe[2-4] that the actual universe is closed, and the evidence suggests[5] that it is isotropic and homogeneous. This implies the existence of a surface of time symmetry. Thus if we assume that this surface of time symmetry is a Cauchy surface, then it follows that the universe can exist for only a finite time. In general, Sec. 3 will be devoted to a discussion of the conditions which must be imposed on a spacetime in order to make it singularity symmetric.

In Sec. 4 I shall briefly discuss some of the properties of totally vicious spacetimes, the class of causally symmetric spacetimes for which $I^+(p)=I^-(p)=M$, the entire spacetime, for any point p in M. These spacetimes provide a counterexample to a theorem by Hawking and Sachs: A causally simple spacetime is stably causal.

Causally symmetric spacetimes have one advantage over Killing symmetric spacetimes. Causal symmetries are quite amendable to analysis by the global techniques developed by Hawking and Penrose, and using these methods it is easy to prove that many of the properties of these spacetimes are "stable." I have placed "stable" in quotes because there are two notions of stability used in general relativity. First, a spacetime property is said to be stable if it still occurs when the initial data is perturbed. Second, a property is said to be stable if it persists when the metric is changed slightly at every point in the spacetime. I shall make these different notions of stability more precise in Sec. 5, showing that the singularity symmetry of some of the spacetimes considered in Sec. 3 is a stable property in the first sense, and that the total viciousness of the spacetimes of Sec. 4 is a stable property in the second sense.

The notation of this paper is the same as that of Hawking and Ellis,[6] hereafter denoted HE. I shall assume that the cosmological constant is zero.

2. DEFINITIONS AND BASIC RELATIONSHIPS

The three basic causal sets, $J^+(S)$, $I^+(S)$, and $D^+(S)$ give rise to the following three definitions of symmetry:

Definition: A spacetime (M,g) will be called *causally symmetric* about a set S if $J^+(S)$ is isometric to $J^-(S)$, written

$$J^+(S)\approx J^-(S).$$

Definition: A spacetime (M,g) will be called *chronologically symmetric* about a set S if

$$I^+(S)\approx I^-(S).$$

Definition: A spacetime (M,g) will be called *Cauchy symmetric* about a set S if

$$D^+(S)\approx D^-(S).$$

Since *all* spacetimes have the above symmetries if $S=M$, some restriction will have to be placed on S for the definitions to be useful. In Sec. 3, we will require S to be a partial Cauchy surface, and on this structure we can define another notion of "Cauchy" symmetry:

Definition: A spacetime (M,g) is called *time symmetric* if there exists a partial Cauchy surface at each point of which the extrinsic curvature χ_{ab} vanishes.

This is the definition of time symmetry as given by Misner, Thorne, and Wheeler.[7] There are other definitions of time symmetry in the literature.[8] For example, in Harrison, Thorne, Wakano, and Wheeler we find the following definition of time symmetry: "A spacelike hypersurface is said to be a hypersurface of time symmetry when the dynamical history and the 4-geometry on the future side of this hypersurface is the time-reversed image of the dynamical history and the 4-geometry in the past.[9]" As I interpret this statement, the above authors claim that a spacelike hypersurface S is a surface of time symmetry if the spacetime is both Cauchy symmetric and causally symmetric about S. (The time reversal of the dynamical history gives rise to the Cauchy symmetry and the time reversal of the 4-geometry gives rise to the causal symmetry.) Note, however, that neither $D^+(S)\approx D^-(S)$ nor $J^+(S)\approx J^-(S)$ imply $\chi_{ab}=0$. For example, let S have the topology R^3 and let (x,y,z) be a Euclidean coordinate system on S, with initial data set (h_{ab},χ_{ab}) on S (h_{ab} is the metric on S). Then if $h_{ab}(x,y,z)=h_{ab}(x,y,-z)$ and $\chi_{ab}(x,y,z)=-\chi_{ab}(x,y,-z)$ with $\chi_{ab}\neq 0$ except at points for which $z=0$, we can evolve this data so that $D^+(S)\approx D^-(S)$ and $J^+(S)\approx J^-(S)$. In other words, this initial data set is *globally* time symmetric [i.e., $J^+(S)\approx J^-(S)$ and $D^+(S)\approx D^-(S)$] but not *locally time* symmetric [i.e., for any point p in S with $z\neq 0$, we have $J^+(p)$ *not isometric* to $J^-(p)$].

Furthermore, $\chi_{ab}=0$ on a partial Cauchy surface S does not imply any of the three causal symmetries. For example, remove the point $(x=y=z=0,\ t=+1)$ from Minkowski space. In the resulting spacetime the hypersurface $t=0$ has $\chi_{ab}=0$, but $D^+(S)$ is not isometric to $D^-(S)$. We do, however, have the following:

Proposition 1: If the spacetime (M,g) is time symmetric about S and if $D(S)\equiv D^+(S)\cup D^-(S)$ is the maximal Cauchy development from S, then $D^+(S)\approx D^-(S)$.

This result follows immediately from the existence and uniqueness of the maximal development from S, proven in Chap. 7 of HE. In a similar manner, we prove

Proposition 2: If S is a time symmetric Cauchy surface, then S is causally symmetric about S.

In the next section we will show that singularities develop both to the past and to the future of a maximal hypersurface[10] S provided there is matter present everywhere on S. Intuitively, the notion of "matter present everywhere on S" means "the energy density is nonzero at each point of S." We can make this intuitive notion precise via one of the following conditions:

Definition: The *weak ubiquitous energy condition* is said to hold on a set S if $T_{ab}V^aV^b>0$ for all timelike or null vectors V^a at each point p in S.

Definition: The *strong ubiquitous energy condition* is said to hold on a set S if $(T_{ab}-\frac{1}{2}g_{ab}T)V^aV^b>0$ for all timelike or null vectors V^a at each point p in S.

All observed matter fields obey both of the above conditions at a point p if $T^{ab}\neq 0$ at p. However, there are certain fields which are often used as approximations to actual fields that do not satisfy one or both of the above conditions if $T_{ab}\neq 0$. For example, a null fluid moving *entirely* in the V^a direction would give $T_{ab}V^aV^b=0$ with $T_{ab}\neq 0$. Furthermore, a massive scalar field could violate the strong ubiquitous energy condition while satisfying the weak ubiquitous energy condition (see p. 95 of HE). Since it is unlikely that the matter at a given point would consist *entirely* of radiation moving in one direction, and since a massive scalar field with $T^{ab}\neq 0$ could violate the strong ubiquitous energy condition only at such extremely high densities that we cannot trust the matter equations, it is reasonable to assume that the above energy conditions hold at a point p whenever $T_{ab}\neq 0$ at p.

The condition $T_{ab}\neq 0$ for all $p\in M$ was apparently originally proposed by Aristotle (nature abhors a vacuum), and later defended by numerous authors, among them Leibniz, who supported it with an argument which is cogent even in the world view of general relativity: At any point in spacetime we expect there will be a little randomly oriented radiation present, even in what would otherwise be a perfect vacuum. The microwave background radiation, for example, is expected to be present everywhere in spacetime, except perhaps where there is matter to shield it out. This random background radiation would be sufficient to satisfy both of the ubiquitous energy conditions; even in radiation shielded regions there would be quantum mechanical zero-point radiation which would in itself be sufficient to satisfy the condition. Thus the above ubiquitous energy conditions seem to be eminently reasonable conditions to impose on the whole of spacetime, though for our purposes we will need to impose them only on an initial spacelike hypersurface.

3. SINGULARITY SYMMETRIC SPACETIMES

We will now show that any spacetime which is time symmetric about a spacelike hypersurface S (or more generally, for which S is a maximal hypersurface) and which has matter everywhere on S has singularities both to the future and to the past of S. The first two theorems will apply to the case in which S is compact, and they require no global causality assumption. The third theorem, which handles the noncompact case, *will* require a causality assumption: S is required to be a Cauchy surface. The first theorem is really a special case of the second. It is included separately for two reasons. First of all, it facilitates comparison with a similar theorem by Brill and Flaherty,[11] and second, since its conclusions depend explicitly on the initial data and not on a more general global generic condition, it will be used to prove the stability of a class of singularity symmetric spacetimes.

Theorem 1: Suppose that a spacetime (M,g) contains a maximal spacelike hypersurface S which is compact and edgeless. Then there is at least one timelike geodesic which is incomplete to the future of S, and at least one timelike geodesic which is incomplete to the past of S, provided:

(1) the Einstein equations hold on (M,g);

(2) the strong energy condition holds on (M,g);

(3) the strong ubiquitous energy condition holds on S.

Theorem 2: Suppose that a spacetime (M,g) contains a maximal spacelike hypersurface S which is compact and edgeless. Then there is at least one timelike geodesic which is incomplete to the future of S, and at least one timelike geodesic which is incomplete to the past of S, provided:

(1) the Einstein equations hold on (M,g);

(2) the strong energy condition holds on (M,g);

(3) on every timelike geodesic γ with $\gamma \cap S \neq \phi$, there are points p,q in $J^+(S)$ and $J^-(S)$ respectively such that at p and q, $V^a V^b V_{[c} R_{d]ab[e} V_{f]} \neq 0$, where V^a is the unit tangent vector to γ.

Proof: Clearly Theorem 1 is a special case of Theorem 2, for let $p=q$ be a point in $\gamma \cap S$. Then $R_{ab} V^a V^b > 0$ at $p=q$ by conditions (1) and (2) of Theorem 1. But this implies $V^a V^b V_{[c} R_{d]ab[e} V_{f]} \neq 0$ at $p=q$ (see p. 540 of Ref. 12), so condition (3) of Theorem 2 holds. Thus we need only prove Theorem 2. (The proof is a modification of the proof of Theorem 4 in HE, p. 273.) It can be shown (HE, pp. 204–5) that there exists a covering manifold $\hat{M}$ to M such that each connected component of the image of S is diffeomorphic to S and is a partial Cauchy surface in $\hat{M}$. If there are incomplete timelike geodesics both to the future and to the past of any one connected component $\hat{S}$ of the image of S, then there will be incomplete timelike geodesics both to the future and to the past of S in M. Therefore, the proof can be carried out in $\hat{M}$. We first show that any timelike geodesic γ which intersects $\hat{S}$ orthogonally will have a point conjugate to $\hat{S}$ both to the future and to the past of $\hat{S}$, provided γ can be extended that far. Recall that a point p on γ is said to be conjugate to $\hat{S}$ along γ if there is a Jacobi field along γ which is not identically zero but vanishes at p and satisfies the initial condition

$$V_{a;b} = \chi_{ab} \tag{3.1}$$

at $\hat{S}$ (HE, pp. 96–100). The Jacobi fields along $\gamma(t)$ which satisfy the above initial condition can be written (HE, p. 99)

$$Z^\alpha = A_{\alpha\beta}(t) Z^\beta \big|_q,$$

where $\alpha,\beta = (1,2,3)$, t is the proper time along $\gamma(t)$ [with $t=0$ at q], and q is the point at which γ intersects $\hat{S}$. At q, $A_{\alpha\beta}$ is the unit matrix, and the point p will be conjugate to $\hat{S}$ along $\gamma(t)$ if and only if the determinant of $A_{\alpha\beta}$ vanishes at p. If we define

$$x^3 \equiv \det(A_{\alpha\beta}),$$

$$\theta = \frac{3}{x}\frac{dx}{dt}, \tag{3.2}$$

$$\sigma_{\alpha\beta} = A^{-1}_{\gamma(\beta} \frac{d}{dt} A_{\alpha)\gamma} - \tfrac{1}{3}\delta_{\alpha\beta}\theta,$$

then it can be shown (HE, pp. 96–101) that $A_{\alpha\beta}$ satisfies

$$\frac{d\theta}{dt} = -R_{ab}V^aV^b - 2\sigma^2 - \tfrac{1}{3}\theta^2, \tag{3.3}$$

where $2\sigma^2 = \sigma_{\alpha\beta}\sigma^{\alpha\beta} \geqslant 0$. Using $\theta = (3/x)dx/dt$, (3.3) can be written

$$\frac{d^2x}{dt^2} + F(t)x = 0, \tag{3.4}$$

where

$$F(t) = \tfrac{1}{3}(R_{ab}V^aV^b + 2\sigma^2). \tag{3.5}$$

Since $x^3 = \det(A_{\alpha\beta})$, $\det(A_{\alpha\beta})$ will be zero at p if and only if $x=0$ at p. At q, we have (HE, p. 100)

$$\theta = V^a_{;a} = \chi^a{}_a = \frac{3}{\chi}\frac{dx}{dt} = 0.$$

Thus, showing that any future-complete timelike geodesic $\gamma(t)$ orthogonal to $\hat{S}$ has a point conjugate to $\hat{S}$ to the future of $\hat{S}$ is equivalent to showing that the solution to (3.4) which satisfies the initial conditions

$$x = 1, \quad \frac{dx}{dt} = 0, \tag{3.6}$$

at $t=0$ has a zero in $(0, +\infty)$.

By conditions (1) and (2), $F(t) \geqslant 0$ in $[0, +\infty)$, and condition (3) implies that there exists a value t_1 in $[0, +\infty)$ for which $F(t) > 0$. Thus, from Eq. (3.4), we have a value t_2 for which

$$\frac{dx}{dt}\bigg|_{t_2} = -\int_0^{t_2} F(t)x(t)dt < 0. \tag{3.7}$$

Since $F(t) \geqslant 0$, this means implies a zero of x for some t in $[0, +\infty)$. A similar argument shows that any past-complete timelike geodesic $\gamma(t)$ orthogonal to $\hat{S}$ has a point conjugate to $\hat{S}$ to the past of $\hat{S}$.

By Proposition 7.24 of Penrose,[13] the location of the first conjugate point to $\hat{S}$ on γ varies continuously with the point at which γ intersects $\hat{S}$ and γ. Thus the proper time length to the first conjugate point of $\hat{S}$ along the future-directed timelike geodesics orthogonal to $\hat{S}$ is a continuous function defined on $\hat{S}$, provided all γ are future complete. Thus it attains its maximum value b on the compact set $\hat{S}$: if $\hat{M}$ were timelike geodesically complete to the future of $\hat{S}$, there would be a point conjugate to $\hat{S}$ on every future-directed geodesic orthogonal to $\hat{S}$ within a proper time distance b. But to every point $q \in D^+(\hat{S})$ there is a future-directed geodesic orthogonal to $\hat{S}$ which does not contain any point conjugate to $\hat{S}$ between $\hat{S}$ and q (HE, p. 217). Let $\beta: \hat{S} \times [0,b] \rightarrow M$ be the differentiable map which takes a point $p \in \hat{S}$ a proper time distance $t \in [0,b]$ along the future-directed geodesic through p orthogonal to $\hat{S}$. Then $\beta(\hat{S} \times [0,b])$ would be compact and would contain $\overline{D^+(\hat{S})}$. Since the intersection of a compact set and a closed set is compact, this implies that $\overline{D^+(\hat{S})}$ and hence $H^+(\hat{S})$ would be compact.

Consider now a point $q \in H^+(\hat{S})$. The function $d(\hat{S},q)$ would be less than or equal to b, since every past-directed nonspacelike curve from q to $\hat{S}$ would consist of a (possibly zero) null geodesic segment in $H^+(\hat{S})$ followed by a nonspacelike curve in $D^+(\hat{S})$. [See HE, p. 215 for the definition of $d(\hat{S},q)$.] Since d is lower semicontinuous, there would exist an infinite sequence of points $r_n \in D^+(\hat{S})$ converging to q such that $d(\hat{S},r_n)$ converged to $d(\hat{S},q)$. There would correspond to each r_n at least one element $\beta^{-1}(r_n)$ of $\hat{S} \times [0,b]$. Furthermore, there would be an element $\beta^{-1}(p,t)$ which would be a limit point of the $\beta^{-1}(r_n)$ since $\hat{S} \times [0,b]$ is compact. By continuity we would have $t = d(\hat{S},q)$ and $\beta(p,t) = q$. Hence to every point $q \in H^+(\hat{S})$ there would be a timelike geodesic of length $d(\hat{S},q)$ from $\hat{S}$. Now let $q_1 \in H^+(\hat{S})$ be a point to the past of q on the same null geodesic generator λ of $H^+(\hat{S})$. If we were to join the geodesic of length $d(\hat{S},q_1)$ from $\hat{S}$ to q_1 to the segment of λ between q_1 and q, we would obtain a nonspacelike curve of length $d(\hat{S},q_1)$ from $\hat{S}$ to q which could be varied to give a longer curve between these endpoints (HE, p. 112). Thus the function $d(\hat{S},q)$, with $q \in H^+(\hat{S})$, would strictly decrease along every past-directed generator of $H^+(\hat{S})$. Now these generators have no past endpoints (HE, p. 203). But this contradicts the fact that $d(\hat{S},q)$, $q \in H^+(\hat{S})$, would have a minimum on the compact set $H^+(\hat{S})$ since $d(\hat{S},q)$ is lower semicontinuous in q. Thus some future-directed timelike geodesic orthogonal to $\hat{S}$ must be incomplete. A similar argument with the past-directed geodesics orthogonal to $\hat{S}$ will show that there is at least one timelike geodesic from $\hat{S}$ which is incomplete in the past direction. □

Condition (3) of Theorem 2 is a very weak condition to impose on a spacetime. We can have $V^a V^b V_{[c} R_{d]ab[e} V_{f]} = 0$ along $\gamma \cap J^+(S)$ only if $R_{ab}V^a V^b$ vanishes at every point of $\gamma \cap J^+(S)$, and then only if the Weyl tensor is related in a very particular way to γ ($C_{abcd}V^b V^c = 0$) at *every* point of $\gamma \cap J^+(S)$. Hawking and Penrose have pointed out[12] that for any physically realistic spacetime, this would not even occur at *any* point of *any* γ!

In order to prove singularity symmetry about a maximal *noncompact* spacelike hypersurface, we will need to impose stronger conditions on the spacetime than were necessary in the compact case. We shall need a causality assumption—S will be assumed to be a Cauchy surface—and we shall need to assume that the matter density is bounded away from zero for some finite proper time for all observers which travel on geodesics hitting S orthogonally. A stronger initial condition on the matter than that imposed in the compact case is a necessary condition for singularity symmetry: There are spacetimes for which $T^{ab}V_a V_b > 0$ *everywhere* on a maximal Cauchy surface S and which is singularity-free. An example would be a static, spherical star which carries an electric charge. Thus the following theorem will not apply to asymptotically flat spacetimes. However, we would expect its condition (3) to hold for a spacetime for which the matter density is roughly constant on S.

Theorem 3: Suppose that (M,g) contains a maximal Cauchy surface S. Then there is at least one timelike geodesic which is incomplete to the future of S, and at least one timelike geodesic which is incomplete to the past of S, provided:

(1) the Einstein equations hold on (M,g);

(2) the strong energy condition holds on (M,g);

(3) there exist positive constants a,b such that

$$\left|\int_0^a (T_{ab} - \tfrac{1}{2}g_{ab}T)V^a V^b dt\right| \geq b$$

for *every* timelike geodesic segment $\gamma \cap J^+(S)$ *and* $\gamma \cap J^-(S)$, where γ is a geodesic intersecting S orthogonally and the proper time t along γ is zero at S.

Proof: We first show that every future-directed timelike geodesic γ orthogonal to S has a conjugate point to S within a proper time distance $(a + 3/8\pi b)$. Suppose not. Then we have $x > 0$ in this interval and

$$\begin{aligned}\left.\frac{dx}{dt}\right|_{t=a} &= -\int_0^a F(t)x(t)dt \\ &= -\frac{8\pi}{3}\int_0^a \left((T_{ab} - \tfrac{1}{2}g_{ab}T)V^a V^b + \frac{2\sigma^2}{8\pi}\right)x\, dt \\ &\leq -\frac{8\pi}{3}x(a)\int_0^a (T_{ab} - \tfrac{1}{2}g_{ab}T)V^a V^b\, dt \leq -\frac{8\pi}{3}x(a)b.\end{aligned}$$

Since $dx/dt \leq dx/dt|_{t=a}$ for all $t \geq a$ before the first zero of x, there must be a zero of x within a distance c of $t = a$, where c is defined by

$$\left.\frac{dx}{dt}\right|_{t=a} = -\frac{x(a)}{c}.$$

Thus

$$c = -\frac{x(a)}{dx/dt|_{t=a}} \leq -x(a)/[-(8\pi/3)x(a)b] = \frac{3}{8\pi b}.$$

Hence a zero of x occurs within a distance $(a + 3/8\pi b)$ from S, and this means a point conjugate to S along γ.

From this result and the fact that to each point $q \in D^+(S)$ there is a future-directed timelike geodesic orthogonal to S of proper time length $d(S,q)$ which does not contain any point conjugate to S between S and q, it follows that there is in $D^+(S)$ no future-directed timelike curve from S with proper time length greater than $(a + 3/8\pi b)$. However, all future-directed timelike curves from S remain in $D^+(S)$ since S is a Cauchy surface. Furthermore, all timelike curves intersect S. Thus, *all* timelike geodesics are incomplete in the future direction, and their lengths from S are less than or equal to $(a + 3/8\pi b)$. A similar result holds for the past direction. Since the maximum proper time distance from S in either time direction is $(a + 3/8\pi b)$, no timelike curve has a length greater than $2(a + 3/8\pi b)$. □

We have also proven:

Corollary: All timelike geodesics are both future and past incomplete, and no timelike curve has a proper time length greater than $2(a + 3/8\pi b)$.

Note that Theorem 3 and its Corollary apply to *all* spacetimes with a maximal Cauchy surface S. If S is compact and conditions (1)—(3) of Theorem 1 hold, then conditions (1)—(3) of Theorem 3 hold.

4. TOTALLY VICIOUS SPACETIMES

Definition: A spacetime (M,g) will be called *totally vicious* if $I^+(q) \cap I^-(q) = M$ for some point q in M. [Notice that if $I^+(q) \cap I^-(q) = M$ is true for *one* point q in M, it will be true for *all* points q in M; every point in M can be connected to every other point by both a future-directed and a past-directed timelike curve.]

The Gödel universe, the Kerr—Newman solution with $a^2 + e^2 > m^2$ ($a \neq 0$), and Minkowski space with the hyperplanes $t = 0$ and $t = 1$ identified, are examples of totally vicious spacetimes. Totally vicious spacetimes are causally and chronologically symmetric about any point and any set in the spacetime. One property of such spacetimes is given by:

Proposition 3: A totally vicious spacetime is causally simple.

Proof: Recall that a spacetime is said to be causally simple if for every compact set K contained in M, $J^+(K)$ and $J^-(K)$ are closed (HE, p. 206). Let q be a point in K. Then we have $I^+(q)=M=I^+(K)=J^+(K)$, so $J^+(K)$ is closed. Similarly, $J^-(K)$ is closed. Hence, totally vicious spacetimes are causally simple. □

This proposition constitutes a counterexample to a theorem of Hawking and Sachs[14]: A causally simple spacetime is stably causal. The *stable causality condition* is said to hold on (M,g) if there are no closed timelike lines in both the metric g originally placed on M *and* in all metrics g' on M which are "near" g. (For a precise definition of "near" see Sec. 5.) Clearly totally vicious spacetimes are *not* stably causal.

The proof of the Hawking—Sachs theorem as given by those authors assumes that causally simple spacetimes are *distinguishing* [a spacetime is said to be distinguishing if for all points q and p, $I^-(q)=I^-(p)$ or $I^+(q)=I^+(p)$ implies $q=p$]. However, this condition is *not* included in the usual definition of causally simple, which is apparently the one used by Hawking and Sachs. Proposition 3 is really a defect in the usual definition of causally simple, for we have

Proposition 4: A spacetime (M,g) which contains closed timelike lines but which is not totally vicious is not causally simple.

Proof: Since (M,g) contains closed timelike lines, there is a point $q\in M$ for which $I^+(q)\cap I^-(q)\neq\emptyset$. Since (M,g) is not totally vicious, $\dot{J}^+(q)\cup\dot{J}^-(q)$ is nonempty. Suppose $\dot{J}^+(q)\neq\emptyset$ and let $p\in\dot{J}^+(q)$. If (M,g) were simply causal, then $\dot{J}^+(q)=E^+(q)=J^+(q)-I^+(q)$, so $q\prec p$, but not $q\ll p$. However, we have $q\ll q$ and $q\prec p$, which imply $q\ll p$. Similarly, we can deduce a contradiction between the assumptions $\dot{J}^-(q)\neq\emptyset$ and causal simplicity. Thus if (M,g) were causally simple, $\dot{J}^+(q)\cup J^-(q)$ would have to be empty, and this is impossible. □

5. STABILITY

As mentioned in the Introduction, there are two notions of stability in general relativity. The first is the continued existence of a spacetime property under perturbations of the initial data. To be more precise,

Definition: A spacetime property will be said to be *D stable* (for development stability) about a spacelike hypersurface S with initial data $(h^0_{ab},\chi^0_{ab},\Psi^0_{(i)})$ if the property exists in all spacetimes maximally developed from initial data $(h_{ab},\chi_{ab},\Psi_{(i)})$ in some neighborhood of the initial data $(h^0_{ab},\chi^0_{ab},\Psi^0_{(i)})$ on S in the original spacetime, where $\Psi_{(i)}$ denotes the other fields and their derivatives on S. We use the original metric h^0_{ab} on S to define a distance function and hence a topology on the space of initial data on S. (I.e., we use the C^∞ open topology on this space—see HE, p. 198 and Ref. 15 for more details.) All sets of initial data are required to satisfy the constraint equations.[16] (All h_{ab} are required to be positive definite.)

We have:

Theorem 4: Singularity symmetry is a D-stable property of the initial data described in Theorem 1. That is, given a compact maximal spacelike hypersurface S with $R_{ab}V^aV^b>0$ everywhere, we can perturb the initial data slightly (so that S is no longer maximal, but still $|\chi^a{}_a|<\epsilon$, for some $\epsilon>0$) and still obtain incomplete timelike geodesics both to the past and to the future of S.

Proof: A change in the metric will change $R_{ab}V^aV^b$, but it still will be bounded away from zero on S for a change sufficiently small. Then there is an $\epsilon>0$ such that when $\chi^a{}_a|<\epsilon$, every timelike geodesic intersecting S orthogonally still has a point conjugate to S both to the past and to the future of S, provided every geodesic is both past and future complete. The existence of a geodesic which is incomplete to the future of S and one which is incomplete to the past of S then follows as in the proof of Theorem 2. □

Similarly, we can show that singularity symmetry still occurs if we relax the maximal hypersurface condition of Theorem 3 to $|\chi^a{}_a|<\epsilon$ on S for some $\epsilon>0$, the precise value of ϵ being determined by the constants a and b. However, it is not possible to prove D stability with the initial data of Theorem 3 because we do not know if S would still be a Cauchy surface when the initial data is perturbed; it is not known if global hyperbolicity is a D-stable property about an S with the initial data of Theorem 3. It probably is not; an arbitrarily small amount of electric field added to Schwarzschild initial data can convert the resulting spacetime from Schwarzschild to Reissner—Nordström, and the former is globally hyperbolic while the latter is not.

The second notion of stability in general relativity is the continued existence of a spacetime property under arbitrary, sufficiently small variations in the metric. To make this notion precise, we follow Geroch[17] and introduce a topology on the collection $\mathcal{G}$ of all Lorentz metrics on M. Let $g'_{ab},\tilde{g}_{ab}\in\mathcal{G}$. We will write $g'_{ab}<\tilde{g}_{ab}$ if every vector which is timelike or null with respect to g'_{ab} is timelike with respect to $\tilde{g}_{ab}$. That is, the light cones of $\tilde{g}_{ab}$ are "larger" than those of g'_{ab}. The set of $g_{ab}\in\mathcal{G}$ such that $g'_{ab}<g_{ab}<\tilde{g}_{ab}$ forms a basis for a topology for $\mathcal{G}$: the C^0 open topology (Ref. 15, HE, p. 198).

Definition: A property of spacetime is said to be *G stable* (for global stability) if given any M, the collection of Lorentz metrics on M which have the given property forms an open set in $\mathcal{G}$.

Theorem 5: Total viciousness is a G-stable property of spacetime.

Proof: Let (M,g) be a totally vicious spacetime. Clearly any metric $\tilde{g}_{ab}$ with $g_{ab}<\tilde{g}_{ab}$ is also totally vicious; if we "expand" the light cones at each point, then any closed timelike line in g_{ab} is also a closed timelike line in $\tilde{g}_{ab}$. To show that there exists a metric g'_{ab} with $g'_{ab}<g_{ab}$ for which (M,g') is totally vicious we proceed as follows. Suppose there is no such metric g'_{ab}. Then there exists a point $p\in M$ through which no closed timelike line passes for *any* $g'_{ab}<g_{ab}$. For if there were no such point p then the chronology violating sets would cover (M,g'_{ab}) for some $g'_{ab}<g_{ab}$. But the set of points at which the chronology condition is violated is the *disjoint* union of open sets of the form $I^+(q,g'_{ab})$

$\cap I^-(q,g'_{ab})$, $q \in M$ (HE, p. 189). Thus if these sets covered M they could not be disjoint unless they consisted only of one set; i.e., $I^+(q,g'_{ab}) \cap I^-(q,g'_{ab}) = M$.

However, the existence of such a point p is impossible, because given any closed timelike line γ (timelike in g_{ab}) through p, there is always a metric g'_{ab} with $g'_{ab} < g_{ab}$ for which γ is still everywhere timelike. (Given a timelike curve γ of finite proper time length, we can always "shrink" the light cones at all the points of γ such that γ is *still* everywhere timelike.)

We have a contradiction, and so there must exist a totally vicious spacetime (M, g') with $g'_{ab} < g_{ab}$. □

Totally vicious spacetimes are in a real sense mirror images of stably causal spacetimes: Both classes are G stable, and both are defined by closed timelike lines—the former by their presence, and the latter by their absence.

ACKNOWLEDGMENTS

I should like to thank Professor D. R. Brill for helpful discussions.

[1] R.H. Gowdy, Phys. Rev. Lett. **27**, 826 (1971).

[2] A. Einstein, *The Meaning of Relativity*, fifth ed. (Princeton U. P., Princeton, New Jersey, 1950), p. 107.

[3] C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), pp. 543, 1181.

[4] It should be mentioned, however, that at present the experimental evidence is *against* closure. See J.R. Gott, J.E. Gunn, D.N. Schramn, and B.M. Tinsley, Astrophys. J. **194**, 543 (1974).

[5] P.J.E. Peebles, *Physical Cosmology* (Princeton U. P., Princeton, New Jersey, 1971).

[6] S. W. Hawking and G.F.R. Ellis, *The Large Scale Stmcture of Spacetime* (Cambridge U. P., Cambridge, 1973).

[7] Ref. 3, p. 535.

[8] The first fruitful use of the concept of time symmetry was made by D.R. Brill in Ann. Phys. **7**, 466 (1959), and by H. Araki in Ann. Phys. **7**, 456 (1959). These authors use the phrase "time symmetry" to mean the existence of a spacelike hypersurface S with $\chi_{ab} = 0$ *and* $D^+(S) \simeq D^-(S)$. (For the precise definition see D.R. Brill, thesis, Princeton University, 1959.)

[9] B.K. Harrison, K.S. Thorne, M. Wakano, and J.A. Wheeler, *Gravitation Theory and Gravitational Collapse* (University of Chicago Press, Chicago, 1965), p. 13.

[10] A maximal hypersurface is one for which the trace of the extrinsic curvature vanishes: $\chi^a{}_a = 0$.

[11] D.R. Brill and F.J. Flaherty, Commun. Math. Phys. **50**, 157 (1976).

[12] S. W. Hawking and R. Penrose, Proc. R. Soc. A **314**, 529 (1970).

[13] R. Penrose, *Techniques of Differential Topology in Relativity*, Vol. 7 of the Regional Conference Series in Applied Mathematics (SIAM, Philadelphia, 1972).

[14] S. W. Hawking and R.K. Sachs, Commun. Math. Phys. **35**, 287 (1974).

[15] S. W. Hawking, Gen. Rel. Grav. **1**, 393 (1971).

[16] D.R. Brill and S. Deser, Commun. Math. Phys. **32**, 291 (1973).

[17] R. Geroch, J. Math. Phys. **11**, 437 (1970).

Asymptotic behavior of integral and integrodifferential equations

Rina Ling

Department of Mathematics, California State University, Los Angeles, California 90032

(Received 17 February 1977)

Long time behavior of integral and integrodifferential equations is studied. Some of them are generalizations of the models for the transport of charged particles in a random magnetic field; the solution of the homogeneous integrodifferential equation has an algebraic-logarithmic decay for long times, whereas the solution of the inhomogeneous equation has a slower logarithmic decay.

1. INTRODUCTION

Equations of integral and integrodifferential type occur in various applied fields; see, for example, Refs. 1–4.

In this work, long time behavior of certain integral and integrodifferential equations is studied. Some of them arise as generalizations of the models for the transport of charged particles in a random magnetic field. Models with particular kernels have been investigated in Ref. 1.

Other integral equations with monotonic increasing kernels are studied; papers in the past[2,5] have considered equations with monotonic decreasing kernels. Other integrodifferential equations studied in this work are of a form similar to that in Ref. 3, but not under the requirements that the kernel in the trnasformed equation be in $L_1[0,\infty)$ and that the initial value of the solution be sufficiently small.

2. INTEGRAL AND INTEGRODIFFERENTIAL EQUATIONS FOR THE TRANSPORT OF CHARGED PARTICLES IN A RANDOM MAGNETIC FIELD

In Ref. 1, a model for the cosmic ray flux $F(t;\alpha)$ is given to be the integrodifferential equation

$$\dot{F}(t;\alpha) = -\sigma(t) - \alpha\int_0^t \frac{1}{1+(t-\tau)}F(\tau;\alpha)\,d\tau,$$

$$t>0,\quad \alpha>0,$$

$$F(0;\alpha)=1,$$

where the source $\sigma(t)$ represents the density gradient and α is a small parameter. When $\sigma=0$, the homogeneous integrodifferential equation can be written in the form of a Volterra equation of the second kind, namely,

$$f(t;\alpha) = 1 - \alpha\int_0^t \ln(1+t-\tau)f(\tau;\alpha)\,d\tau.$$

Consider more generally the equation

$$f(t;\alpha) = 1 - \alpha\int_0^t K(t-\tau)f(\tau;\alpha)\,d\tau,$$

where α is a small position parameter. [For simplicity, the argument α will be dropped from the various functions from now on; for example, $f(t;\alpha)$ will be written as $f(t)$. Also $K*f$ will be used to denote the convolution integral of K and f.] The following theorem is on the long time behavior of the solution for fixed α. It would be assumed that all involved derivatives of $K(t)$ are continuous for $0\leq t<\infty$.

Theorem 2.1: If (1) $K(t)>0$, $t>0$, (2) $a/(t+b) \leq K'(t)\leq c/(t+d)$, where a, b, c, d are positive constants, (3) $-a/(t+b)^2\leq K''(t)<0$, and (4) $K''(t)/K'(t)$ is nondecreasing [i.e., $\ln K'(t)$ is a convex function], then $f(t)\sim -1/\alpha abt(\ln t)^2$ as $t\to\infty$.

Proof: The proof is in the same spirit as that in Ref. 1. In a previous paper,[6] it was shown that

$$f(t) = 2\,\mathrm{Re}[b_1 l^{s_0 t}] + I(t),$$

where s_0 is a pole of $[D(S)]^{-1}$,

$$D(s) = s(1+\alpha\bar{K}(s)),$$

$\bar{K}(s)$ is the Laplace transform of $K(t)$,

$$b_1 = \lim_{s\to s_0}[(s-s_0)/D(s)],$$

and

$$I(t) = -\alpha\int_0^\infty \exp(-xt)B(x)\,dx,$$

$$B(x) = abe^{-x}/G(x),$$

$$G(x) = \{x + \alpha abe^{-x}[\ln x + \gamma + e_2(x)]$$
$$- \alpha K(0) + \alpha x\bar{\theta}(-x)\}^2 + (\alpha ab\pi e^{-x})^2,$$

γ is the Euler constant and $e_2(x) = \int_0^x[(e^y-1)/y]\,dy$,

$\bar{\theta}(s)$ is the Laplace transform of $\theta(t)$,

$$0\leq\theta(t)\leq c\ln(1+t/d) - a\ln(1+t/b).$$

Now

$$I(t) = -\alpha\int_0^\infty dx\,ab\exp[-(t+1)x]/G(x),$$

$$G(x) = g^2(x) + (\alpha abe^{-x})^2,$$

$$g(x) = x + \alpha abe^{-x}h(x) - \alpha K(0) + \alpha x\bar{\theta}(-x),$$

$$h(x) = \ln x + \gamma + \int_0^x[(e^y-1)/y]\,dy.$$

Let $z=1+t$, $\mu_1 = 1/z\ln z$, $\mu_2 = (\ln z)/z$, $\xi = zx$; then

$$I = I_1 + I_2 + I_3,$$

where

$$I_1 = -\alpha z^{-1} \int_{\mu_1 z}^{\mu_2 z} \frac{abe^{-\xi}}{G(\xi z^{-1})} d\xi,$$

$$I_2 = -\alpha \int_0^{\mu_1} \frac{ab \exp(-zx)}{G(x)} dx,$$

$$I_3 = -\alpha \int_{\mu_2}^{\infty} \frac{ab \exp(-zx)}{G(x)} dx.$$

Estimate for I_2: Differentiating $G(x)$, we get

$$G'(x) = 2g(x)g'(x) - 2\alpha^2 a^2 b^2 \pi^2 \exp(-2x),$$

where

$$g'(x) = 1 + \alpha ab \exp(-x)h'(x) - \alpha ab \exp(-x)h(x) + \alpha\bar{\theta}(-x) - \alpha x\bar{\theta}'(-x)$$
$$= 1 + \alpha ab/x - \alpha ab \exp(-x)h(x) + \alpha\bar{\theta}(x) - \alpha x\bar{\theta}'(-x).$$

As $x \to 0+$, $g(x) \sim \alpha ab \ln x$ and $g'(x) \sim \alpha ab/x$, so that $G'(x) < 0$ for sufficiently small x and $G(x) \geqslant G(\mu_1)$ for $0 \leqslant x \leqslant \mu_1$ for sufficiently large z. Therefore,

$$|I_2| \leqslant \frac{\alpha ab}{G(\mu_1)} \int_0^{\mu_1} dx$$
$$\sim \frac{\mu_1}{\alpha ab(\ln\mu_1)^2} \sim \frac{1}{\alpha abz(\ln z)^3} \ll \frac{1}{\alpha abz(\ln z)^2}$$

as $z \to \infty$.

Estimate for I_3: Since $G(x) \geqslant (\alpha ab\, \pi \exp(-x))^2$,

$$|I_3| \leqslant \frac{1}{\alpha ab\, \pi^2} \int_{\mu_2}^{\infty} \exp[-(z-2)x]\, dx$$
$$\sim \frac{1}{\alpha ab\, \pi^2 z^2} \ll \frac{1}{\alpha abz(\ln z)^2}$$

as $z \to \infty$.

Estimate for I_1: Note that

$$I_1 = -\frac{1}{\alpha abz(\ln z)^2} - \frac{1}{\alpha abz(\ln z)^2} I_{12} + \frac{1}{\alpha abz(\ln z)^2}[1 - \exp(-\mu_1 z) + \exp(-\mu_2 z)],$$

where

$$I_{12} = \int_{\mu_1 z}^{\mu_2 z} \left[\frac{G(\xi z^{-1}) - \alpha^2 a^2 b^2 \ln^2 z}{G(\xi z^{-1})}\right] \exp(-\xi)\, d\xi.$$

Since for ξz^{-1} small, $G(\xi z^{-1}) \sim [\alpha ab \ln(\xi z^{-1})]^2$, it follows that $|G - \alpha^2 a^2 b^2 \ln^2 z|/G = O(1)$ as $z \to \infty$ uniformly on $[\mu_1 z, \mu_2 z]$ and

$$|I_{12}| = O(1) \int_{\mu_1 z}^{\mu_2 z} \exp(-\xi)\, d\xi = O(1) \quad \text{as } z \to \infty.$$

Also since $\exp(-\mu_1 z) \leqslant 1 - \mu_1 z$,

$$[1 - \exp(-\mu_1 z) + \exp(-\mu_2 z)] \leqslant 1/\ln z + 1/z = O(1)$$
$$\text{as } z \to \infty.$$

Estimate for $I(t)$: From the estimates for I_1, I_2 and I_3, we get

$$I(t) \sim -1/\alpha abz(\ln z)^2$$
$$\sim -1/\alpha abt(\ln t)^2 \quad \text{as } t \to \infty,$$
$$f(t) \sim -1/\alpha abt(\ln t)^2 \quad \text{as } t \to \infty.$$

Consider now the inhomogeneous integrodifferential equation with constant source term $\sigma(t) = \sigma_0$ and kernel $K_1(t)$,

$$\dot{F}(t) = -\sigma_0 - \alpha K_1 * F,$$

the following theorem is on the large time behavior of the solution $F(t)$. It would be assumed that all involved derivatives of $K_1(t)$ are continuous.

Theorem 2.2: If (1) $a/(t+b) \leqslant K_1(t) \leqslant c/(t+d)$, where a, b, c, d are positive constants; (2) $a/(t+b)^2 \leqslant K_1'(t) < 0$ and (3) $K_1'(t)/K_1(t)$ is nondecreasing (i.e., $\log K_1$ is convex), then $F(t) \sim -\sigma_0/\alpha ab \ln t$ as $t \to \infty$

Proof: The proof is in the same spirit as that in Ref. 1. By quadrature,[4] $F(t)$ can be expressed in terms of the solution $f(t)$ to the homogeneous integrodifferential equation,

$$F = f - \sigma_0^* f$$
$$= f - \sigma_0 \tilde{f}, \quad \tilde{f}(t) = \textstyle\int_0^t f(\tau)\, d\tau,$$
$$= f - \sigma_0{}^* (\tilde{f}(\infty) - \textstyle\int_t^{\infty} f(\tau)\, d\tau).$$

By a Tauberian theorem,[7]

$$\tilde{f}(\infty) = \bar{f}(0+),$$

where $\bar{f}(s) = L[f(t); s]$ is the Laplace transform of $f(t)$. But

$$\bar{f}(s) = 1/s(1 + \bar{K}(s)),$$

where $\bar{K}(s) = L[K(t); s]$ is the Laplace transform of the kernel $K(t)$ in Theorem 2.1, namely,

$$\bar{K}(s) = K(0)/s + ab\bar{K}_0(bs) + \bar{\theta}(s),$$

where

$$\bar{K}_0(x) = L[\ln(1+t); x]$$
$$= (e^x/x)E_1(x), \quad E_1(x) = \textstyle\int_x^{\infty} dt/te^t;$$

therefore,

$$\bar{f}(s) = 1/[s + \alpha K(0) + \alpha a \exp(bs)E_1(bs) + \alpha s\bar{\theta}(s)].$$

Since $E_1(z) = -\ln z - \gamma + e_1(z)$, where $e_1(z)$ is analytic, it follows that

$$\bar{f}(s) \sim 1/-\alpha a \ln s \to 0 \quad \text{as } s \to 0+,$$

and $\tilde{f}(\infty) = 0$.

By Theorem 2.1, the following is true:

$$f(t) \sim -1/\alpha abt(\ln t)^2 \quad \text{as } t \to \infty;$$

therefore

$$\textstyle\int_t^{\infty} f(\tau)\, d\tau \sim -1/\alpha ab \int_t^{\infty} d\tau/\tau(\ln\tau)^2 \quad \text{as } t \to \infty$$
$$= -1/\alpha ab(\ln t).$$

Hence $F(t) \sim -\sigma_0/\alpha ab(\ln t)$ as $t \to \infty$.

Remark: The solution of the inhomogeneous integrodifferential equation has a slower logarithmic decay, whereas the solution of the homogeneous integrodifferential equation has an algebraic-logarithmic decay for long times.

3. OTHER INTEGRAL AND INTEGRODIFFERENTIAL EQUATIONS

In this section, long time behavior of the solutions to other integral and integrodifferential equations with monotonic increasing or decreasing kernels is studied.

Consider first the following integral equation with positive increasing kernel $K(t)$:

$$f(t) = \phi(t) - K*f. \qquad (3.1)$$

It would be assumed that all involved derivatives of $\phi(t)$ and $K(t)$ are continuous for $0 \leqslant t < \infty$.

Theorem 3.1: If (1) $K(t) > 0$, $K'(t) > 0$, $K''(t) < 0$, (2) $a^2 \geqslant 4b$, where $a = K(0)$, $b = K'(0)$, and (3) $\phi(\infty)$ is finite, then $f(t) \to 0$ as $t \to \infty$.

Proof: The equation is $f(t) = \phi(t) - K*f$.

Consider the equation with source term 1,

$$g(t) = 1 - K*g.$$

By equivalence theorem,[6]

$$g(t) = \exp(-\gamma t) - L*g,$$

where $L(t) = (a - \gamma)\exp(-\gamma t) + K' * \exp(-\gamma t)$, γ any constant. Since $L'(t) = (\gamma^2 - a\gamma + b)\exp(-\gamma t) + K''*\exp(-\gamma t)$, by choosing $\gamma = [a \pm (a^2 - 4b)^{1/2}]/2$, we get $L(t) > 0$, $L'(t) < 0$. Now let $\psi(t) = \exp(-\gamma t)$; clearly $\psi'(t) \in L_1[0, \infty)$.

It is shown in Ref. 6 that if $K(t) > 0$, $K'(t) > 0$, $K''(t) < 0$, and $a^2 \geqslant 4b$, then the solution $g(t)$ is bounded.

Consider now the following two cases.

Case 1: $K(\infty) = \infty$: Since $\int_0^\infty L(t)\,dt = -1 + K(\infty)/\gamma$, $L \notin L_1[0, \infty)$ and by Theorem 1,[2] $g(\infty) = 0$.

Case 2: $K(\infty) < \infty$: Since $\int_0^\infty L(t)\,dt = -1 + K(\infty)/\gamma$, $L \in L_1[0, \infty)$ and by Theorem 1,[2]

$$g(\infty) = \psi(\infty)/[1 - \int_0^\infty L(t)\,dt] = 0.$$

So in any case, $g(\infty) = 0$.

The original equation is

$$f(t) = \phi(t) - K*f;$$

by the convolution theorem,[8] $f(t)$ is related to $g(t)$ by

$$f(t) = g(t)\phi(0) + \phi' * g(t),$$

so that $f(\infty) = g(\infty)\phi(0) + g(\infty)[\phi(\infty) - \phi(0)] = 0$.

Integrodifferential equations of the form

$$f'(t) = -mf(t) - \int_0^t k(t-\tau)f(\tau)\,d\tau, \quad f(0) = f_0, \qquad (3.2)$$

where m is a constant, can be transformed into integral equations of the form (3.1) studied in Theorem 3.1. It would be assumed that all involved derivatives of $k(t)$ are continuous for $0 \leqslant t < \infty$. Long time behavior of Eq. (3.2) with positive decreasing kernel $k(t)$ is investigated in the following theorem.

Theorem 3.2: If (1) $k(t) > 0$, $k'(t) < 0$ and (2) $m^2 \geqslant 4k(0)$, $m > 0$, then $f(t) \to 0$ as $t \to \infty$.

Proof: The equation is

$$f'(t) = -mf(t) - \int_0^t k(t-\tau)f(\tau)\,d\tau, \quad f(0) = f_0.$$

It can be written in the form

$$f(t) = f_0 - \int_0^t K(t-\tau)f(\tau)\,d\tau,$$

where

$$K(t) = m + \int_0^t k(\tau)\,d\tau.$$

To apply Theorem 3.1, note that $K(t) > 0$, $K'(t) = k(t) > 0$, $K''(t) = k'(t) < 0$, $a = m$ here, $b = k(0)$ and so $a^2 \geqslant 4b$; by Theorem 3.1 with $\phi(t) = f_0$, we obtain $f(\infty) = 0$.

Consider next the integrodifferential equation (3.2) with negative kernel $k(t)$. Under certain additional condition, similar conclusion can be obtained.

Theorem 3.3: If (1) $k(t) < 0$ and (2) $m \geqslant -\int_0^\infty k(t)\,dt$, then $f(t)$ tends to a constant as $t \to \infty$. {In fact, if $K(t) \notin L_1[0, \infty)$, where $K(t) = m + \int_0^t k(\tau)\,d\tau$, then $f(t) \to 0$ as $t \to \infty$. And if $K(t) \in L_1[0, \infty)$, then $\lim_{t\to\infty} f(t)[1 - \int_0^\infty K(t)\,dt] = f_0$.}

Proof: The equation is

$$f'(t) = -mf(t) - \int_0^t k(t-\tau)\,d\tau, \quad f(0) = f_0.$$

As in Theorem 3.2, let $K(t) = m + \int_0^t k(\tau)\,d\tau$; then

$$f(t) = f_0 - \int_0^t K(t-\tau)f(\tau)\,d\tau.$$

To apply Theorem 1 in Ref. 2, note that $K(t) > 0$, $K'(t) = k(t) < 0$. The source term here is constant and so its derivative is in $L_1[0, \infty)$. Finally the fact that $K(t) > 0$, $K'(t) < 0$ and the source term is constant implies that $f(t)$ is bounded, see Ref. 6. By Theorem 1 in Ref. 2, if $K(t) \notin L_1[0, \infty)$, then

$$\lim_{t\to\infty} f(t) = 0;$$

if $K(t) \in L_1[0, \infty)$, then

$$\lim_{t\to\infty} f(t)[1 - \int_0^\infty K(t)\,dt] = f_0.$$

Remark: The assumptions that $m + \int_0^t k(\tau)\,d\tau$ are in $L_1[0, \infty)$ and f_0 is sufficiently small are not required in Theorems 3.2 and 3.3. They are in the hypotheses of Theorem 5 in Ref. 3.

[1] F. B. Hanson, A. Klimas, G. V. Ramanathan, and G. Sandri, J. Math. Phys. 14, 1592—1600 (1973).

[2] S. O. Londen, "On the Solutions of a Nonlinear Volterra Equation," J. Math. Anal. Appl. 39, 564—73 (1972).

[3] R. K. Miller, "On the Linearization of Volterra Integral Equations," J. Math. Anal. Appl. 23, 198—208 (1968).

[4] G. V. Ramanathan and G. Sandri, J. Math. Phys. 10, 1763—73 (1969).

[5] A. Friedman, "On Integral Equations of Volterra Type," J. d'analyse Math. XI, 381—413 (1963).

[6] R. Ling, "Uniformly Valid Solutions to Volterra Integral Equation," Ph.D. thesis, University of Illinois, 1976.

[7] D. V. Widder, *The Laplace Transform* (Princeton U. P., Princeton, New Jersey, 1946).

[8] R. Bellman and K. L. Cooke, Differential-Difference Equations (Academic, New York, 1963).

Group-theoretical foundations of classical and quantum mechanics. I. Observables associated with Lie algebras

L. Martinez Alonso

Departamento de Física Teórica, Universidad Complutense de Madrid, Madrid, Spain

(Received 28 July 1976)

This paper is a first attempt to explore the relationship between classical and quantum mechanics from a group-theoretical point of view. We deal here with the algebraic aspects of the sets of classical and quantum observables in the framework of the algebraic structures associated with finite-dimensional Lie algebras. In particular, we investigate the canonical structure of the quotient fields predicted by the Gel'fand–Kirillov and Vergne conjectures in order to study the types of observables that emerge from a given Lie algebra.

1. INTRODUCTION

In recent years,[1] group-theoretical methods have been applied for the mathematical foundations of classical mechanics. Many of the results that have been obtained are similar to the corresponding quantum ones. The analysis of this fact is essential for a deeper understanding of both classical and quantum mechanics, On the other hand, the passage from one to another context provides an important mathematical framework which includes for example the theory of "geometric quantization" and is related to important problems as the relationship between the Gel'fand—Kirillov[2] and Vergne[3] conjectures.

The aim of this series of papers concerned with the group-theoretical foundations of classical and quantum mechanics is to build up a unified formalism for the construction of the classical and quantum mechanics associated with a connected Lie group $\mathcal{G}$. In particular we shall apply the method to the Galilei, Poincaré, and Weyl Lie groups.

Paper I deals with the problem of the determination of classical and quantum observables associated with a given connected Lie group $\mathcal{G}$. It turns out that the natural mathematical tool for this analysis is the theory of algebraic structures associated with Lie algebras. In this context, symmetric and enveloping algebras[4] play an important role. We give in Sec. 2 a short review of its more important properties, and we introduce the concept of "characteristic dimensions" of a Lie algebra G. In a forthcoming paper we shall see how the characteristic dimensions of the Lie algebra of a connected Lie group $\mathcal{G}$ allow us to know the degrees of freedom of the elementary systems associated with $\mathcal{G}$ and the number of parameters required for labeling them. Section 3 is devoted to the description of the Gel'fand—Kirillov and Vergne conjectures on the rational structures associated with algebraic Lie algebras. In Secs. 4, 5, and 6 the above considerations are applied to the extended Galilei, Poincaré, and Weyl Lie algebras respectively.

2. RATIONAL STRUCTURES

Henceforth G will denote any finite-dimensional Lie algebra over the field $\mathbb{R}$ (real numbers) with the commutations relations

$$[A_\alpha, A_\beta] = \sum_\mu c^\nu_{\alpha\beta} A_\nu$$

in a given basis $B = \{A_\alpha\}_1^N$. We consider now the following rational structures associated with G over the field $\mathbb{C}$ (complex numbers).

A. The symmetric algebra and its quotient field

The complex symmetric algebra[4] S of G is the polynomial ring $\mathbb{C}[a_1, \ldots, a_N]$ in N commutative variables $\{a_\alpha\}_1^N$. The quotient field $D(S)$ of S is the field $\mathbb{C}(a_1, \ldots, a_N)$ of rational functions in the variables $\{a_\alpha\}_1^N$. The Poisson bracket of two elements $h_1, h_2 \in D(S)$ is defined by

$$\{h_1, h_2\} = \sum_{\alpha,\beta,\nu} c^\nu_{\alpha\beta} a_\nu \frac{\partial h_1}{\partial a_\alpha} \frac{\partial h_2}{\partial a_\beta}.$$

The field $D(S)$ equipped with the Poisson bracket defines a Lie algebra. Evidently S is a Lie subalgebra of $D(S)$.

There are two important subsets:

$$S^I = \{p \in S/\{p, p'\} = 0, \ \forall p' \in S\},$$

$$D(S)^I = \{h \in D(S)/\{h, h'\} = 0, \forall h' \in D(S)\},$$

which are the invariants in S and $D(S)$ respectively under the adjoint action of G.

B. The enveloping algebra and its quotient field

The complex enveloping algebra[4] $\mathcal{U}$ of G is the set of all (noncommutative in general) polynomials in N elements $\{A_\alpha\}_1^N$ satisfying the relations

$$A_\alpha A_\beta - A_\beta A_\alpha = i \sum_\nu c^\nu_{\alpha\beta} A_\nu.$$

It is a well-known fact[4] that $\mathcal{U}$ is a Noetherian ring without zero divisors. Then, one can construct its quotient field, denoted $\mathcal{D}(\mathcal{U})$; the elements of $\mathcal{D}(\mathcal{U})$ are of the form uv^{-1} with $u, v \in \mathcal{U}$ and $v \neq 0$. Any pair of fractions $r_1, r_2 \in \mathcal{D}(\mathcal{U})$ can be reduced to a common denominator; therefore, it is possible to define on $\mathcal{D}(\mathcal{U})$ all required operations to make $\mathcal{D}(\mathcal{U})$ a noncommutative field. If u and $v \neq 0$ are two elements of $\mathcal{U}$ such that $uv = vu$, then it follows that $uv^{-1} = v^{-1}u$; in this case we shall denote $u/v = uv^{-1} = v^{-1}u$.

The commutator of two elements $r_1, r_2 \in \mathcal{D}(\mathcal{U})$ is defined in the usual form $[r_1, r_2] = r_1 r_2 - r_2 r_1$. For instance, if $u, v \in \mathcal{U}$ and $v \neq 0$, we get $[u, v^{-1}] = -v^{-1}[u, v]v^{-1}$. One can show that $\mathcal{D}(\mathcal{U})$ is a Lie algebra and $\mathcal{U}$ a Lie subalgebra of $\mathcal{D}(\mathcal{U})$. Given $r_1, r_2 \in \mathcal{D}(\mathcal{U})$, we define the anticommutator by $[r_1, r_2]_+ = r_1 r_2 + r_2 r_1$.

Now, the sets of invariants in $\mathcal{U}$ and $\mathcal{D}(\mathcal{U})$ under the adjoint action of G are given by

$$\mathcal{U}^I = \{u \in \mathcal{U} \mid [u, u'] = 0, \forall u' \in \mathcal{U}\}$$

$$\mathcal{D}(\mathcal{U})^I = \{r \in \mathcal{D}(\mathcal{U}) \mid [r, r'] = 0, \forall r' \in \mathcal{D}(\mathcal{U})\}.$$

The elements of $\mathcal{U}^I$ are the familiar "Casimir invariants" of G.

C. The characteristic dimensions

Let M_G be the matrix with elements $(M_G)_{\alpha\beta} = \sum_\nu c^\nu_{\alpha\beta} a_\nu$. We write

$$r(G) = \sup_{(a_1, \dots, a_N)} \operatorname{rank} M_G.$$

By definition $r(G)$ is always an even integer (rank of an antisymmetric matrix). Let dim$G$ be the vectorial dimension of G; we define the characteristic dimensions of G by the integers:

$$n(G) = \tfrac{1}{2} r(G), \quad s(G) = \dim G - r(G).$$

We call $n(G)$ and $s(G)$, respectively, canonical and central dimensions of G.

It is known[4] that S^I and $\mathcal{U}^I$ are isomorphic algebras; therefore, maximal algebraically independent sets in S^I and $\mathcal{U}^I$ have an equal number of elements τ. In the same way[4] $D(S)^I$ and $\mathcal{D}(\mathcal{U})^I$ are isomorphic fields, we denote τ' the common number of elements of their maximal algebraically independent sets. We have the following theorem[5]:

Theorem 1: (i) $\tau \leq \tau' \leq s(G)$ for every Lie algebra G.

(ii) If G is an algebraic Lie algebra, then $\tau' = s(G)$.

(iii) $[G, G] = G \Rightarrow \tau = \tau' = s(G)$.

In general, equality between τ, τ', and $s(G)$ will not be accessible, because of the existence of nonrational invariant functions over the adjoint action of G.

D. Representations

From the algebraic point of view, the natural spaces for the representations of the symmetric algebras are the spaces $C^\infty(V)$ of complex valued C^∞-functions defined on real symplectic manifolds V. It is well known that $C^\infty(V)$ equipped with the Poisson bracket $\{\ ,\ \}_V$ associated with V defines a Lie algebra. By a representations of S we shall mean a linear mapping $p \rightarrow \tilde{p}$ from S to the C^∞-functions in some symplectic manifold V which has the following two properties:

(i) $\widetilde{p_1 p_2} = \tilde{p}_1 \tilde{p}_2$ for all $p_1, p_2 \in S$,

(ii) $\{\widetilde{p_1, p_2}\} = \{p_1, p_2\}_V$ for all $p_1, p_2 \in S$.

Given $h = p_1/p_2$ $(p_1, p_2 \in S)$ in $D(S)$, we shall denote by $\tilde{h}$ the function $\tilde{p}_1/\tilde{p}_2$. Clearly, not every h in $D(S)$ leads to a function $\tilde{h}$ in $C^\infty(V)$.

Let K be a linear subspace of a complex Hilbert space $\mathcal{H}$. We denote by $\mathcal{O}(K)$ the set of linear operators $A : K \rightarrow K$. By a representation of $\mathcal{U}$ we shall mean a linear mapping $u \rightarrow \tilde{u}$ from $\mathcal{U}$ to $\mathcal{O}(K)$ in some dense domain K such that $\widetilde{u_1 u_2} = \tilde{u}_1 \tilde{u}_2$ for all $u_1, u_2 \in \mathcal{U}$. This condition implies that $[\widetilde{u_1, u_2}] = [\tilde{u}_1, \tilde{u}_2]_K$ denotes the commutator of two elements in $\mathcal{O}(K)$. Given $r = u_1 u_2^{-1}$ $(u_1, u_2 \in \mathcal{U})$ in $\mathcal{D}(\mathcal{U})$ such that u_2 is an invertible operator on K, then we shall denote by $\tilde{r}$ the operator $\tilde{u}_1 \tilde{u}_2^{-1}$ defined on K.

By the nature of their representations, it is evident that S and $D(S)$ are algebraic structures of classical observables. On the other hand, $\mathcal{U}$ and $\mathcal{D}(\mathcal{U})$ form algebraic structures of quantum observables.

3. CANONICAL VARIABLES

Let n, s be two nonnegative integers, we shall denote by $D_{n,s}$ the field $\mathbb{C}(x_1, \dots, x_s, q_1, \dots, q_n, p_1, \dots, p_n)$ of rational functions in $2n + s$ commuting variables $x_1, \dots, x_s$, $q_1, \dots, q_n$, $p_1, \dots, p_n$. The field $D_{n,s}$ has a Lie algebra structure associated with the Poisson bracket given by

$$\{f_1, f_2\} = \sum_i \frac{\partial f_1}{\partial q_i} \frac{\partial f_2}{\partial p_i} - \frac{\partial f_1}{\partial p_i} \frac{\partial f_2}{\partial q_i}, \quad f_1, f_2 \in D_{n,s}.$$

Similarly, we shall denote $\mathcal{D}_{n,s}$ the noncommutative field generated over the field $\mathbb{C}(X_1, \dots, X_s)$ by $2n$ elements $Q_1, \dots, Q_n$, $P_1, \dots, P_n$, satisfying

$$Q_i P_j - P_j Q_i = \delta_{ij}, \quad Q_i Q_j - Q_j Q_i = P_i P_j - P_j P_i = 0.$$

The commutator of two elements $r_1, r_2 \in \mathcal{D}_{n,s}$ is defined by $[r_1, r_2] = r_1 r_2 - r_2 r_1$.

Given a Lie algebra G the algebraic differences between their associated quotient fields $D(S)$ and $\mathcal{D}(\mathcal{U})$ are parallel to those between the fields $D_{n,s}$ and $\mathcal{D}_{n,s}$. We write $D(S) \approx D_{n,s}$ if there is a field isomorphism $\varphi : D(S) \rightarrow D_{n,s}$ such that $\varphi(\{h_1, h_2\}) = \{\varphi(h_1), \varphi(h_2)\}$ for all h_1, h_2 in $D(S)$. On the other hand, we write $\mathcal{D}(\mathcal{U}) \approx \mathcal{D}_{n,s}$ if there is a field isomorphism $\phi : \mathcal{D}(\mathcal{U}) \rightarrow \mathcal{D}_{n,s}$. In this case the algebraic character of ϕ implies that $\phi([r_1, r_2]) = [\phi(r_1), \phi(r_2)]$ for all r_1, r_2 in $\mathcal{D}(\mathcal{U})$.

Gel'fand and Kirillov[2] have established a conjecture on the relationship between the quotient field $\mathcal{D}(\mathcal{U})$ associated with an algebraic[6] Lie algebra G and the standard fields $\mathcal{D}_{n,s}$. Vergne[3] has formulated a corresponding conjecture that can be applied to the quotient field $D(S)$. These conjectures can be described in the following terms:

$$G \text{ algebraic} \Rightarrow \begin{cases} D(S) \approx D_{n,s} & \text{(A)} \\ \mathcal{D}(\mathcal{U}) \approx \mathcal{D}_{n,s} & \text{(B)} \end{cases}$$

n and s being the canonical and the central dimensions of G respectively.

The conjecture (B) was verified by Gel'fand and Kirillov[2] for GL(n), SL(n), every nilpotent G, and in a modified form for G semisimple.[7] Joseph[8] has proved it for G solvable. The conjecture (A) was verified by Vergne[3] for the case G nilpotent. On the other hand, Abellanas and Martinez Alonso[9] have showed the failure of the corresponding versions of these conjectures for the real quotient fields associated with real algebraic Lie algebras. This fact is related to the nature of the field $\mathbb{R}$ of real numbers which is not algebraically closed.

Example: Let G be the SU(2) Lie algebra, with the commutation relations

$$[A_i,A_j]=\epsilon_{ijk}A_k \quad (i,j,k=1,2,3).$$

We have

$$r(G)=\sup_{(a_1,a_2,a_3)}\operatorname{rank}\begin{bmatrix} 0 & a_3 & -a_2 \\ -a_3 & 0 & a_1 \\ a_2 & -a_1 & 0 \end{bmatrix}=2;$$

then $n(G)=s(G)=1$. The quotient field $D(S)$ is generated by three commuting variables $\{a_i\}_1^3$ with the Poisson brackets $\{a_i,a_j\}=\epsilon_{ijk}a_k$. If we adopt the following coordinates:

$$q=a_+, \quad p=a_3/ia_+, \quad x=a_+a_-+a_3^2,$$

where $a_\pm=a_1\pm ia_2$, we see that these coordinates satisfy

$$\{q,p\}=1, \quad \{x,p\}=\{x,q\}=0.$$

On the other hand, the variables $\{a_i\}_1^3$ can be found from $\{x,q,p\}$ via the relations

$$a_+=q, \quad a_-=(x+q^2p^2)/q, \quad a_3=iqp.$$

Then, it is evident that $D(S)\approx D_{1,1}$.

We now turn our attention to the quotient field $\mathcal{D}(\mathcal{U})$, it is generated by three elements $\{A_i\}_1^3$ with the commutations relations $[A_i,A_j]=\epsilon_{ijk}A_k$. If we choose

$$Q=A_+, \quad P=A_3(iA_+)^{-1}, \quad X=\tfrac{1}{2}[A_+,A_-]_+ +A_3^2,$$

where $A_\pm=A_1\pm iA_2$, we have

$$[Q,P]=1, \quad [X,P]=[X,Q]=0.$$

Also, the generators $\{A_i\}_1^3$ can be found from $\{X,Q,P\}$ via the relations

$$A_+=Q, \quad A_-=Q^{-1}(X+(PQ)^2+PQ), \quad A_3=iPQ.$$

Then, we conclude that $\mathcal{D}(\mathcal{U})\approx\mathcal{D}_{1,1}$.

4. NONRELATIVISTIC MECHANICS

Here the relevant Lie algebra is the extended Galilei Lie algebra generated by $\{\mathcal{M},\mathcal{H},\mathcal{P},\mathcal{K},\mathcal{J}\}$ with the commutation relations

$$[\mathcal{J}_i,\mathcal{K}_j]=\epsilon_{ijk}\mathcal{K}_k, \quad [\mathcal{J}_i,\mathcal{P}_j]=\epsilon_{ijk}\mathcal{P}_k, \quad [\mathcal{J}_i,\mathcal{J}_j]=\epsilon_{ijk}\mathcal{J}_k,$$
$$[\mathcal{K}_i,\mathcal{P}_j]=-\delta_{ij}\mathcal{M}, \quad [\mathcal{K}_i,\mathcal{H}]=-\mathcal{P}_i;$$

all other commutators are zero. Now, the characteristic dimensions of G are $n(G)=4$ and $s(G)=3$.

A. Classical observables

The quotient field $D(S)$ is given by the field $\mathbb{C}(m,h,\mathbf{p},\mathbf{k},\mathbf{j})$ of complex rational functions in the commuting variables $\{m,h,\mathbf{p},\mathbf{k},\mathbf{j}\}$ with Poisson brackets

$$\{j_i,k_j\}=\epsilon_{ijk}k_k, \quad \{j_i,p_j\}=\epsilon_{ijk}p_k, \quad \{j_i,j_j\}=\epsilon_{ijk}j_k,$$
$$\{k_i,p_j\}=-\delta_{ij}m, \quad \{k_i,h\}=-p_i.$$

It is well known[5] that the set of rational invariants $D(S)^I$ is given by the subfield generated by the elements

$$x_1=m, \quad x_2=h-\mathbf{p}^2/2m, \quad x_3=[\mathbf{j}+(\mathbf{k}/m)\times\mathbf{p}]^2.$$

If we define

$$\mathbf{q}=-\mathbf{k}/m, \quad \mathbf{s}=\mathbf{j}+(1/m)(\mathbf{k}\times\mathbf{p}),$$

we find the following Poisson brackets:

$$\{j_i,q_j\}=\epsilon_{ijk}q_k, \quad \{q_i,p_j\}=\delta_{ij},$$
$$\{j_i,s_j\}=\{s_i,s_j\}=\epsilon_{ijk}s_k,$$
$$\{q_i,q_j\}=\{q_i,s_j\}=\{p_i,s_j\}=0.$$

In terms of the variables $\{x_1,x_2,\mathbf{q},\mathbf{p},\mathbf{s}\}$ we have the relations

$$\mathbf{k}=-x_1\mathbf{q}, \quad \mathbf{j}=\mathbf{q}\times\mathbf{p}+\mathbf{s}, \quad x_3=\mathbf{s}^2.$$

It is evident that $\mathbf{q}$ and $\mathbf{s}$ have the algebraic properties of the position and spin observables respectively in classical nonrelativistic mechanics. We also note that the quotient field $D(S)$ is equal to $\mathbb{C}(x_1,x_2,\mathbf{q},\mathbf{p},\mathbf{s})$. On the other hand, $\mathbf{s}$ has null Poisson brackets with $\{x_1,x_2,\mathbf{q},\mathbf{p}\}$ and, according to the example discussed in Sec. 3, $\mathbb{C}(\mathbf{s})\approx D_{1,1}$. Then we find $D(S)\approx D_{4,3}$.

In this context, it is easily proved that $D(S)$ does not admit an element with the algebraic properties of a time observable. Indeed, given $t\in D(S)$ such that

$$\{h,t\}=1, \quad \{\mathbf{p},t\}=0,$$

these relations imply that $\{x_2,t\}=1$, which contradicts the fact that x_2 belongs to $D(S)^I$.

B. Quantum observables

Now we consider the quotient field $\mathcal{D}(\mathcal{U})$: It is useful in quantum mechanics to use the complex basis $\{M,H,\mathbf{P},\mathbf{K},\mathbf{J}\}$ with the commutation relations

$$[J_i,K_j]=i\epsilon_{ijk}K_k, \quad [J_i,P_j]=i\epsilon_{ijk}P_k, \quad [J_i,J_j]=i\epsilon_{ijk}J_k,$$
$$[K_i,P_j]=-i\delta_{ij}M, \quad [K_i,H]=-iP_i;$$

all other commutators are zero. The subfield $\mathcal{D}(\mathcal{U})^I$ of rational invariants is generated by the elements

$$X_1=M, \quad X_2=H-\mathbf{P}^2/2M, \quad X_3=[\mathbf{J}+(1/M)\mathbf{K}\times\mathbf{P}]^2.$$

We define

$$\mathbf{Q}=-\mathbf{K}/M, \quad \mathbf{S}=\mathbf{J}+(1/M)\mathbf{K}\times\mathbf{P}.$$

Their commutations relations are

$$[J_i,Q_j]=i\epsilon_{ijk}Q_k, \quad [Q_i,P_j]=i\delta_{ij},$$
$$[J_i,S_j]=[S_i,S_j]=i\epsilon_{ijk}S_k,$$
$$[Q_i,Q_j]=[Q_i,S_j]=[P_i,S_j]=0,$$

and we can write

$$\mathbf{K}=-X_1\mathbf{Q}, \quad \mathbf{J}=\mathbf{Q}\times\mathbf{P}+\mathbf{S}, \quad X_3=\mathbf{S}^2,$$

which are in agreement with the algebraic properties of the position and the spin observables in quantum nonrelativistic mechanics.

It is easily seen $\mathcal{D}(\mathcal{U})\approx\mathcal{D}_{4,3}$ and furthermore that $\mathcal{D}(\mathcal{U})$ does not admit an element with the algebraic properties of a time observable.

5. RELATIVISTIC MECHANICS

The relevant Lie algebra in relativistic mechanics is the Poincaré Lie algebra. We choose the basis $\{\mathcal{H},\mathcal{P},\mathcal{K},\mathcal{J}\}$ with the commutation relations

$$[\mathcal{J}_i,\mathcal{K}_j]=\epsilon_{ijk}\mathcal{K}_k,\ [\mathcal{J}_i,\mathcal{P}_j]=\epsilon_{ijk}\mathcal{P}_k,\ \ [\mathcal{J}_i,\mathcal{J}_j]=\epsilon_{ijk}\mathcal{J}_k,$$
$$[\mathcal{K}_i,\mathcal{P}_j]=-\delta_{ij}\mathcal{H},\ \ [\mathcal{K}_i,\mathcal{H}]=-\mathcal{P},\ \ [\mathcal{K}_i,\mathcal{K}_j]=-\epsilon_{ijk}\mathcal{J}_k,$$

and all other commutators are zero. The characteristic dimensions are $n(G)=4$ and $s(G)=2$. One can prove[10] that $D(S)\approx D_{4,2}$ and $\mathcal{D}(\mathcal{U})\approx\mathcal{D}_{4,2}$, but we shall see now that the physical position and spin observables are outside of these quotient fields.

A. Classical observables

The quotient field $D(S)$ is generated by ten variables $\{h,\mathbf{p},\mathbf{k},\mathbf{j}\}$. Let w^μ be the Pauli—Lubanski quadrivector $(-\mathbf{j}\mathbf{p},\ -h\mathbf{j}+\mathbf{p}\times\mathbf{k})$, and let $m=(h^2-\mathbf{p}^2)^{1/2}$. In terms of the generators of $D(S)$ the physical position and spin observables in classical relativistic mechanics[11] can be written

$$\mathbf{q}=-\frac{\mathbf{k}}{h}+\frac{\mathbf{p}\times\mathbf{w}}{mh(h+m)},\quad \mathbf{s}=-\frac{\mathbf{w}}{\mathbf{m}}+\frac{(\mathbf{w}\mathbf{p})\mathbf{p}}{mh(h+m)}.$$

It is straightforward to verify the following relations:

$$\{j_i,q_j\}=\epsilon_{ijk}q_k,\quad \{q_i,p_j\}=\delta_{ij},$$
$$\{j_i,s_j\}=\{s_i,s_j\}=\epsilon_{ijk}s_k,$$
$$\{q_i,q_j\}=\{q_i,s_j\}=\{p_i,s_j\}=0.$$

Also we have

$$h^2=m^2+\mathbf{p}^2,\quad \mathbf{k}=-\mathbf{q}h+\frac{\mathbf{s}\times\mathbf{p}}{h+m},\quad \mathbf{j}=\mathbf{q}\times\mathbf{p}+\mathbf{s}.$$

However, since $m\notin D(S)$, the components of $\mathbf{q}$ and $\mathbf{s}$ does not belong to $D(S)$. Therefore, it is necessary to consider more general algebraic structures than the field $D(S)$. For example we can construct a suitable algebra S^* as follows. Let S_m be the extended symmetric algebra $\mathbb{C}[m,h,\mathbf{p},\mathbf{k},\mathbf{j}]$, where m is a new variable such that $m\in S_m^I$, and let L be the two-sided ideal generated by the element $m^2-h^2+\mathbf{p}^2$ in S_m. We can define the quotient algebra $S^*=S_m/L$. This algebra is a commutative integral domain[12]; therefore, we can construct its quotient field $D(S^*)$. Obviously the components of $\mathbf{q}$ and $\mathbf{s}$ are elements of $D(S^*)$. We also note that there are two independent invariants in $D(S^*)$ given by m and $\mathbf{s}^2$.

It is easy to check that $D(S^*)$ does not admit an element with the algebraic properties of a time observable. Indeed, given t satisfying

$$\{h,t\}=1,\quad \{\mathbf{p},t\}=0,$$

we obtain $\{m^2,t\}=2h$, which is absurd as $m^2\in D(S^*)^I$.

B. Quantum observables

Let $\{H,\mathbf{P},\mathbf{K},\mathbf{J}\}$ be the generators of the enveloping algebra $\mathcal{U}$. As before we consider an extended structure $\mathcal{U}_M$, with a new generator M such that $M\in\mathcal{U}_M^I$, and we define the quotient algebra $\mathcal{U}^*=\mathcal{U}_M/\mathcal{L}$, where $\mathcal{L}$ is the two-sided ideal generated by the element $M^2-H^2+\mathbf{P}^2$ in $\mathcal{U}_M$. The modified enveloping algebra $\mathcal{U}^*$ is an Ore ring[2,12]; Then there exists a quotient field $\mathcal{D}(\mathcal{U}^*)$ associated with $\mathcal{U}^*$. In this field we can define the elements

$$\mathbf{Q}=-\tfrac{1}{2}\left[\mathbf{K},\frac{1}{H}\right]_+ +\frac{\mathbf{P}\times\mathbf{W}}{MH(H+M)},\quad \mathbf{S}=-\frac{\mathbf{W}}{M}+\frac{(\mathbf{W}\mathbf{P})\mathbf{P}}{MH(H+M)},$$

where $\mathbf{W}$ is the spatial part of the Pauli—Lubanski quadrivector $W^\mu=(-\mathbf{J}\mathbf{P},\ -H\mathbf{J}+\mathbf{P}\times\mathbf{K})$. Performing some routine calculations, we get

$$[J_i,Q_j]=i\epsilon_{ijk}Q_k,\quad [Q_i,P_j]=i\delta_{ij},$$
$$[J_i,S_j]=[S_i,S_j]=i\epsilon_{ijk}S_k,$$
$$[Q_i,Q_j]=[Q_i,S_j]=[P_i,S_j]=0.$$

Moreover, we have the relations

$$H^2=M^2+\mathbf{P}^2,\quad \mathbf{K}=-\tfrac{1}{2}[\mathbf{Q},H]_+ +\frac{\mathbf{S}\times\mathbf{P}}{H+M},\quad \mathbf{J}=\mathbf{Q}\times\mathbf{P}+\mathbf{S}.$$

Obviously, $\mathbf{Q}$ and $\mathbf{S}$ have the physical meaning of the position and the spin respectively in quantum relativistic mechanics.

The invariants in $\mathcal{D}(\mathcal{U}^*)$ are generated by M and $\mathbf{S}^2$. Also, it can be proved using the same arguments as in the case of $D(S^*)$ that there are no time observables in $\mathcal{D}(\mathcal{U}^*)$.

6. OBSERVABLES ASSOCIATED WITH THE WEYL GROUP

The Weyl group Lie algebra is generated by $\{\mathcal{D},\mathcal{P}^\mu,\mathcal{M}^{\mu\nu}=-\mathcal{M}^{\nu\mu}:\ \mu,\nu=0,1,2,3\}$ with the commutation relations

$$[\mathcal{M}^{\lambda\rho},\mathcal{M}^{\mu\nu}]=-(g^{\lambda\mu}\mathcal{M}^{\rho\nu}+g^{\rho\nu}\mathcal{M}^{\lambda\mu}-g^{\rho\mu}\mathcal{M}^{\lambda\nu}-g^{\lambda\nu}\mathcal{M}^{\rho\mu}),$$
$$[\mathcal{M}^{\lambda\rho},\mathcal{P}^\mu]=-(g^{\lambda\mu}\mathcal{P}^\rho-g^{\rho\mu}\mathcal{P}^\lambda),\quad [\mathcal{D},\mathcal{P}^\mu]=-\mathcal{P}^\mu,$$

where $g^{\mu\nu}$ is given by $g^{00}=1$, $g^{ij}=-\delta^{ij}$, $g^{0i}=g^{i0}$. Its characteristic dimensions are $n(G)=5$, $s(G)=1$.

A. Classical observables

The quotient field $D(S)$ is generated by the variables $\{d,p^\mu,m^{\mu\nu}=-m^{\nu\mu}\}$. The invariants are generated by the element w^2/p^2, where w^μ is $\tfrac{1}{2}\epsilon^{\mu\nu\lambda\rho}m_{\mu\nu}p_\rho$. If we define[13,14]

$$r^\mu=(1/p^2)(dp^\mu+m^{\mu\nu}p_\nu),\quad w^{\mu\nu}=(1/p^2)\epsilon^{\mu\nu\lambda\rho}p_\lambda w_\rho,$$

we have the following Poisson brackets relations:

$$\{d,r^\mu\}=r^\mu,\quad \{m^{\lambda\rho},r^\mu\}=-(g^{\lambda\mu}r^\rho-g^{\rho\mu}r^\lambda),$$
$$\{p^\mu,r^\nu\}=g^{\mu\nu},\quad \{r^\mu,r^\nu\}=w^{\mu\nu}/p^2.$$

The Poisson brackets of r^μ with $\{d,p^\mu,m^{\mu\nu}\}$ give the behavior of a space—time observable under Poincarê transformations and dilatations; nevertheless, we have that the different components of r^μ have nonzero Poisson brackets. However, if we define

$$q^\mu=r^\mu+iw^\mu/p^2,\quad s^{\mu\nu}=w^{\mu\nu}+(i/p^2)(p^\mu w^\nu-p^\nu w^\mu),$$

we find that

$$\{d,q^\mu\}=q^\mu,\quad \{m^{\lambda\rho},q^\nu\}=-(g^{\lambda\nu}q^\rho-g^{\rho\nu}q^\lambda),$$
$$\{p^\mu,q^\nu\}=g^{\mu\nu},\quad \{q^\mu,q^\nu\}=\{s^{\mu\nu},q^\rho\}=\{s^{\mu\nu},p^\rho\}=\{s^{\mu\nu},d\}=0,$$
$$\{m^{\lambda\rho},s^{\mu\nu}\}=\{s^{\lambda\rho},s^{\mu\nu}\}=-(g^{\lambda\mu}s^{\rho\nu}+g^{\rho\nu}s^{\lambda\mu}-g^{\rho\mu}s^{\lambda\nu}-g^{\lambda\nu}s^{\rho\mu}).$$

On the other hand, we obtain

$$d=qp,\quad m^{\mu\nu}=q^\mu p^\nu-q^\nu p^\mu+s^{\mu\nu}.$$

Therefore, q^μ is a relativistic space-time observable, and $s^{\mu\nu}$ is a relativistic spin tensor, but they are not real variables.

We see that $D(S)=\mathbb{C}(q^\mu,p^\mu,s^{\mu\nu})$; furthermore, $\mathbb{C}(q^\mu,p^\mu)\approx D_{4,4}$, and $s^{\mu\nu}$ verifies that

$$s^{\mu\nu}=(1/2i)\epsilon^{\mu\nu\lambda\rho}s_{\lambda\rho},\quad s^{\mu\nu}s_{\mu\nu}=-4w^2/p^2,$$

where the first equation shows that $\mathbb{C}(s^{\mu\nu})=\mathbb{C}(s^{ij})$.

Since $\{s_{ij}\}$ generates an SU(2) Lie algebra, it is clear that $\mathbb{C}(s^{\mu\nu}) \approx D_{1,1}$. Therefore, we conclude that $D(S) \approx D_{5,1}$.

B. Quantum observables

The analysis of $\mathcal{D}(\mathcal{U})$ is completely analogous to the one we carried out for $D(S)$. In terms of the complex basis of G we define[14]

$$R^{\mu} = \tfrac{1}{2}[D, P/P^2]_+ + (1/2P^2)[M^{\mu\nu}, P_\nu]_+,$$

$$W^{\mu\nu} = (1/P^2)\epsilon^{\mu\nu\lambda\rho} P_\lambda W_\rho,$$

$$Q^{\mu} = R^{\mu} + iW^{\mu}/P^2, \quad S^{\mu\nu} = W^{\mu\nu} + (i/P^2)(P^{\mu}W^{\nu} - P^{\nu}W^{\mu}).$$

It is straightforward to verify the following commutation relations:

$$[D, Q^{\mu}] = iQ^{\mu}, \quad [M^{\lambda\rho}, Q^{\mu}] = -i(g^{\lambda\mu}Q^{\rho} - g^{\rho\mu}Q^{\lambda}),$$

$$[P^{\mu}, Q^{\nu}] = ig^{\mu\nu}, \quad [Q^{\mu}, Q^{\nu}] = [S^{\mu\nu}, Q^{\rho}] = [S^{\mu\nu}, P^{\rho}] = [S^{\mu\nu}, D] = 0,$$

$$[M^{\lambda\rho}, S^{\mu\nu}] = [S^{\lambda\rho}, S^{\mu\nu}] = -i(g^{\lambda\mu}S^{\rho\nu} + g^{\rho\nu}S^{\lambda\mu} - g^{\rho\mu}S^{\lambda\nu} - g^{\lambda\nu}S^{\rho\mu}).$$

Also, we have

$$D = \tfrac{1}{2}[Q^{\mu}, P_{\mu}]_+, \quad M^{\mu\nu} = Q^{\mu}P^{\nu} - Q^{\nu}P^{\mu} + S^{\mu\nu},$$

and it is easily seen that $\mathcal{D}(\mathcal{U}) \approx \mathcal{D}_{5,1}$.

[1] J. M. Souriau, *Structure de systemes dynamiques* (Dunod, Paris, 1970); E. C. G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974); M. Pauri and G. M. Prosperi, J. Math. Phys. **16**, 1503 (1975); L. Martinez Alonso, J. Math. Phys. **17**, 1177 (1976).

[2] I. M. Gel'fand and A. A. Kirillov, Inst. Hautes Etudes Scient. Publ. Math. **31**, 5 (1966).

[3] M. Vergne, Bull. Soc. Math. France **100**, 301 (1972).

[4] J. Dixmier, *Algebres enveloppantes* (Gauthier-Villars, Paris, 1974).

[5] L. Abellanas and L. Martinez Alonso, J. Math. Phys. **16**, 1580 (1975).

[6] C. Chevalley, *Theorie des groupes de Lie. II* (Hermann, Paris, 1951).

[7] I. M. Gel'fand and A. A. Kirillov, Funkcional. Anal. Prilozen 3, 7 (1969).

[8] A. Joseph, Proc. Am. Math. Soc. **45**, 1 (1974).

[9] L. Abellanas and L. Martinez Alonso, Comm. Math. Phys. **43**, 69 (1975).

[10] L. Martinez Alonso, Tesis Doctoral (Universidad Complutense de Madrid, 1975).

[11] E. C. G. Sudarshan *et al.*, see Ref. 1.

[12] N. Jacobson, *Lectures in Abstract Algebra I* (Van Nostrand, Princeton, N. J., 1951).

[13] H. Bacri, Comm. Math. Phys. 5, 97 (1967).

[14] D. J. Almond, Ann. Inst. H. Poincaré **XIX**, 2, 105 (1973).

Nonlocal interactions: The generalized Levinson theorem and the structure of the spectrum[a]

Roger G. Newton

Physics Department, Indiana University, Bloomington, Indiana 47401

(Received 17 January 1977)

Fredholm theory is applied to the Lippmann–Schwinger equation for nonlocal potentials without spherical symmetry. For a specified large set of trace-class interactions it is proved that when, for real $k\neq 0$, the Fredholm determinant vanishes, k^2 is the energy of a bound state. The point $k=0$ is examined and the analog of the distinction between zero-energy bound states and zero-energy resonances for local central potentials is found. A generalized Levinson theorem is proved.

1. INTRODUCTION AND SUMMARY

In a recent paper[1] the distinction between s-wave zero-energy resonances or "half-bound" states and zero-energy bound states for higher angular momenta, well known for the Schrödinger equation with a local central potential, was generalized to the case of local noncentral potentials. For a large class of such interactions it was also proved that there are no positve-energy exceptional points. As a result, a generalized Levinson theorem followed in its full generality, including the possibility of an extra $\frac{1}{2}\pi$ for a "half-bound" state.

In the present paper we carry out a similar kind of analysis for a large class of nonlocal interactions. We find that all the essential results hold in this case too. Again there may be "half-bound" states of zero energy and they enter similarly into the generalized Levinson theorem. However, in the present case there may be positive-energy bound states, i.e., eigenvalues embedded in the continuous spectrum. We, nevertheless, show that there can be no exceptional points of the second kind on the real axis, except at the origin.

Levinson's theorem was first generalized to nonlocal interactions by Jauch.[2] Other work on such a generalization includes papers by Gourdin and Martin,[3] Martin,[4] Ida,[5] Bertero *et al.*,[6a] Buslaev,[6b] Horwitz and Marchand,[7] Wollenberg,[8] Dreyfus,[9] and Bagchi *et al.*[10] Their results are more restrictive in their assumptions than those of the present paper, and none deals with the full generality of the point $E=0$. The manner in which the bound states of positive energy enter into the generalized Levinson theorem has been a matter of some controversy[8] and we discuss this in Sec. 7.

The assumptions on the interaction that are used in this paper are stated at the beginning of Sec. 2, in (2.2), (2.6), and (2.8). We do not assume V to be invariant under rotations. The paper is organized in parallel to Ref. 1, which we shall refer to as I.

2. THE INTEGRAL EQUATION

We assume that the interaction term V in the Schrödinger equation

$$H\psi=(-\Delta+V)\psi=k^2\psi,$$

where H is regarded as acting in $L^2(\mathbb{R}^3)$, is an operator in the *trace class*. The necessary and sufficient condition for this is that there exist two Hilbert-Schmidt operators V_1 and V_2 such that

$$V=V_1V_2. \tag{2.1}$$

Their kernels will be denoted by $V_i(\mathbf{x},\mathbf{y})$, $i=1, 2$, and $V(\mathbf{x},\mathbf{y})$, so that

$$V(\mathbf{x},\mathbf{y})=\int(d\mathbf{z})\,V_1(\mathbf{x},\mathbf{z})V_2(\mathbf{z},\mathbf{y}) \tag{2.2}$$

and

$$\int(d\mathbf{x})(d\mathbf{y})\,|V_i(\mathbf{x},\mathbf{y})|^2<\infty,\quad i=1,2. \tag{2.3}$$

If we define for almost all $\mathbf{x}$

$$u_i(\mathbf{x})=\left[\int(d\mathbf{y})\,|V_i(\mathbf{x},\mathbf{y})|^2\right]^{1/2}, \tag{2.4}$$

$$v_i(\mathbf{x})=\left[\int(d\mathbf{y})\,|V_i(\mathbf{y},\mathbf{x})|^2\right]^{1/2}, \tag{2.5}$$

then (2.3) is equivalent to u_i, $v_i\in L^2(\mathbb{R}^3)$, $i=1,2$. In addition, we will assume that $u_i, v_i\in L^1(\mathbb{R}^3)$ and that $|\mathbf{x}|u_i$, $|\mathbf{x}|v_i\in L^1(\mathbb{R}^3)$. In other words, our assumptions are (2.2) and[11]

$$u_i,\,v_i\in L^1(\mathbb{R}^3)\cap L^2(\mathbb{R}^3),\quad |\mathbf{x}|u_i,\,|\mathbf{x}|v_i\in L^1(\mathbb{R}^3). \tag{2.6}$$

Some results, as will be noted, will have a more conventional form if we make the somewhat stronger assumption that there is a function $\bar{u}_1(|\mathbf{x}|)$ such that

$$|\mathbf{x}|^2\bar{u}_1\in L^1(0,\infty),\quad |\mathbf{x}|^3\bar{u}_1\in L^1(0,\infty),\quad u_1(\mathbf{x})\leq\bar{u}_1(|\mathbf{x}|). \tag{2.6'}$$

We note that (2.2) implies by Schwarz's inequality that

$$|V(\mathbf{x},\mathbf{y})|\leq u_1(\mathbf{x})v_2(\mathbf{y}). \tag{2.7}$$

It will also be assumed that V is self-adjoint[12] $V=V^\dagger$, and that time-reversal invariance holds, i.e., that V is real,

$$V(\mathbf{x},\mathbf{y})=V(\mathbf{y},\mathbf{x})=V^*(\mathbf{x},\mathbf{y}). \tag{2.8}$$

As in I, the starting point of the analysis is the Lippmann–Schwinger equation

$$\psi=\psi_0+GV\psi, \tag{2.9}$$

where

$$\psi_0(\mathbf{k},\mathbf{x})=\exp(i\mathbf{k}\cdot\mathbf{x})$$

and[13]

$$G(k;\mathbf{x},\mathbf{y})=-\frac{\exp(ik|\mathbf{x}-\mathbf{y}|)}{4\pi|\mathbf{x}-\mathbf{y}|}.$$

[a] Supported in part by the National Science Foundation.

Letting V_2 act on (2.9) gives, by (2.1)

$$\varphi = \varphi_0 + K\varphi, \tag{2.10}$$

where

$$\varphi_0 = V_2\psi_0 \tag{2.11}$$

and

$$K = V_2 G V_1. \tag{2.12}$$

Explicitly

$$\varphi_0(\mathbf{k}, \mathbf{x}) = \int (d\mathbf{y}) V_2(\mathbf{x}, \mathbf{y}) \exp(i\mathbf{k}\cdot\mathbf{y}).$$

Because of (2.6) this defines φ_0 for every real $\mathbf{k}$ as a function of $\mathbf{x}$ in $L^2(\mathbb{R}^3)$. It is shown in Appendix A that for $\mathrm{Im}k \geqslant 0$, K is in the Hilbert-Schmidt class,

$$\mathrm{tr}KK^\dagger < \infty, \tag{2.13}$$

that $\mathrm{tr}K$ exists, and that

$$\lim_{|k|\to\infty} \mathrm{tr}K = \lim_{|k|\to\infty} \mathrm{tr}KK^\dagger = 0. \tag{2.14}$$

As a consequence, Fredholm theory is applicable to (2.10) for $\mathrm{Im}k = 0$ and the Fredholm determinant[14]

$$D(k) = \det(1 - \lambda K)\big|_{\lambda=1} \tag{2.15}$$

exists for $\mathrm{Im}k \geqslant 0$ as an absolutely convergent power series in λ. Furthermore

$$\lim_{|k|\to\infty} D(k) = 1 \tag{2.16}$$

for $\mathrm{Im}k \geqslant 0$. Since also (see Appendix A)

$$\mathrm{tr}\frac{\partial K}{\partial k}\frac{\partial K^\dagger}{\partial k} < \infty \tag{2.17}$$

and $\mathrm{tr}\partial K/\partial k$ exists for $\mathrm{Im}k \geqslant 0$, each term in the power series expansion of D is an analytic function of k and therefore $D(k)$ is an analytic function, regular in the open upper half-plane. On the real axis, $D(k)$ is continuous and differentiable. The self-adjointness of V leads to

$$D(k) = D^*(-k^*). \tag{2.18}$$

The Fredholm alternative assures us that if $k = k_0$ is an exceptional point,

$$D(k_0) = 0, \tag{2.19}$$

then the homogeneous form

$$\varphi = K\varphi \tag{2.20}$$

of (2.10) has a solution $\varphi \in L^2$. If k_0 is real, then (2.10) has a solution $\varphi \in L^2$ if and only if φ_0 is orthogonal to all solutions φ' of the equation

$$\varphi' = K^\dagger\varphi'.$$

Assumption (2.8) implies that then

$$\widetilde{V}_2\varphi'^* = V_1\varphi,$$

where $\varphi = V_2 G\widetilde{V}_2\varphi'^* \in L^2(\mathbb{R}^3)$ solves (2.20). Hence

$$(\varphi', \varphi_0) = (\psi_0, V_1\varphi),$$

and the necessary and sufficient condition for (2.9) to have a solution when (2.19) holds is that for the[15] $\hat{k}$ in (2.9),

$$\int (d\mathbf{x})(d\mathbf{y}) \exp(ik_0\hat{k}\cdot\mathbf{x}) V_1(\mathbf{x}, \mathbf{y})\varphi(\mathbf{y}) = 0. \tag{2.21}$$

For nonexceptional values of k, (2.10) has a unique solution $\varphi \in L^2(\mathbb{R}^3)$. We define

$$\psi = \psi_0 + \psi_s, \tag{2.22}$$

where

$$\psi_s = GV_1\varphi. \tag{2.23}$$

Then ψ satisfies (2.9) and since by Schwarz's inequality

$$\left|\int (d\mathbf{y}) V_1(\mathbf{x}, \mathbf{y})\varphi(\mathbf{y})\right| \leqslant u_1(\mathbf{x})\|\varphi\|_2, \tag{2.24}$$

it follows that

$$|\psi_s(\mathbf{k}, \mathbf{x})| \leqslant C\int (d\mathbf{y}) \frac{u_1(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}. \tag{2.25}$$

Thus, although ψ is not necessarily a continuous function of $\mathbf{x}$, it follows from (2.6) and Sobolev's inequality[16] that integrals of the form

$$\int (d\mathbf{x})\psi'^*(\mathbf{k}, \mathbf{x}) f(\mathbf{x})$$

are absolutely convergent for every $f \in L^q(\mathbb{R}^3)$, $1 \leqslant q \leqslant \frac{3}{2}$. [If we make the stronger assumption (2.6′), then $\psi_s(\mathbf{k}, \mathbf{x})$ is continuous.]

The function ψ can be written in the form

$$\psi(\mathbf{k}, \mathbf{x}) = \exp(i\mathbf{k}\cdot\mathbf{x}) + \frac{\exp(ik|\mathbf{x}|)}{|\mathbf{x}|} T(\hat{x}, \mathbf{k}) + R(\mathbf{x}), \tag{2.26}$$

where

$$T(\hat{k}', \mathbf{k}) = -\frac{1}{4\pi}\int (d\mathbf{x})(d\mathbf{y}) \exp(-ik\hat{k}'\cdot\mathbf{x}) V(\mathbf{x}, \mathbf{y})\psi(\mathbf{k}, \mathbf{y}). \tag{2.27}$$

As $|\mathbf{x}| \to \infty$ the remainder $R(\mathbf{x})$ is shown in Appendix B to be $o(|\mathbf{x}|^{-1})$ in an *average* sense, that is,[17]

$$\int d\hat{\mathbf{x}}\,|R(\mathbf{x})| = o(|\mathbf{x}|^{-1}). \tag{2.28}$$

[If we assume (2.6′), then $R(\mathbf{x}) = o(|\mathbf{x}|^{-1})$.] Equations (2.25) and (2.6) imply that for every nonexceptional real k, $T(\hat{k}', \mathbf{k})$ is well defined as a continuous function of $\hat{k}$ and $\hat{k}'$, and that the generalized optical theorem[18] holds.

3. NEGATIVE-ENERGY BOUND STATES

For $\mathrm{Im}k > 0$ the operator GV is in the Hilbert–Schmidt class. Hence the situation is exactly as in the local case, described in Sec. 3 of I. There is a one-to-one correspondence between negative eigenvalues k_0^2 at which there are nontrivial L^2 solutions of the homogeneous form of (2.9),

$$\psi = GV\psi, \tag{3.1}$$

and zeros k_0 of $D(k)$ in the upper half-plane. Because the spectrum of H is real, $D(k)$ has no zeros in $\mathrm{Im}k > 0$ except on the imaginary axis. It follows from the analyticity of $D(k)$, together with (2.16) and (5.11)–(5.13) below, that the number of negative eigenvalues must be finite. The multiplicity of a zero of $D(k)$ at $k = k_0$ equals the degeneracy of the eigenvalue k_0^2.[19]

4. REAL EXCEPTIONAL $k_0 \neq 0$

If $k_0 \neq 0$, $\mathrm{Im}k_0 = 0$, is an exceptional point (2.19), then there exists a solution $\varphi \in L^2$ of (2.20) which is of the form $\varphi = V_2\psi$, and ψ satisfies the homogeneous form of (2.9). It can be written in the form

$$\psi(\mathbf{x}) = -\left[\exp(ik_0|\mathbf{x}|)/4\pi|\mathbf{x}|\right]$$
$$\times \int (d\mathbf{y})(d\mathbf{z}) \exp(-ik_0\hat{x}\cdot\mathbf{y}) V(\mathbf{y},\mathbf{z})\psi(\mathbf{z}) + R(\mathbf{x}), \qquad (4.1)$$

where $R(\mathbf{x}) \in L^2(\mathbb{R}^3)$, as is shown in Appendix C. Therefore, $\psi \in L^2(\mathbb{R}^3)$ if and only if for almost all $\hat{n}$

$$\int (d\mathbf{y})(d\mathbf{z}) \exp(ik_0\hat{n}\cdot\mathbf{y}) V(\mathbf{y},\mathbf{z})\psi(\mathbf{z}) = 0. \qquad (4.2)$$

As in I, k_0 is called an exceptional point of the first kind if there is a solution ψ of (3.1) for which (4.2) holds for almost all $\hat{n}$; if there is a solution of (3.1) for which this is not true, then k_0 is called an exceptional point of the second kind. Thus, if and only if $k_0 \neq 0$ is an exceptional point of the first kind, k_0^2 is a positive eigenvalue, i.e., a bound state in the continuum. In contrast to the local case in I, such bound states cannot be ruled out in the present situation. Their degeneracy is equal to the "multiplicity" of the corresponding zero of D. (See Appendix D for a proof and the appropriate definition of "multiplicity" in this case.) Because K is compact, this multiplicity is necessarily finite. It follows from this, together with (2.16) and (5.11)–(5.13) below, that the number of positive-energy bound states is finite.

We note that (4.2) is identical with (2.21). Consequently a bound state in the continuum does not prevent the existence of a scattering solution of (2.9). This solution will not be unique, but comparison of (2.27) with (4.2) shows that $T(\hat{k}', \mathbf{k})$ will nevertheless be unique for almost all directions. Thus the T matrix is a continuous function of $k, \hat{k}$, and $\hat{k}'$, even at k_0.

Now suppose that k_0 is an exceptional point of the second kind, so that (4.2) fails for a set of directions $\hat{n}$ of positive Lebesgue measure. That this is impossible is proved exactly as in I by means of the optical theorem.[20] We therefore have the following theorem.

Theorem: If the interaction V satisfies (2.2), (2.6), and (2.8), *then the set of real exceptional points of the second kind contains at most the point* $k=0$.

An implication of this theorem is that the singular continuous spectrum of the Hamiltonian H is empty. (This result is not new. It is equivalent to the theorem of Kato mentioned in Ref. 16.)

5. THE POINT $k = 0$

The procedure here is exactly as in Sec. 5 of I, and we will not repeat all of it. If

$$D(0) = 0, \qquad (5.1)$$

then there are solutions $\varphi \in L^2$ and $\chi \in L^2$ of

$$K_0\varphi = \varphi, \qquad (5.2)$$

$$B_0\chi = \chi, \qquad (5.3)$$

where

$$B_0 = -AVA, \qquad (5.4)$$

$$A = (-G_0)^{1/2}, \qquad (5.5)$$

and G_0 and K_0 are G and K for $k=0$. To every solution χ of (5.3) there corresponds

$$\varphi = V_2A\chi \qquad (5.6)$$

which solves (5.2), and

$$(\varphi, \varphi) = (\chi, AV_2^2A\chi).$$

But

$$\mathrm{tr}(AV_2^2A)^2 = \mathrm{tr}(G_0V_2^2)^2 \leqslant \left(\int (d\mathbf{x})(d\mathbf{y}) \frac{u_2(\mathbf{x})u_2(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\right)^2 < \infty$$

by the Sobolev inequality and (2.6). Consequently φ of (5.6) is in $L^2(\mathbb{R}^3)$. Conversely, if $\varphi \in L^2(\mathbb{R}^3)$ solves (5.2), then

$$\chi = AV_1\varphi \qquad (5.7)$$

solves (5.3), and

$$(\chi, \chi) = -(\varphi, V_1G_0V_1\varphi).$$

But

$$\mathrm{tr}(V_1G_0V_1)^2 \leqslant \left(\int (d\mathbf{x})(d\mathbf{y}) \frac{u_1(\mathbf{x})u_1(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\right)^2 < \infty$$

by (2.6) and the Sobolev inequality. Thus $\chi \in L^2$, and there is a one-to-one correspondence between nontrivial L^2-solutions of (5.2) and (5.3).[21]

We then define

$$\psi = -A\chi = G_0V_1\varphi \qquad (5.8)$$

and it obeys the homogeneous equation

$$\psi = G_0V\psi. \qquad (5.9)$$

It is not identically zero, and (2.25), as well as the discussion following it, apply to it.

The remaining arguments of I, Sec. 5, go through exactly as before. The necessary estimates are given in Applendix E. We state the results as follows.

If (5.1) holds, then there are only the following three possibilities:

Case 1. Equation (5.3) has $n \geqslant 1$ linearly independent nontrivial solutions and all of the solutions ψ of (5.9) that correspond to them by (5.8) obey the analog of (4.2) for $k_0 = 0$,

$$\int (d\mathbf{y})(d\mathbf{z}) V(\mathbf{y},\mathbf{z})\psi(\mathbf{z}) = 0. \qquad (5.10)$$

Then 1 is an n-fold degenerate eigenvalue of G_0V, zero is an n-fold degenerate eigenvalue of H, and $k=0$ is an exceptional point of the first kind. The number n is necessarily finite. The behavior of D in the vicinity of $k=0$ for $\mathrm{Im}k \geqslant 0$ then is

$$D(k) = ck^{2n} + o(k^{2n}), \qquad (5.11)$$

where c is real and $c \neq 0$. The Lippmann–Schwinger equation (2.9) now has a well-defined solution for $k=0$ and the scattering cross section is finite.

Case 2. Equation (5.3) has exactly one nontrivial solution χ, and the function ψ that corresponds to it by (5.8) does not satisfy (5.10). Then 1 is not an eigenvalue of G_0V, zero is not an eigenvalue of H, we call ψ a half-bound state, and $k=0$ is an exceptional point of the second kind. The behavior of D near $k=0$, $\text{Im}k \geq 0$, is in this case

$$D(k) = ick + o(k), \tag{5.12}$$

where c is real and $c \neq 0$. Equation (2.9) now has no solution for $k=0$ and the scattering cross section is infinite.

Case 3. This is a combination of cases 1 and 2. Equation (5.3) has $n+1>1$ linearly independent nontrivial solutions such that the ψ's that correspond via (5.8) to the first n of them satisfy (5.10), but the ψ that corresponds to the last, does not. Then 1 is an n-fold degenerate eigenvalue of G_0V, there is an n-fold degenerate bound state of zero energy, and in addition there is a half-bound state. The point $k=0$ is now an exceptional point of the first and second kind. The behavior of D near $k=0$, $\text{Im}k \geq 0$, now is

$$D(k) = ick^{2n+1} + o(k^{2n+1}), \tag{5.13}$$

where c is real and $c \neq 0$. Equation (2.9) now has no solution for $k=0$ and the scattering cross section is infinite.

As for the resolvent $(k^2 - H)^{-1}$, or the complete Green's function, we note (as in I) that in cases 2 and 3 it diverges as k^{-1} at $k \to 0$.

6. THE DETERMINANT OF THE *S* matrix

The S matrix is defined by

$$S = 1 + \frac{ik}{2\pi} T$$

as an integral operator on the unit sphere, where the kernel of T is given by (2.27). As in I we now derive

$$\det S = D^*/D. \tag{6.1}$$

The extra exponential term of I (6.4) now does not arise because D is the unmodified Fredholm determinant. Since for large $|k|$, $\text{Im}k = 0$,

$$\frac{k}{2\pi}\text{tr}T = \frac{1}{2\pi i}\int (d\mathbf{x})(d\mathbf{y}) \frac{V(\mathbf{x},\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} \sin k|\mathbf{x}-\mathbf{y}| + o(1) \tag{6.2}$$

and by the Sobolev inequality and (2.6) this remains finite as $k \to \infty$, (6.1) is consistent with (2.16) and the exponential factor of I (6.4) is not needed.

Note that (6.1), (5.12), and (5.13) show that if $k=0$ is an exceptional pont of the second kind (cases 2 or 3), then

$$\det S(0) = -1, \tag{6.3}$$

whereas otherwise

$$\det S(0) = 1. \tag{6.4}$$

We define

$$\eta(k) = \arg D(k) \tag{6.5}$$

uniquely in the first quadrant by specifying

$$\lim_{|k| \to \infty} \eta(k) = 0 \tag{6.6}$$

as we may by (2.16). Because D is analytic and has no zeros in the open first quadrant, η is continuous there. On the real and imaginary axes D is continuous; hence so is η, except for the points at which $D=0$. If there is an n-fold degenerate positive-energy bound state at $k=k_0>0$, then D has an n-fold zero there and η has a downward discontinuity of πn. We may define the phase $\delta(k)$ on the positive real axis so that

$$\delta(k) = -\eta(k) (mod\, \pi) \tag{6.7}$$

and is *continuous* for all $0<k<\infty$. If we require

$$\lim_{k\to\infty} \delta(k) = 0, \tag{6.8}$$

then it follows that

$$\delta(0) = \lim_{k\to 0+} \delta(k) = -\lim_{k\to 0+} \eta(k) + \pi n_+, \tag{6.9}$$

where n_+ is the number of positive-energy eigenvalues, each counted as many times as its degeneracy.

Equations (6.1) and (6.7) lead to

$$\det S = \exp(2i\delta), \tag{6.10}$$

which, together with (6.8) and its continuity defines δ uniquely in terms of S for all $0 \leq k < \infty$.

7. THE GENERALIZED LEVINSON THEOREM

Since (2.16) forces D to be real on the positive imaginary axis we can immediately state the value of its phase η at $k=\epsilon$, arrived at via the first quadrant to the right of the imaginary axis. If n_- is the total number of negative-energy bound states (including their degeneracies) so that D has n_- zeros (including their multiplicities) on the positive imaginary axis, and if D has an n_0-fold zero at $k=0$, then by (6.6)

$$\lim_{\epsilon\to 0+} \eta(\epsilon) = -\pi(n_- + \tfrac{1}{2}n_0). \tag{7.1}$$

According to (6.9) it follows that

$$\delta(0) = \pi(n_+ + n_- + \tfrac{1}{2}n_0). \tag{7.2}$$

This can be written

$$\delta(0) = \pi(n + \tfrac{1}{2}q), \tag{7.3}$$

where n is the total number of bound states (including their degeneracy and including those of positive, negative, and zero energy), and $q=1$ if there is a half-bound state at $k=0$; $q=0$ otherwise.

We remark that the generalized Levinson theorem (7.3) looks somewhat different from its statement by some other authors.[8] The question of whether the positive-energy bound states are to be included on the right-hand side has been a matter of some confusion. We emphasize that (7.3) is based on a definition of δ that makes it *continuous* for all $0<k<\infty$. That such a definition is possible follows from the continuity of D, which we have proved. That it is desirable follows from the fact that a positive-energy bound state at k_0 is not at all recognizable from the scattering near k_0. As we have seen in Sec. 4, it has no effect on the T matrix.

That the positive-energy bound states must be included in (7.3) in the way they are can be recognized heuristically as follows. Assume first that V is such that there are no positive-energy bound states, but that there are some sharp resonances. We then change V so as to maneuver the resonance zeros of D onto the real axis. In this limit, what previously was a sharp upward change of δ by π at each resonance becomes an upward discontinuity. We now redefine δ to be continuous. Then the redefined phase must differ at $E=0$ from the old value $\pi(n_- + \frac{1}{2}q)$ by $n_+\pi$, and we get (7.3).

APPENDIX A

We first want to show (2.13). We have

$$\mathrm{tr}KK^\dagger = \int (d\mathbf{x})(d\mathbf{y})(d\mathbf{z})(d\mathbf{x}')(d\mathbf{y}')(d\mathbf{t})\, V_2(\mathbf{z},\mathbf{x}) V_2^*(\mathbf{x}',\mathbf{z})$$

$$\times V_1(\mathbf{y},\mathbf{t}) V_1^*(\mathbf{t},\mathbf{y}') \frac{\exp(ik|\mathbf{x}-\mathbf{y}| - ik^*|\mathbf{x}'-\mathbf{y}'|)}{16\pi^2 |\mathbf{x}-\mathbf{y}|\,|\mathbf{x}'-\mathbf{y}'|}. \tag{A1}$$

The integral will be shown to be absolutely convergent and hence the order of integrations is immaterial. For the $\mathbf{z}$ and $\mathbf{t}$ integrals we have by Schwarz's inequality

$$\left|\int (d\mathbf{z}) V_i(\mathbf{z},\mathbf{x}) V_i^*(\mathbf{x}',\mathbf{z})\right| \leqslant u_i(\mathbf{x}') v_i(\mathbf{x}).$$

Thus (A1) becomes, for $\mathrm{Im}k \geqslant 0$,

$$\mathrm{tr}KK^\dagger \leqslant \left(\int (d\mathbf{x})(d\mathbf{y}) \frac{v_2(\mathbf{x}) u_1(\mathbf{y})}{4\pi|\mathbf{x}-\mathbf{y}|}\right)^2.$$

Since (2.6) implies that $v_2, u_1 \in L^{6/5}(\mathbb{R}^3)$ the Sobolev inequality[16] implies that the right-hand side is finite, and (2.13) follows.

The above argument shows that the $(\mathbf{x},\mathbf{y})$ and the $(\mathbf{x}',\mathbf{y}')$ integrands in (A1) are in L^1. We then merely change variables,

$$\mathbf{x}-\mathbf{y} = \mathbf{X}, \quad \mathbf{x}+\mathbf{y} = \mathbf{Y},$$

and are able to conclude from Lebesgue's lemma that the right-hand side of (A1) vanishes as $\mathrm{Re}k \to \pm\infty$. As $\mathrm{Im}k \to \infty$ the same result follows from Lebesgue's dominated convergence theorem. Thus the second part of (2.14).

We have

$$\mathrm{tr}K = -\frac{1}{4\pi}\int (d\mathbf{x})(d\mathbf{y})(d\mathbf{z})\, V_2(\mathbf{x},\mathbf{y}) \frac{\exp(ik|\mathbf{y}-\mathbf{z}|)}{|\mathbf{y}-\mathbf{z}|} V_1(\mathbf{z},\mathbf{x}).$$

Therefore, for $\mathrm{Im}k \geqslant 0$,

$$|\mathrm{tr}K| \leqslant \frac{1}{4\pi}\int (d\mathbf{y})(d\mathbf{z}) \frac{v_2(\mathbf{y}) u_1(\mathbf{z})}{|\mathbf{y}-\mathbf{z}|} < \infty$$

again by the Schwarz and Sobolev inequalities. The same argument as above then allows us to conclude that $\mathrm{tr}K \to 0$ as $\mathrm{Re}k \to \pm\infty$ or as $\mathrm{Im}k \to \infty$. Thus the first part of (2.14).

Equation (2.16) follows from (2.14) because each term in the power series expansion of $\det(1-\lambda K)$ uniformly tends to zero.

Next we consider

$$\mathrm{tr}\frac{\partial K}{\partial k} = \frac{1}{4\pi i}\int (d\mathbf{x})(d\mathbf{y})(d\mathbf{z})\, V_2(\mathbf{x},\mathbf{y}) \exp(ik|\mathbf{y}-\mathbf{z}|) V_1(\mathbf{z},\mathbf{x})$$

and hence for $\mathrm{Im}k \geqslant 0$, by Schwarz's inequality,

$$\left|\mathrm{tr}\frac{\partial K}{\partial k}\right| \leqslant \frac{1}{4\pi}\int (d\mathbf{y})(d\mathbf{z}) v_2(\mathbf{y}) u_1(\mathbf{z}) < \infty$$

by assumption (2.6). Similarly,

$$\mathrm{tr}\frac{\partial K}{\partial k}\frac{\partial K^\dagger}{\partial k}$$

$$= \left(\frac{1}{4\pi}\right)^2 \int (d\mathbf{x})(d\mathbf{y})(d\mathbf{z})(d\mathbf{x}')(d\mathbf{y}')(d\mathbf{z}')\, V_2(\mathbf{z},\mathbf{x}) V_2^*(\mathbf{x}',\mathbf{z})$$

$$\times V_1(\mathbf{y},\mathbf{z}') V_1^*(\mathbf{z}',\mathbf{y}') \exp(ik|\mathbf{x}-\mathbf{y}| - ik^*|\mathbf{x}'-\mathbf{y}'|)$$

$$\leqslant \left(\frac{1}{4\pi}\right)^2 \left[\int (d\mathbf{x})(d\mathbf{y}) v_2(\mathbf{x}) u_1(\mathbf{y})\right]^2 < \infty$$

by (2.6).

APPENDIX B

The remainder $R(\mathbf{x})$ in (2.26) consists of two terms, $R = R_1 + R_2$,

$$R_1(\mathbf{x}) = -\frac{1}{4\pi}\int (d\mathbf{y})(d\mathbf{z}) \exp(ik|\mathbf{x}-\mathbf{y}|)$$

$$\times V_1(\mathbf{y},\mathbf{z})\varphi(\mathbf{z})\left(\frac{1}{|\mathbf{x}-\mathbf{y}|} - \frac{1}{|\mathbf{x}|}\right),$$

$$R_2(\mathbf{x}) = -\frac{1}{4\pi|\mathbf{x}|}\int (d\mathbf{y})(d\mathbf{z})\, V_1(\mathbf{y},\mathbf{z})\varphi(\mathbf{z})$$

$$\times [\exp(ik|\mathbf{x}-\mathbf{y}|) - \exp(ik(|\mathbf{x}| - \hat{x}\cdot\mathbf{y}))].$$

For the first term we have by (2.24)

$$|\mathbf{x}|\,|R_1(\mathbf{x})| \leqslant C\int (d\mathbf{y}) u_1(\mathbf{y}) \left|\frac{|\mathbf{x}|}{|\mathbf{y}-\mathbf{x}|} - 1\right|$$

$$\leqslant C\int (d\mathbf{y})|\mathbf{y}| \frac{u_1(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}.$$

If we assume (2.6′), then the fact that

$$\frac{1}{4\pi}\int d\hat{y}\, |\mathbf{x}-\mathbf{y}|^{-1} = \begin{cases} \dfrac{1}{|\mathbf{x}|} & \text{if } |\mathbf{x}| > |\mathbf{y}|, \\ \dfrac{1}{|\mathbf{y}|} & \text{if } |\mathbf{x}| < |\mathbf{y}|, \end{cases} \tag{B1}$$

readily leads to $|\mathbf{x}| R_1(\mathbf{x}) = o(1)$ as $|\mathbf{x}| \to \infty$. On the other hand, if we assume only (2.6) then this may be true only in an average sense,

$$\int (d\hat{x})|\mathbf{x}|\,|R_1(\mathbf{x})| \leqslant C\int (d\mathbf{y})|\mathbf{y}| u_1(\mathbf{y}) \int (d\hat{x})|\mathbf{x}-\mathbf{y}|^{-1}$$

$$= C\left(\int_0^{|\mathbf{x}|} (d\mathbf{y})|\mathbf{y}| u_1(\mathbf{y}) |\mathbf{x}|^{-1} + \int_{|\mathbf{x}|}^{\infty} (d\mathbf{y}) u_1(\mathbf{y})\right)$$

$$= o(1).$$

In R_2 we use the estimate

$$|\exp(i|k|\,|\mathbf{x}-\mathbf{y}|) - \exp[i|k|(|\mathbf{x}| - \hat{x}\cdot\mathbf{y})]|$$

$$\leqslant C\frac{|\mathbf{y}|^2}{|\mathbf{x}| + |\mathbf{y}| + |\mathbf{y}|^2} \tag{B2}$$

and (2.24) to get

$$|\mathbf{x}|\,|R_2(\mathbf{x})|$$

$$\leqslant C\int (d\mathbf{y}) u_1(\mathbf{y}) \frac{|\mathbf{y}|^2}{|\mathbf{x}| + |\mathbf{y}| + |\mathbf{y}|^2}$$

$$\leqslant C\Big[|\mathbf{x}|^{-1/2}\int_{|\mathbf{y}|<|\mathbf{x}|^{1/2}} (d\mathbf{y}) u_1(\mathbf{y})|\mathbf{y}|$$

$$+\int_{|\mathbf{y}|>|\mathbf{x}|^{1/2}} (d\mathbf{y}) u_1(\mathbf{y}) |\mathbf{y}| = o(1).$$

APPENDIX C

We want to show here that in (4.1) $R \in L^2(\mathbb{R}^3)$. We have $R = R_1 + R_2$, where R_1 and R_2 are given in Appendix B. We proceed as in Appendix A of I. From (2.24),

$$|R_1(\mathbf{x})| \leqslant C \int (d\mathbf{y}) u_1(\mathbf{y}) \left| \frac{1}{|\mathbf{x}-\mathbf{y}|} - \frac{1}{|\mathbf{x}|} \right|.$$

Therefore, by Schwarz's inequality, (A6) of I, and (2.6)

$$\int (d\mathbf{x}) |R_1(\mathbf{x})|^2 \leqslant C[\int (d\mathbf{y}) u_1(\mathbf{y}) |\mathbf{y}|^{1/2}]^2 < \infty. \tag{C1}$$

In R_2 we use (2.2) and (B2),

$$|R_2(\mathbf{x})| \leqslant C \int (d\mathbf{y}) u_1(\mathbf{y}) \frac{1}{|\mathbf{x}|} \frac{|\mathbf{y}|^2}{|\mathbf{x}| + |\mathbf{y}| + |\mathbf{y}|^2}.$$

Therefore, by Schwarz's inequality, (A9) of I, and (2.6)

$$\int (d\mathbf{x}) |R_2(\mathbf{x})|^2 \leqslant C[\int (d\mathbf{y}) u_1(\mathbf{y}) |\mathbf{y}|]^2 < \infty. \tag{C2}$$

APPENDIX D

We want to examine here the question whether the degeneracy of a positive-energy bound state equals the multiplicity of the corresponding zero of $D(k)$ at a real value of k, at which D need not be analytic. The proof for negative eigenvalues is as follows: For Im$k > 0$

$$\frac{d}{dk} \ln D = \frac{d}{dk} \ln \det(1 - GV)$$

$$= -\operatorname{tr}(1 - GV)^{-1} \frac{dG}{dk} V = \operatorname{tr} \mathcal{G} G V,$$

where $\mathcal{G} = (k^2 - H)^{-1}$. Hence if $D(k_0) = 0$,

$$\lim_{k \to k_0} (k - k_0) \frac{d}{dk} \ln D = n \tag{D1}$$

if n is the degeneracy. If D is analytic at k_0, it follows that

$$D(k) = (k - k_0)^n [c + o(1)]$$

near $k = k_0$.

For real k_0, (D1) still follows for $k \to k_0$ via the upper half-plane. Consequently

$$\frac{d}{dk} \ln D = (k - k_0)^{-1} [n + g(k)],$$

where $g(k) = o(1)$ as $k \to k_0$. Therefore,

$$\ln D = \ln(k - k_0)[n + f(k)] + c + o(1), \tag{D2}$$

where $f(k) = o(1)$. We can therefore assert that the zero of D at $k = k_0$ is of order n in the sense that near $k = k_0$, for Im$k > 0$,

$$D(k) = (k - k_0)^{m(k)} [c + o(1)], \tag{D3}$$

where $m(k) = n + o(1)$.

Equation (D2) shows that, as k goes around k_0 from $k_0 - \epsilon$ to $k_0 + \epsilon$ in the upper half-plane along a semicircle of vanishing radius, the phase of $D(k)$ changes by $-n\pi$. This is the essential result needed for the proof of Levinson's theorem.

APPENDIX E

In this Appendix we will supply the estimates needed to duplicate the contents of I, Sec. 5, that is, the estimates of I, Appendix B.

We have

$$\int (d\mathbf{x})(d\mathbf{y})(d\mathbf{z})(d\mathbf{t}) \frac{|V(\mathbf{x},\mathbf{y}) V(\mathbf{t},\mathbf{z}) \psi(\mathbf{x}) \psi(\mathbf{z})| \, |\mathbf{y}-\mathbf{t}|}{1 + |k| \, |\mathbf{y}-\mathbf{t}|}$$

$$\leqslant C \int (d\mathbf{y})(d\mathbf{t}) |\mathbf{y}-\mathbf{t}| v_2(\mathbf{y}) u_1(\mathbf{t}) < \infty$$

by (2.24) and its analog, and (2.6). Equation (I.B2) follows.

We also have

$$\int (d\mathbf{x})(d\mathbf{y})(d\mathbf{z})(d\mathbf{t}) \frac{|V(\mathbf{x},\mathbf{y}) V(\mathbf{t},\mathbf{z}) \psi(\mathbf{x}) \psi(\mathbf{z})| \, |k| \, |\mathbf{y}-\mathbf{t}|^2}{1 + |k| \, |\mathbf{y}-\mathbf{t}|}$$

$$\leqslant C \int (d\mathbf{y})(d\mathbf{t}) \frac{|k| \, |\mathbf{y}-\mathbf{t}|^2}{1 + |k| \, |\mathbf{y}-\mathbf{t}|} v_2(\mathbf{y}) u_1(\mathbf{t})$$

$$\leqslant C \Big[\int_{|\mathbf{y}|+|\mathbf{t}|<|k|^{-1/2}} (d\mathbf{y})(d\mathbf{t}) (|\mathbf{y}| + |\mathbf{t}|) v_2(\mathbf{y}) u_1(\mathbf{t}) |k|^{1/2}$$

$$+ \int_{|\mathbf{y}|+|\mathbf{t}|>|k|^{-1/2}} (d\mathbf{y})(d\mathbf{t}) (|\mathbf{y}| + |\mathbf{t}|) v_2(\mathbf{y}) u_1(\mathbf{t}) \Big]$$

$$= o(1) \text{ as } k \to 0,$$

by (2.6). Equation (I.B4) follows.

APPENDIX F

We want to give here a simple derivation of the generalized optical theorem, i.e., of the unitarity of the S matrix, that does not depend on the nature of the spectrum of $H = -\Delta + V$.

Inserting (2.9) in (2.27) yields

$$-4\pi T(\hat{k}', \mathbf{k}) = (\psi_{\mathbf{k}'}^{(0)}, V\psi_{\mathbf{k}})$$

$$= (\psi_{\mathbf{k}'}, V\psi_{\mathbf{k}}) - (\psi_{\mathbf{k}'}, VG^{\dagger} V\psi_{\mathbf{k}}),$$

where $\mathbf{k}' = k\hat{k}'$, and hence

$$-4\pi[T(\hat{k}', \mathbf{k}) - T^*(\hat{k}, \mathbf{k}')] = (\psi_{\mathbf{k}'}, V(G - G^{\dagger}) V\psi_{\mathbf{k}}).$$

But

$$(G - G^{\dagger})(\mathbf{x}, \mathbf{y}) = -\frac{ik}{2(2\pi)^2} \int d\hat{k} \exp[i\mathbf{k} \cdot (\mathbf{x} - \mathbf{y})]$$

and hence

$$T(\hat{k}', \mathbf{k}) - T^*(\hat{k}, \mathbf{k}') = \frac{ik}{2\pi} \int d\hat{k}'' T(\hat{k}', \mathbf{k}'') T^*(\hat{k}, \mathbf{k}''). \tag{F1}$$

The special case of $\hat{k} = \hat{k}'$ is the optical theorem.

This derivation shows that, at a given value of k, the unitarity of the S matrix depends on the existence of a solution to (2.9) only. It does not depend on any kind of completeness of the spectrum of H.

[1] R. G. Newton, "Noncentral potentials: The generalized Levinson theorem and the structure of the spectrum," J. Math. Phys. 18, 1348 (1977).

[2] J. M. Jauch, Helv. Phys. Acta 30, 143 (1957).

[3] M. Gourdin and A. Martin, Nuovo Cimento 6, 757 (1957).

[4] A. Martin, Nuovo Cimento 7, 607 (1958).

[5] M. Ida, Progr. Theor. Phys. 21, 625 (1959).

[6] (a) M. Bertero, G. Valenti, and G. A. Viano, Nucl. Phys. A 113, 625 (1968); (b) V. S. Buslaev, Top. Math. Phys. 4, 43 (1969).

[7] L. Horwitz and J.-P. Marchand, Rocky Mount. J. Math. 1, 225 (1971).

[8] M. Wollenberg, "Levinson theorem. instabile Eigenwerte und Streuguerschnittmaxima," to be published in Math. Nachrichten.

[9] T. Dreyfus, thesis, Univ. of Geneva, 1976, unpublished preprint, and J. Phys. A: Math. Gen. 9, L 187 (1976).

[10] B. Bagchi, T. O. Krause, and B. Mulligan, Ohio State Univ. preprint. See also L. G. Arnold, B. Bagchi, and B. Mulligan, Ohio State Univ. preprint; and B. Mulligan *et al.*, Phys. Rev. C 13, 2131 (1976). The "spurious states" of these papers should not be confused with exceptional points of the second kind.

[11] These assumptions are somewhat different from those of M. Bertero, G. Valenti, and G. A. Viano, Nuovo Cimento 62, 27 (1969). This accounts for some slightly less conventional results in Sec. 2.

[12] We write † for the adjoint and * for the complex conjugate.

[13] If $\mathbf{k}$ is a vector in $\mathbb{R}^3$ then $k = |\mathbf{k}|$.

[14] Note that, in contrast to I, D is here defined as the *unmodified* Fredholm determinant. This is possible because tr$K$ exists; in I it did not.

[15] We write $\hat{k} = \mathbf{k}/|k|$.

[16] See, for example, B. Simon, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms* (Princeton U. P., Princeton, New Jersey, 1971), p. 9.

[17] This means that in a given direction $\hat{x}$, $R(\mathbf{x})$ may be nonnegligible compared to the $|\mathbf{x}|^{-1}$ term for large values of $|\mathbf{x}|$. But as $|\mathbf{x}| \to \infty$ the angular widths of such possible spikes of the remainder must tend to zero. Since all physical counters have finite sizes, this is sufficient for physical purposes.

[18] Eq. (F1). Thus, for every nonexceptional real k, the S matrix is unitary as an integral operator on the unit sphere. In view of our results in Secs. 4 and 5, S is, in fact, unitary for *all* real k. [T. Kato, *Perturbation Theory of Linear Operators* (Springer, New York, 1966), p. 540, Theorem 4.4, proves that the scattering operator for trace-class interactions is unitary. This implies only that the S matrix is unitary for almost all k.] Because the optical theorem plays an important role in the proof in Sec. 4 that there are no real exceptional points $k_0 \neq 0$ of the second kind, it is important to realize that its validity does not depend on the nature of the spectrum of H, even though its usual derivation makes it appear that it does. We give a suitable derivation in Appendix F.

[19] The proof of this proposition is essentially the same as in I, except that D is now the unmodified determinant, which simplifies it slightly.

[20] See Ref. 18 and Appendix F. What is needed is the validity of the optical theorem when the interaction is λV, near $\lambda = 1$. If $\det(1 - GV) = 0$ for $k = k_0$, then the analyticity of $\det(1 - \lambda GV)$ as a function of λ guarantees the existence of a real neighborhood of $\lambda = 1$ in which $\det(1 - \lambda GV) \neq 0$ and hence the optical theorem holds. One then lets $\lambda \to 1$ and shows that (4.2) must hold for almost all $\hat{n}$.

[21] Note that if (5.6) entailed $\varphi = 0$, then (5.3) would imply $\chi = 0$; if (5.7) meant that $\chi = 0$, then (5.2) would lead to $\varphi = 0$.

Geometrodynamics with tensor sources. IV

Karel Kuchař

Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 4 November 1976)

We develop the Hamiltonian hypersurface dynamics of the gravitational field derivatively coupled to general tensor sources. The closing of constraints follows from the independence of the hypersurface action on the path in the space of embeddings. The derivative coupling breaks the DeWitt supermetric in Riem ($\mathscr{m}$).

1. INTRODUCTION

In our previous papers,[1,2] we have proposed a geometrical language enabling us to describe field dynamics in a spacetime manifold as dynamics of hypertensor fields in hyperspace. Along these lines, we have developed a general theory of tensor fields propagating on a given Riemannian background.[3]

We now want to apply the same techniques to the Dirac-ADM dynamics of a free gravitational field, and to build up the hypersurface formalism for the gravitational field derivatively coupled to tensor sources. While nonderivatively coupled sources were soon embraced by the Dirac-ADM theory,[4] it was long doubted[5] whether the derivatively coupled sources can be treated in the same way. Our explicit construction shows they can, but at the same time draws attention to some unexpected features of the resulting formalism. In particular, the derivative gravitational coupling destroys DeWitt's "Riemannian structure" in the configuration space Riem($\mathscr{m}$) of gravitational variables. The significance of the breakdown is discussed in Sec. 5. Hamiltonian dynamics of tensor sources interacting with gravity is now independently studied by Fisher and Marsden; see Ref. 6 for a preliminary discussion of their method.

Our construction is carried on the same general level as the Belinfante—Rosenfeld analysis of the relationship between the symmetrical and canonical energy—momentum tensors. The Lagrangian potential Λ of the source fields is left arbitrary, so that we are able to treat source fields described by nonlinear equations. On this general level, we are concerned neither about the positive definiteness of the field energy, nor about the difficulty to keep the spin of a Pauli—Fierz field pure under the derivative gravitational coupling. We did not set out to resolve these difficulties, but we show that they can be spelled out and discussed within the framework of Hamiltonian hypersurface dynamics.

In hypersurface dynamics, the field equations naturally split into the initial value constraints and the Hamilton equations. The initial value constraints are preserved by Hamilton's equations in the course of dynamical evolution by virtue of the closing relations between the constraint functions. Our work thus separates the initial value problem from the evolution problem in the Cauchy problem for the gravitational field derivatively coupled to tensor sources. The Cauchy problem for nonderivatively coupled fields has been treated in full detail (see Bruhat[7] for an excellent discussion), but its extension to fields with nonderivative coupling looks far from straightforward. The main complication is again the breakdown of DeWitt's "Riemannian structure," implying that source-field variables explicitly enter into the evolution operator determining the development of the metric.

Because the present paper is the last one in a series,[1-3] we freely quote our previous results and use the old notation without explanation. The sections and equations of Refs. 1—3 are referred to by prefixing the Roman numerals I, II, III before the section and equation numbers; for notation, consult the Sec. I.2.

The material is organized into five sections, including this Introduction. In Sec. 2, we review the Dirac-ADM rearrangement of the Hilbert action in the hypertensor formalism. In Sec. 3, we couple the gravitational field to arbitrary tensor sources and cast the total action into the hypersurface form. The derivative gravitational coupling leads to a source contribution which must be added to the gravitational momentum to complete it into a true momentum canonically conjugate to the metric. The situation closely parallels the standard treatment of a charged particle moving in an electromagnetic field. We show in detail how the hypersurface action generates the appropriate projections of the Einstein law and of the field equations of the sources. In Sec. 4, we derive the Poisson brackets between the constraint functions from the independence of the hypersurface action on the choice of path in the space of embeddings. The derivation is simpler than for generalized Hamiltonian dynamics of the fields on a given Riemannian background, because the geometrodynamical action is in "an already parametrized form." Finally, in Sec. 5 we analyze the complications caused by the elimination of λ multipliers in the passage from the first-order to the second-order formalism. There we see how the derivative coupling precludes the introduction of a supermetric in Riem($\mathscr{m}$) independent of the source-field variables.

2. GRAVITATIONAL ACTION

Einstein's law in the vacuum can be obtained by varying the Hilbert[8] action

$$S^g[\underline{g}] = \int_{\mathcal{M}} \underline{\eta}\, {}^4R \tag{2.1}$$

with respect to the spacetime metric $g_{\alpha\beta}(X)$. We shall show now how to bring the Hilbert action principle into an equivalent Dirac-ADM[9] hypersurface form.

For this purpose, we pick up a curve $e(t)$ in the space of embeddings $\mathcal{E}$, connecting an initial embedding $\underset{1}{e}=e(\underset{1}{t})$ with a final embedding $\underset{2}{e}=e(\underset{2}{t})$. The deformation $\mathcal{E}$-vector tangent to this curve is

$$\mathbf{N}=\frac{de(t)}{dt}=\epsilon N^{x}\delta_{\perp x}+N^{ax}\delta_{ax}, \tag{2.2}$$

its components with respect to the normal hyperbasis $\{\delta_{\perp x},\delta_{ax}\}$ being the lapse function $N(x)$ and the shift vector $N^a(x)$ (see Sec. I.4). On each embedding e, the spacetime geometry $\underset{\sim}{g}$ induces the intrinsic geometry $\underset{\sim}{g}(x)[e]$ and the extrinsic curvature $\underset{\sim}{K}(x)[e]$. Along the embedding curve $e(t)$, the extrinsic curvature is connected with the rate of change $\dot{g}_{ab}(x)[e(t)]=\delta_{N}g_{ab}(x)[e(t)]$ of the intrinsic geometry $g_{ab}(x)[e]$ by the formula

$$\begin{aligned}\delta_{\mathrm{N}}g_{ab}(x)[e]&=\delta_{\mathrm{N}}g_{ab}(x)[e]+\delta_{\tilde{N}}g_{ab}(x)[e]\\&=-2K_{ab}(x)[e]N+2N_{(a|b)}.\end{aligned} \tag{2.3}$$

The integrand of the Hilbert action (2.1) can be expressed as a functional of the hypertensors $g_{ab}(x)[e]$, $K_{ab}(x)[e]$, $N(x)[e]$, $N^a(x)[e]$ along the embedding curve $e(t)$. Referring back to Eq. (II.8.3) for the projected form of the Riemann curvature scalar ${}^4R$, and using the equation

$$\delta_{\mathrm{N}}g^{1/2}=-\underset{\vee}{K}N+\underset{\vee}{N}{}^{a}{}_{,a} \tag{2.4}$$

for the rate of change of the space volume density $g^{1/2}$, we get

$$\begin{aligned}|{}^4g|^{1/2}\,{}^4R=Ng^{1/2}\,{}^4R&=Ng^{1/2}[-\epsilon(K_{ab}K^{ab}-K^2)+R]\\&\quad-2\epsilon(g^{1/2}g^{ab}N_{,b})_{,a}-2\epsilon(\underset{\vee}{K}N^a)_{,a}+2\epsilon\delta_{\mathrm{N}}\underset{\vee}{K}.\end{aligned} \tag{2.5}$$

The action S^g contained in a space region m between the initial and final embeddings $\underset{1}{e}$ and $\underset{2}{e}$ is obtained by integrating the expression (2.5),[1]

$$\begin{aligned}&S^g[\underset{\sim}{g}[e(t)],N[e(t)],\tilde{N}[e(t)]]\\&=\int_{\underset{1}{t}}^{\underset{2}{t}}dt\{\int_{x\in m}Ng^{1/2}[-\epsilon(K_{ab}K^{ab}-K^2)+R]\\&\quad+\int_{\partial m}d\sigma_a(-2\epsilon g^{ab}N_{,b}-2\epsilon KN^a)\}\\&\quad+2\epsilon\int_{x\in m}(\underset{\vee}{K}(x)[\underset{2}{e}]-\underset{\vee}{K}(x)[\underset{1}{e}]).\end{aligned} \tag{2.6}$$

The term in braces is the hypersurface gravitational Lagrangian $\delta_{\mathrm{N}}S^g$. The last three terms in Eq. (2.5) yielded the boundary terms in the action (2.6); one taken over the spatial boundary ∂m with the surface element $d\sigma_a$ [see Eq. (III.5.15)], and another one over the initial and final embeddings.

The extrinsic curvature in Eq. (2.6) is thought to be expressed through N, N^a, and $\delta_{\mathrm{N}}g_{ab}$ by means of Eq. (2.3), so that the action becomes a functional of the variables N, N^a, g_{ab} along the embedding curve. There is a one-to-one correspondence between the spacetime metric $g_{\alpha\beta}(X)$ and the hypertensors $N(x)[e(t)]$, $N^a(x)[e(t)]$, $g_{ab}(x)[e(t)]$,

$$\begin{aligned}N(x)[e(t)]&=\epsilon n_\alpha(x,t)\dot{e}^\alpha(x,t),\\N^a(x)[e(t)]&=e^a_\alpha(x,t)\dot{e}^\alpha(x,t),\\g_{ab}(x)[e(t)]&=g_{\alpha\beta}(e(x,t))e^\alpha_a(x,t)e^\beta_b(x,t).\end{aligned} \tag{2.7}$$

Because the Hilbert action (2.1) evaluated for a spacetime metric $g_{\alpha\beta}(X)$ is numerically equal to the hypersurface action (2.6) evaluated for the projections (2.7), we get correct field equations by varying the action (2.6) with respect to the hypertensor variables (2.7).

While N and N^a enter the hypersurface action merely as Lagrange multipliers, g_{ab} occurs there together with its rate of change $\delta_{\mathrm{N}}g_{ab}$. Introducing the momentum $\pi^{ab}(x)[e]$ canonically conjugate to the metric $g_{ab}(x)[e]$,

$$\pi^{ab}(x)=\frac{\delta(\delta_{\mathrm{N}}S^g)}{\delta(\delta_{\mathrm{N}}g_{ab}(x))}=\epsilon(\underset{\vee}{K}{}^{ab}(x)-\underset{\vee}{K}(x)g^{ab}(x)) \tag{2.8}$$

[cf. Eq. (II.8.7)], we can cast the action (2.6) into canonical form. The geometrodynamical Hamiltonian is formed by applying the Legendre dual transformation to the Lagrangian form (2.6) of the action, taking into account Eqs. (2.3) and (2.8),

$$\begin{aligned}&\pi^{abx}\delta_{\mathrm{N}}g_{abx}-\delta_{\mathrm{N}}S^g\\&=\int_m\{\pi^{ab}\delta_{\mathrm{N}}g_{ab}-Ng^{1/2}[-\epsilon(K_{ab}K^{ab}-K^2)+R]\}\\&\quad+\int_{\partial m}d\sigma_a(2\epsilon g^{ab}N_{,b}+2\epsilon KN^a)\\&=\int_m(NH^g+N^aH^g{}_a)\\&\quad+\int_{\partial m}d\sigma_a(2\epsilon g^{ab}N_{,b}+2g^{-1/2}(\pi^a_b-\tfrac{1}{2}\pi\delta^a_b)N^b).\end{aligned} \tag{2.9}$$

The gravitational super-Hamiltonian H^g and supermomentum $H^g{}_a$,

$$\begin{aligned}H^g&=-\epsilon G_{ab\,cd}\pi^{ab}\pi^{cd}-g^{1/2}R,\\G_{ab\,cd}&\equiv\tfrac{1}{2}g^{-1/2}(g_{ac}g_{bd}+g_{ad}g_{bc}-g_{ab}g_{cd}),\end{aligned} \tag{2.10}$$

$$H^g{}_a=-2\pi^b_{a|b}, \tag{2.11}$$

are identical with the expressions (II.8.12) and (II.8.13) we have found earlier by the direct rearrangement of the Einstein tensor. We thus arrive at the well-known Dirac-ADM form of the gravitational action,

$$\begin{aligned}&S^g[g_{ab}(x)[e(t)],\pi^{ab}(x)[e(t)];N[e(t)],N^a[e(t)]]\\&=\int_{\underset{1}{t}}^{\underset{2}{t}}dt\{\int_m(\pi^{ab}\delta_{\mathrm{N}}g_{ab}-NH^g-N^aH^g{}_a)\\&\quad+\int_{\partial m}d\sigma_a(-2\epsilon g^{ab}N_{,b}-2g^{-1/2}(\pi^a_b-\pi\delta^a_b)N^b)\}\\&\quad+\int_m(\pi(x)[\underset{1}{e}]-\pi(x)[\underset{2}{e}]).\end{aligned} \tag{2.12}$$

Similarly as in Sec. III.5, where we have studied a tensor field propagating on a given Riemannian background, we now disregard boundary terms; the discussion may be found, e.g., in the references in[10] III.

Varying the action (2.12) with respect to the canonical variables $g_{ab}(x)[e]$, $\pi^{ab}(x)[e]$, we obtain the Hamilton equations. We write them as variational equations in hyperspace, explicitly splitting the normal dynamics from the tangential dynamics, following the pattern of Eqs. (III.6.5) and (III.6.6),

$$\delta_N g_{ab}(x)[e] = [g_{ab}(x), H^g{}_{x'}]N^{x'}, \tag{2.13}$$

$$\delta_N \pi^{ab}(x)[e] = [\pi^{ab}(x), H^g{}_{x'}]N^{x'},$$

$$\begin{aligned} \delta_{\tilde{N}} g_{ab}(x)[e] &= [g_{ab}(x), H^g{}_{cx'}]N^{cx'}, \\ \delta_{\tilde{N}} \pi^{ab}(x)[e] &= [\pi^{ab}(x), H^g{}_{cx'}]N^{cx'}. \end{aligned} \tag{2.14}$$

As discussed in Sec. II.5, the tangential equations (2.14) merely reproduce the Lie-derivative change of the dynamical variables. On the other hand, the normal equations (2.13) carry the true information about the dynamical evolution of geometry.

Varying the action with respect to the Lagrange multipliers N and N^a, we obtain the initial value constraints

$$H^g(x) = 0 = H^g{}_a(x). \tag{2.15}$$

Their presence indicates that the canonical variables g_{ab}, π^{ab} carry implicit information about the embedding variables e^α, p_α, the whole scheme resembling the generalized Hamiltonian dynamics of spacetime hypertensors discussed in Sec. III.7. We have analyzed this particular aspect of the geometrodynamical formalism in an earlier paper.[10]

3. SWITCHING ON THE SOURCES

In Ref. 3, we have studied the dynamics of tensor fields propagating on a given geometrical background. We now couple these fields to geometry by adding the field action S^Φ to the Hilbert gravitational action S^g,

$$S[\underline{g}, \underline{\phi}, \overline{\lambda}] = S^g[\underline{g}] + S^\Phi[\underline{g}; \underline{\phi}, \overline{\lambda}]. \tag{3.1}$$

Our units are chosen so that the coupling constant $2\kappa = 16\pi G c^{-3} = 1$; the Einstein gravitational law then becomes

$$\overline{G} = \tfrac{1}{2}\overline{T}. \tag{3.2}$$

The factor $\frac{1}{2}$ on the right-hand side of Eq. (3.2) is to be remembered when interpreting the projected terms. The field action S^Φ is taken in the first-order form which, similarly as for the fields propagating on a given background, substantially simplifies the hypersurface projections. The price to pay are the λ multipliers which enter the Hamiltonian formalism; their elimination is discussed in Sec. 5.

To follow the joint dynamics of geometry and tensor sources, we must cast the total action (3.1) into the hypersurface form. The rearrangement of the source action S^Φ brings no surprises. It proceeds exactly as if the background was fixed, which allows us to take directly the results of Secs. III.4—III.6. For the representative case of a covector field, the hypersurface form of the source action is [see Eq. (III.5.13)]

$$\begin{aligned} &S^\Phi[\phi_\perp, \pi^\perp, \phi_a, \pi^a; \lambda^{\perp a}, \lambda^{ab}; g_{ab}, K_{ab}] \\ &= \int_{t_1}^{t_2} dt\, \delta_{\mathrm{N}} S^\Phi, \ \delta_{\mathrm{N}} S^\Phi = \int_m (\pi^\perp \delta_{\mathrm{N}} \phi_\perp + \pi^a \delta_{\mathrm{N}} \phi_a - N \overset{\circ}{H}{}^\Phi - N^a \overset{\circ}{H}{}^\Phi{}_a) \\ &\quad + \int_{\partial m} d\sigma_a (N \overset{\circ}{P}{}^{\Phi\, a} + N^b \overset{\circ}{P}{}^{\Phi\, a}{}_b). \end{aligned} \tag{3.3}$$

The field super-Hamiltonian $\overset{\circ}{H}{}^\Phi$ and supermomentum $\overset{\circ}{H}{}^\Phi{}_a$ are given by expressions (III.5.16)—(III.5.21). At the present moment, we want to stress that the field supermomentum does not depend on the geometrical variables g_{ab}, K_{ab}, while the field super-Hamiltonian depends on them in a very special way; it is linear in the extrinsic curvature for fields with derivative gravitational coupling,

$$\overset{\circ}{H}{}^\Phi = H^\Phi + 2P^{ab}K_{ab}, \tag{3.4}$$

and the term with extrinsic curvature drops out for fields with nonderivative gravitational coupling (see Sec. III.11).

The importance of these observations becomes apparent when we pass from the source action to the total action (3.1) and try to cast it into the Hamiltonian form with respect to the g_{ab}. The Hilbert action S^g is easily rearranged into the hypersurface from (2.6). However, when we try to define the momentum $p^{ab}(x)[e]$ canonically conjugate to the metric $g_{ab}(x)[e]$,

$$\begin{aligned} p^{ab}(x) &\equiv \frac{\delta(\delta_{\mathrm{N}} S)}{\delta(\delta_{\mathrm{N}} g_{ab}(x))} = -\frac{1}{2N(x)} \frac{\delta(\delta_{\mathrm{N}} S)}{\delta K_{ab}(x)} \\ &= -\frac{1}{2N(x)} \left(\frac{\delta(\delta_{\mathrm{N}} S^g)}{\delta K_{ab}(x)} + \frac{\delta(\delta_{\mathrm{N}} S^\Phi)}{\delta K_{ab}(x)} \right), \end{aligned} \tag{3.5}$$

it becomes painfully obvious that not only the gravitational action S^g, but also the source field action S^Φ contributes to $p^{ab}(x)[e]$. This contribution amounts to $P^{ab}(x)$ and comes directly from the extrinsic curvature term $2P^{ab}K_{ab}$ in the field super-Hamiltonian (3.4).

To describe the situation, let us use the old symbol π^{ab} for the momentum (2.8) stemming from the Hilbert action S^g. Because this action characterizes the gravitational field without interaction with sources, we will call π^{ab} *the gravitational momentum*. On the other hand, the momentum p^{ab} canonically conjugate to the metric g_{ab} determines the dynamics of geometry coupled to sources within the Hamiltonian scheme and deserves thus the name of *geometrodynamical momentum*. We see from Eq. (3.5) that the geometrodynamical momentum p^{ab} differs from the gravitational momentum π^{ab} by the term P^{ab} constructed from the source-field variables and attributable to the derivative gravitational coupling,

$$\pi^{ab} = p^{ab} - P^{ab}. \tag{3.6}$$

Unfortunately, in one respect the proposed terminology is misleading. It is the gravitational momentum π^{ab} which is purely geometrical, being constructed from the extrinsic curvature in the standard way (2.8), and it is the geometrodynamical momentum p^{ab} which is extrageometrical, containing a contribution P^{ab} from the sources.

The reason why the source-field momentum is the same for the field interacting with geometry as for the field propagating on a given geometrical background is obvious: The gravitational action S^g does not contain any source-field variables. In contrast with this, the source-field action S^Φ contains the geometrical variables. Therefore, when projected into the hypersurface form, it exhibits the term $-2N\,P^{ab}\,K_{ab}$ linear in the geometrodynamical velocity $\delta_{\mathrm{N}} g_{ab}$. This term then leads to the difference between the geometrodynamical and the gravitational momentum.

The whole situation closely resembles the Hamiltonian dynamics of a charged particle moving in an electromagnetic field. Here, the momentum $\tilde{p}$ canonically conjugate to the position q of the particle also differs from the mechanical momentum $\underset{\sim}{\pi} = m\dot{q}$ by the term $e\underset{\sim}{A}$ containing the vector potential $\underset{\sim}{A}$ of the electromagnetic field. This term can be traced back to the interaction term $e\,A_a\,\dot{q}^a$ (linear in the particle velocity $\dot{q}^a$) in the particle Lagrangian, similarly as our source term P^{ab} can be traced back to the interaction term $P^{ab}\delta_{\rm N}g_{ab}$ (linear in the geometrodynamical velocity $\delta_{\rm N}g_{ab}$) in the field Lagrangian.

With the field action S^Φ already in the canonical form (3.3) in the source-field variables, and the geometrodynamical momentum (3.6) at our disposal, it is easy to bring the total action (3.1) into the canonical form with respect to the remaining dynamical variable g_{ab}. The Hamiltonian is again obtained from the hypersurface Langrangian $\delta_{\rm N}S = \delta_{\rm N}S^g + \delta_{\rm N}S^\Phi$ by the Legendre dual transformation,

$$\begin{aligned}
p^{abx}\delta_{\rm N}g_{abx} &+ \pi^{\perp x}\delta_{\rm N}\phi_{\perp x} + \pi^{ax}\delta_{\rm N}\phi_{ax} - \delta_{\rm N}S \\
&= \int_m \{p^{ab}\delta_{\rm N}g_{ab} - Ng^{1/2}(-\epsilon(K_{ab}K^{ab} - K^2) + R) \\
&\quad + 2NP^{ab}K_{ab} + NH^\Phi + N^a\mathring{H}^\Phi{}_a\} \\
&\quad + \int_{\partial m} d\sigma_a\{(2\epsilon g^{ab}N_{,b} + 2\epsilon KN^a) - (N\mathring{P}^{\Phi\,a} + N^b\mathring{P}^{\Phi\,a}{}_b)\} \\
&= \int_m (NH + N^aH_a) \\
&\quad + \int_{\partial m} d\sigma_a\{(2\epsilon g^{ab}N_{,b} + 2g^{-1/2}(\pi^a_b - \tfrac{1}{2}\pi\delta^a_b)N^b) \\
&\quad - (N\mathring{P}^{\Phi\,a} + N^b(\mathring{P}^{\Phi\,a}{}_b - g^{-1/2}P^a_b))\},
\end{aligned} \tag{3.7}$$

in course of which the extrinsic curvature K_{ab} gets eliminated by means of the inverted Eqs. (2.8), (3.6),

$$K_{ab} = \epsilon G_{ab\,cd}(p^{cd} - P^{cd}). \tag{3.8}$$

Collecting the terms, we get the total super-Hamiltonian and supermomentum

$$H = H^g + H^\Phi, \tag{3.9}$$

$$H_a = -2p^b_{a\,|b} + \mathring{H}^\Phi{}_a. \tag{3.10}$$

The super-Hamiltonian H is the sum of the gravitational super-Hamiltonian (2.10) and the field super-Hamiltonian H^Φ. Note that H^Φ is not identical with the super-Hamiltonian $\mathring{H}^\Phi$ of the field propagating on a given background, because the term $2P^{ab}K_{ab}$ is missing in $\mathring{H}^\Phi$. Remember also that it was H^Φ and not $\mathring{H}^\Phi$ which we have identified in Eq. (III.8.14) with the $\perp\perp$ projection of the symmetrical energy—momentum tensor,

$$\underset{\sim}{T}^{\perp\perp} = -\epsilon H^\Phi. \tag{3.11}$$

Similarly, the gravitational super-Hamiltonian H^g was identified in Eq. (III.8.11) with the $\perp\perp$ projection of the Einstein tensor,

$$\underset{\sim}{G}^{\perp\perp} = \tfrac{1}{2}\epsilon H^g, \tag{3.12}$$

the factor $\frac{1}{2}$ reflecting our choice of units [cf. Eq. (3.2)]. However, π^{ab} in the gravitational super-Hamiltonian must not be confused with the momentum p^{ab} canonically conjugate to g_{ab}. It is to be considered merely as an abbreviation (3.6). Through the substitution (3.6), the derivative gravitational coupling finds its way into the total super-Hamiltonian, in spite of its seemingly additive form (3.9). This is again completely analogous to the situation of a charged particle moving in an electromagnetic field.

In Sec. II.5, we have learned that $\mathring{H}^\Phi{}_a$ generates the appropriate Lie-derivative change of the source-field variables $\phi_\perp$, $\pi^\perp$, ϕ_a, π^a under the tangential deformation of an embedding, and $-2p^b_{a\,|b}$ generates the appropriate Lie-derivative change of the canonical variables g_{ab}, p^{ab}. This explains the additive form of the total supermomentum (3.10). This form is geared to the purposes of the canonical formalism. However, in complete analogy with a different split (3.9) of the super-Hamiltonian, we can write the supermomentum H_a also as the sum

$$H_a = H^g{}_a + H^\Phi{}_a \tag{3.13}$$

of the gravitational supermomentum (2.11), which is directly related by Eq. (II.8.11) to the $\perp\parallel$ projection of the Einstein tensor,

$$\underset{\sim}{G}_{\perp a} = \tfrac{1}{2}H^g{}_a, \tag{3.14}$$

and that part, $H^\Phi{}_a$, of the field supermomentum,

$$H^\Phi{}_a = \mathring{H}^\Phi{}_a - 2P^b_{a\,|b}, \tag{3.15}$$

which is directly related by Eq. (III.8.15) to the $\perp\parallel$ projection of the symmetrical energy—momentum tensor,

$$\underset{\sim}{T}_{\perp a} = -H^\Phi{}_a. \tag{3.16}$$

This split of the supermomentum is useful when comparing the supermomentum constraint with the $\perp\parallel$ projection of the Einstein law (3.2).

Having now the Hamiltonian (3.7), we can write down the final form of the hypersurface action,

$$\begin{aligned}
S[g_{ab}, p^{ab}; \phi_\perp, \pi^\perp, \phi_a, \pi^a; \lambda^{\perp b}, \lambda^{ab}; N, N^a] \\
&= \int_{t_1}^{t_2} dt\{\int_m (p^{ab}\delta_{\rm N}g_{ab} + \pi^\perp\delta_{\rm N}\phi_\perp \\
&\quad + \pi^a\delta_{\rm N}\phi_a - NH - N^aH_a) \\
&\quad + \int_{\partial m} d\sigma_a(-2\epsilon g^{ab}N_{,b} - 2g^{-1/2}(\pi^a_b - \tfrac{1}{2}\pi\delta^a_b)N^b \\
&\quad + N\mathring{P}^{\Phi\,a} + N^b(\mathring{P}^{\Phi\,a}{}_b - 2g^{-1/2}P^a_b))\} \\
&\quad + \int_m (\pi(x)[\underset{1}{e}] - \pi(x)[\underset{2}{e}]).
\end{aligned} \tag{3.17}$$

For completeness, we have included all the boundary terms, though we do not want to discuss them here in detail.

Let us see now how the hypersurface action (3.17) generates the appropriate projections of the field equations and of the Einstein law of gravitation. All variables which we have written down in the functional dependence of the action (3.17) are to be varied independently. We will discuss first the variation with respect to the source-field variables $\phi_\perp$, $\pi^\perp$, ϕ_a, π^a, $\lambda^{\perp a}$, λ^{ab}. Because

$$\delta_{(\phi,\pi,\lambda)}H^g = 2\epsilon G_{ab\,cd}\pi^{cd}\delta_{(\phi,\pi,\lambda)}P^{ab}$$

$$= 2K_{ab}\delta_{(\phi,\pi,\lambda)}P^{ab} \qquad (3.18)$$

in view of Eqs. (3.6) and (3.8), we get the relation

$$\delta_{(\phi,\pi,\lambda)}H\big|_{\underset{\sim}{g},\tilde{p}\text{ fixed}} = \delta_{(\phi,\pi,\lambda)}\overset{\circ}{H}\big|_{\underset{\sim}{g},\underset{\sim}{K}\text{ fixed}} \qquad (3.19)$$

which implies that equations for the tensor sources of the gravitational field are the same as if the source fields were propagating on a given gravitational background. This applies to the Hamilton equations as well as to the λ equations. The Hamilton equations can again be written as variational equations in hyperspace, Eqs. (III.6.5)—(III.6.6), with H, H_a playing the role of $\overset{\circ}{H}$, $\overset{\circ}{H}_a$. The λ equations have the explicit form (III.10.2); replacing the extrinsic curvature by the geometrodynamical momentum p^{ab}, they read

$$g^{-1/2}\,\frac{\partial H}{\partial\lambda^{\perp a}} = \frac{\partial\Lambda}{\partial\lambda^{\perp a}} - \epsilon G_{ab\,cd}\phi^b(p^{cd} - P^{cd}) - \epsilon\phi_{\perp|a} = 0,$$

$$g^{-1/2}\,\frac{\partial H}{\partial\lambda^{ab}} = \frac{\partial\Lambda}{\partial\lambda^{ab}} + \epsilon G_{ab\,cd}\phi_\perp(p^{cd} - P^{cd}) - \phi_{a|b} = 0. \qquad (3.20)$$

They will serve us to eliminate the λ multipliers in Sec. 5.

Next, varying the lapse function and the shift vector, we get the constraints

$$H = 0 = H_a. \qquad (3.21)$$

Due to the additive nature (3.9) and (3.13) of the constraint functions, H, H_a, and the identification (3.11), (3.12), (3.14), and (3.16) of the projections of the Einstein tensor and the symmetrical energy—momentum tensor, Eqs. (3.21) are easily seen to yield the ⊥⊥ and ⊥ ∥ projections of the Einstein law (3.2).

Finally, vary the geometrodynamical variables p^{ab}, g_{ab}. The variation of p^{ab} leads back to its connection (3.8) with the extrinsic curvature. It remains to be shown that the variation of g_{ab} reproduces the ∥ ∥ projection of the Einstein law. Of course, we know that this variation gives the Hamilton equation

$$\delta_N p^{ab}(x) = [p^{ab}(x), H_{x'}]N^{x'}$$

$$= [p^{ab}(x), H^g{}_{x'}]N^{x'} + [p^{ab}(x), H^\phi{}_{x'}]N^{x'}. \qquad (3.22)$$

In the Poisson bracket with $H^g{}_{x'}$, we can keep $\tilde{\pi}$ fixed instead of keeping $\tilde{p}$ fixed, picking up an additional term,

$$[p^{ab}(x), H^g(x')]$$

$$= -\frac{\delta H^g(x')}{\delta g_{ab}(x)}\bigg|_{\tilde{p}\text{ fixed}}$$

$$= -\frac{\delta H^g(x')}{\delta g_{ab}(x)}\bigg|_{\tilde{\pi}\text{ fixed}} - 2\epsilon G_{cd\,ef}(x')\pi^{ef}(x')\,\frac{\delta P^{cd}(x')}{\delta g_{ab}(x)}$$

$$= [\pi^{ab}(x), H^g(x')[\underset{\sim}{g}, \tilde{\pi}]] - 2K_{cd}\,\frac{\partial P^{cd}}{\partial g_{ab}}\,\delta(x', x). \qquad (3.23)$$

We then use the geometrokinematical identity (II.8.19),

$$\delta_N\pi^{ab}(x) = [\pi^{ab}(x), H^g{}_{x'}[\underset{\sim}{g}, \tilde{\pi}]]N^{x'} + \underset{\smile}{G}^{ab}N, \qquad (3.24)$$

to introduce the G^{ab} projection of the Einstein tensor into our game. Putting Eqs. (3.6), (3.22), (3.23), and (3.24) together, and isolating $\underset{\smile}{G}^{ab}$ on one side, we get

$$\underset{\smile}{G}^{ab}N = \tfrac{1}{2}\left(-2\,\frac{\delta H^\phi{}_{x'}}{\delta g_{ab}}\,N^{x'} - 2\delta_N P^{ab} - 4K_{cd}\,\frac{\partial P^{cd}}{\partial g_{ab}}\,N\right). \qquad (3.25)$$

This is just the ab projection of the Einstein law (3.2), the terms in the bracket giving the $N\underset{\smile}{T}^{ab}$ component of the symmetrical energy—momentum tensor according to Eq. (III.8.17). We have thus directly verified that the hypersurface action (3.17) generates all projections of the Einstein law of gravitation.

This concludes our proof that the action for the gravitational field derivatively coupled to tensor sources can be given a hypersurface form. When sources are described by the first-order action, the resulting formalism closely resembles the situation of a charged particle moving in an electromagnetic field. The elimination of the λ multipliers, however, introduces a number of complications which we will discuss in Sec. 5.

4. CLOSING OF CONSTRAINTS

In Sec. III.12, we have studied the Poisson brackets between the constraint functions $\overset{\circ}{H}$ and $\overset{\circ}{H}_a$ of the generalized Hamiltonian dynamics of tensor fields propagating on a given background. We concluded that these Poisson brackets are expressible as certain universal combinations of the original constraint functions $\overset{\circ}{H}$, $\overset{\circ}{H}_a$ and of the λ equations generated by $\overset{\circ}{H}$. This ensured the preservation of the initial value constraints from one embedding to another.

In geometrodynamics with tensor sources, we again encounter initial value constraints—Eqs. (3.21). The new constraint functions, H and H_a, are quite different from the old constraint functions, $\overset{\circ}{H}$ and $\overset{\circ}{H}_a$. In particular, they depend on the geometrodynamical variables g_{ab}, p^{ab} instead of on the hypersurface variables e^α, p_α. However, the Poisson brackets between H and H_a are the same combinations of H, H_a and of the new λ equations as the Poisson brackets between $\overset{\circ}{H}$ and $\overset{\circ}{H}_a$ were of $\overset{\circ}{H}$, $\overset{\circ}{H}_a$ and of the old λ equations.

There is a good reason why the Poisson brackets have the same structure whether the sources are or are not coupled to geometry; in each case, the structure follows from the fact that the hypersurface action does not depend on the path $e(t,x)$ in $\mathcal{E}$ when the spacetime fields are kept fixed. The only difference is that $g(x)$ is included among the dynamical fields in geometrodynamics. Incidentally, this simplifies the evaluation of the Poisson brackets, since the lapse function and the shift vector are considered as independent variables from the beginning, while in the field theory on a background they are added to other variables later, during the "parametrization process."

We will now show how the structure of the Poisson brackets between the geometrodynamical constraint functions H, H_a follows from the invariance of the total action S under the change of path. For definiteness, we will carry out the proof for the typical case of single scalar field interacting with gravity. (The scalar field has nonderivative gravitational coupling, but the difference between derivatively and nonderivatively coupled fields has no bearing on our argument.) The generalization to tensor fields is straightforward and proceeds as in Sec. III.12. Our derivation also gives the Poisson brackets between the gravitational super-Hamiltonian H^g and supermomentum $H^g{}_a$ in pure geometrodynamics; we only have to put the scalar field variables equal to zero in the final result. These Poisson brackets, (4.18)–(4.20), do not contain any λ multipliers, pure geometrodynamics being treated by us from the beginning in the second-order form. Similarly, we can bring geometrodynamics with arbitrary tensor sources into second-order form, by eliminating λ multipliers from the constraint functions. The modified constraint functions then close in the same universal way as the gravitational constraint functions H^g, $H^g{}_a$.

Our derivation is based on the equivalence of the spacetime action with the hypersurface action. We thus start from the spacetime action of the scalar field interacting with gravity,

$$S[\phi,\lambda^\alpha,g_{\alpha\beta}]=\int_{\mathcal{M}} d^4X\,|{}^4g|^{1/2}({}^4R+\lambda^\alpha\phi_{,\alpha}-\Lambda(\phi,\lambda^\alpha,g_{\alpha\beta})). \tag{4.1}$$

To cast it into the hypersurface form (see Secs. 2 and III.11 A), we choose a path $e^\alpha(t,x)$ in $\mathcal{E}$, represent the spacetime fields by their hypersurface projections, introduce the field momenta, and perform the Lengendre dual transformation

$$\begin{aligned}S[g_{ab},p^{ab};\phi,\pi;\lambda^a,N,N^a]\\ =\int_{\underset{1}{t}}^{\underset{2}{t}} dt\int_m (p^{ab}\delta_{\mathbf{N}}g_{ab}+\pi\delta_{\mathbf{N}}\phi-NH-N^aH_a)\\ +\text{boundary terms.}\end{aligned} \tag{4.2}$$

Conversely, to return from Eq. (4.2) to Eq. (4.1), we define the field momenta $p^{ab}(x)[e]$, $\pi(x)[e]$ in terms of the field velocities $\delta_{\mathbf{N}}g_{ab}(x)[e]$, $\delta_{\mathbf{N}}\phi(x)[e]$ from the first set of Hamilton's equations

$$\begin{aligned}\delta_{\mathbf{N}}g_{ab}(x)[e]&=\frac{\delta H_{x'}}{\delta p^{ab}(x)}N^{x'}+\frac{\delta H_{cx'}}{\delta p^{ab}(x)}N^{cx'},\\ \delta_{\mathbf{N}}\phi(x)[e]&=\frac{\delta H_{x'}}{\delta\pi(x)}N^{x'}+\frac{\delta H_{cx'}}{\delta\pi(x)}N^{cx'}.\end{aligned} \tag{4.3}$$

If we eliminate the field momenta from the hypersurface action (4.2), it becomes numerically equal to the spacetime action (4.1). Let us emphasize that Eqs. (4.3) do not play the role of field equations in our argument, but express the Legendre dual transformation which connects the hypersurface action (4.2) with the spacetime action (4.1). This connection holds for arbitrary fields $\phi(X)$, $g_{\alpha\beta}(X)$, not only for the extremal fields satisfying the field equations.

We know that the spacetime action remains the same if we vary the path $e^\alpha(t,x)$ keeping the spacetime fields $g_{\alpha\beta}(X)$, $\phi(X)$, $\lambda^\alpha(X)$ fixed. Such a variation $\delta e^\alpha(t,x) \equiv M^\alpha(t,x)$, induces the variation $\delta_{\mathbf{M}}$ of the hypersurface projections $g_{ab}(x)[e]$, $N(x)[e]$, $N^a(x)[e]$, $\phi(x)[e]$, $\lambda^a(x)[e]$, and through Eqs. (4.3), of the field momenta $p^{ab}(x)[e]$, $\pi(x)[e]$. The hypersurface action (4.2), being numerically equal to the spacetime action (4.1), remains unchanged by this variation,

$$\begin{aligned}\delta_{\mathbf{M}}S=\int_{\underset{1}{t}}^{\underset{2}{t}} dt\int_m (\delta_{\mathbf{M}}p^{ab}\delta_{\mathbf{N}}g_{ab}-\delta_{\mathbf{N}}p^{ab}\delta_{\mathbf{M}}g_{ab}\\ +\delta_{\mathbf{M}}\pi\delta_{\mathbf{N}}\phi-\delta_{\mathbf{N}}\pi\delta_{\mathbf{M}}\phi-H\delta_{\mathbf{M}}N-H_a\delta_{\mathbf{M}}N^a\\ -N\delta_{\mathbf{M}}H-N^a\delta_{\mathbf{M}}H_a)\equiv 0.\end{aligned} \tag{4.4}$$

The identity (4.3) expresses the path independence of the hypersurface action.

The symplectic form appeared in Eq. (4.4) through integration by parts. Because $\mathbf{N}$ is the deformation vector along the original path $e(t,x)$, its action on an arbitrary functional $F[e]$ of e is given by the time derivative,

$$\delta_{\mathbf{N}}F[e]=\dot{F}[e(t)]. \tag{4.5}$$

This enables us to integrate certain terms by parts with respect to time, like in the expression

$$\begin{aligned}&\int_{\underset{1}{t}}^{\underset{2}{t}} dt\int_m p^{ab}\delta_{\mathbf{M}}\delta_{\mathbf{N}}g_{ab}\\ &\quad=\int_m [p^{ab}(x)[e]\delta_{\mathbf{M}}g_{ab}(x)[e]]_{\underset{1}{e}}^{\underset{2}{e}}-\int_{\underset{1}{t}}^{\underset{2}{t}} dt\int_m \delta_{\mathbf{N}}p^{ab}\delta_{\mathbf{M}}g_{ab}.\end{aligned}$$

At this point we assume that the path e is not varied at the two faces, $\underset{1}{e}$ and $\underset{2}{e}$, of the spacetime sandwich, so that the boundary terms drop out.

The changes $\delta_{\mathbf{M}}N$, $\delta_{\mathbf{M}}N^a$ of the lapse function and the shift vector under the change $\mathbf{M}$ of the path e are easily specified. To start with, the change of the deformation $\mathcal{E}$-vector $\mathbf{N}$ is given by the formula (see Fig. 1)

$$\delta_{\mathbf{M}}\mathbf{N}=\mathbf{N}\mathbf{M}=\mathbf{M}\mathbf{N}+[\mathbf{N},\mathbf{M}]. \tag{4.6}$$

Taking its components with respect to the normal hyperbasis,

$$\delta_{\mathbf{M}}N(x)=\mathbf{M}N(x)+[\mathbf{N},\mathbf{M}]^{\perp x},$$

$$\delta_{\mathbf{M}}N^a(x)=\mathbf{M}N^a(x)+[\mathbf{N},\mathbf{M}]^{ax},$$

and using Eqs. (I.6.19)–(I.6.20) for the $\mathcal{E}$-vector Lie brackets, we get

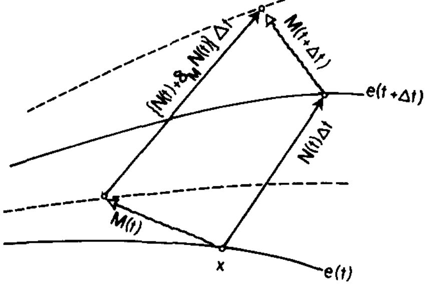


FIG. 1. Change of the deformation $\mathcal{E}$-vector $\mathbf{N}$ under the displacement $\mathbf{M}(t)$ of the path $e(t)$. From the vector quadrangle in the picture, $\delta_{\mathbf{M}}\mathbf{N}(t)\doteq\lim_{\Delta t\to 0}(\Delta t)^{-1}(\mathbf{M}(t+\Delta t)-\mathbf{M}(t))=\mathbf{N}\mathbf{M}$.

$$\delta_{\mathbf{M}}N(x)=\mathbf{N}M(x)+N_{,a}M^a-M_{,a}N^a,$$

$$\delta_{\mathbf{M}}N^a(x)=\mathbf{N}M^a(x)+\epsilon(NM^{,a}-N^{,a}M)-[\tilde{N},\tilde{M}]^a. \tag{4.7}$$

Recalling Eq. (4.5), we can integrate the terms $-\dot{M}H-\dot{M}^aH_a$ by parts with respect to time,

$$\begin{aligned}-\int_{t_1}^{t_2}dt\int_m(H(x)\mathbf{N}M(x)+H_a(x)\mathbf{N}M^a(x))\\ =\int_{t_1}^{t_2}dt\int_m(M(x)\delta_{\mathbf{N}}H(x)+M^a(x)\delta_{\mathbf{N}}H_a(x)),\end{aligned} \tag{4.8}$$

bringing the identity (4.4) into the form

$$\begin{aligned}\delta_{\mathbf{M}}S=\int_{t_1}^{t_2}dt\int_m\{&\delta_{\mathbf{M}}p^{ab}\delta_{\mathbf{N}}g_{ab}-\delta_{\mathbf{N}}p^{ab}\delta_{\mathbf{M}}g_{ab}\\ &+\delta_{\mathbf{M}}\pi\delta_{\mathbf{N}}\phi-\delta_{\mathbf{N}}\pi\delta_{\mathbf{M}}\phi+M\delta_{\mathbf{N}}H-N\delta_{\mathbf{M}}H\\ &+M^a\delta_{\mathbf{N}}H_a-N^a\delta_{\mathbf{M}}H_a+(M_{,a}N^a-N_{,a}M^a)H\\ &+(\epsilon(N^{,a}M-NM^{,a})+[\tilde{N},\tilde{M}]^a)H_a\}\equiv 0.\end{aligned} \tag{4.9}$$

The variations $\delta_{\mathbf{N}}H(x)$, $\delta_{\mathbf{N}}H_c(x)$ and $\delta_{\mathbf{M}}H(x)$, $\delta_{\mathbf{M}}H_c(x)$ in Eq. (4.9) can be written down term by term,

$$\begin{aligned}\delta_{\mathbf{N}}H(x)=&\frac{\delta H(x)}{\delta g_{abx'}}\delta_{\mathbf{N}}g_{abx'}+\frac{\delta H(x)}{\delta p^{abx'}}\delta_{\mathbf{N}}p^{abx'}+\frac{\delta H(x)}{\delta\phi_{x'}}\delta_{\mathbf{N}}\phi_{x'}\\ &+\frac{\delta H(x)}{\delta\pi^{x'}}\delta_{\mathbf{N}}\pi^{x'}+\frac{\partial H(x)}{\partial\lambda^a(x)}\delta_{\mathbf{N}}\lambda^a(x),\\ \delta_{\mathbf{N}}H_c(x)=&\frac{\delta H_c(x)}{\delta g_{abx'}}\delta_{\mathbf{N}}g_{abx'}+\frac{\delta H_c(x)}{\delta p^{abx'}}\delta_{\mathbf{N}}p^{abx'}\\ &+\frac{\delta H_c(x)}{\delta\phi_{x'}}\delta_{\mathbf{N}}\phi_{x'}+\frac{\delta H_c(x)}{\delta\pi^{x'}}\delta_{\mathbf{N}}\pi^{x'}.\end{aligned} \tag{4.10}$$

Rearrange first the multiplier terms

$$(M\delta_{\mathbf{N}}\lambda^a-N\delta_{\mathbf{M}}\lambda^a)\frac{\partial H}{\partial\lambda^a}$$

which appear in the integrand. The change of λ^a is given by the kinematical equation (II.2.6),

$$\begin{aligned}\delta_{\mathbf{N}}\lambda^a&=\delta_N\lambda^a+\delta_{\tilde{N}}\lambda^a\\ &=(\epsilon\lambda^a{}_{;\perp}+K^a_b\lambda^b)N+\epsilon\lambda^\perp N^{,a}+[\tilde{N},\tilde{\lambda}]^a,\end{aligned} \tag{4.11}$$

in which $g^{1/2}\lambda^\perp$ may be identified [see Eq. (III.11.5)] with the scalar field momentum. Interchanging the role of **M** and **N**, we get

$$\begin{aligned}&(M\delta_{\mathbf{N}}\lambda^a-N\delta_{\mathbf{M}}\lambda^a)\frac{\partial H}{\partial\lambda^a}\\ &\quad=\{\epsilon g^{-1/2}\pi(MN^{,a}-NM^{,a})+M[\tilde{N},\tilde{\lambda}]^a-N[\tilde{M},\tilde{\lambda}]^a\}\frac{\partial H}{\partial\lambda^a}.\end{aligned} \tag{4.12}$$

Next, replace the field velocities $\delta_{\mathbf{N}}g_{ab}$, $\delta_{\mathbf{N}}\phi$, $\delta_{\mathbf{M}}g_{ab}$, $\delta_{\mathbf{M}}\phi$ by the field momenta, substituting for them the expressions (4.3). The remaining terms then neatly combine into Poisson brackets and the identity (4.9) becomes

$$\begin{aligned}\int_{t_1}^{t_2}dt\int_{x\in m}\int_{x'\in m}\{&M(x)N(x')[H(x),H(x')]\\ &+M(x)N^b(x')[H(x),H_b(x')]\\ &+M^a(x)N(x')[H_a(x),H(x')]\\ &+M^a(x)N^b(x')[H_a(x),H_b(x')]\}\\ +\int_{t_1}^{t_2}dt\int_m\{&(M_{,a}N^a-N_{,a}M^a)H\\ &+(\epsilon(N^{,a}M-NM^{,a})+[\tilde{N},\tilde{M}]^a)H_a\\ &+(\epsilon g^{-1/2}\pi(MN^{,a}-NM^{,a})+M[\tilde{N},\tilde{\lambda}]^a\\ &-N[\tilde{M},\tilde{\lambda}]^a)\frac{\partial H}{\partial\lambda^a}\}\equiv 0.\end{aligned} \tag{4.13}$$

With this, all integrals are explicitly expressed as some bilinear combinations of N, N^a and M, M^a.

We now only use the arbitrariness of N, N^a and M, M^a, implying that all coefficients of this bilinear combination must vanish. This yields the desired Poisson brackets,

$$\begin{aligned}&[H(x),H(x')]\\ &\quad=-\epsilon\{(H^a(x)+I^a(x))\delta_{,a}(x,x')-(x\longleftrightarrow x')\},\end{aligned} \tag{4.14}$$

$$\begin{aligned}[H_a(x),H(x')]=&H(x)\delta_{,a}(x,x')\\ &+\frac{\partial H}{\partial\lambda^b}(x')\lambda^b{}_{,a}(x')\delta(x',x)\\ &-\frac{\partial H}{\partial\lambda^a}(x')\lambda^b(x')\delta_{,b}(x',x),\end{aligned} \tag{4.15}$$

$$[H_a(x),H_b(x')]=H_b(x)\delta_{,a}(x,x')-(ax\longleftrightarrow bx'), \tag{4.16}$$

with

$$I^a\equiv g^{-1/2}\pi g^{ab}\frac{\partial H}{\partial\lambda^b}. \tag{4.17}$$

As anticipated, the Poisson brackets (4.14)–(4.16) have the same structure as the corresponding Poisson brackets (III.12.31)–(III.12.33) for the generalized Hamiltonian dynamics of the scalar field propagating on a given background. Putting the scalar field variables equal to zero, $\phi=0$, $\pi=0$, $\lambda^a=0$, our derivation shows that the gravitational super-Hamiltonian H^g and supermomentum $H^g{}_a$ close according to the formulas

$$[H^g(x),H^g(x')]=-\epsilon(H^{g\ a}(x)\delta_{,a}(x,x')-(x\longleftrightarrow x')), \tag{4.18}$$

$$[H^g{}_a(x),H^g(x')]=H^g(x)\delta_{,a}(x,x'), \tag{4.19}$$

$$[H^g{}_a(x),H^g{}_b(x')]=H^g{}_b(x)\delta_{,a}(x,x')-(ax\longleftrightarrow bx'). \tag{4.20}$$

Equations (4.18) and (4.19) were shown to coincide with the contracted Bianchi identities in Sec. II.10.

As we have already mentioned, the closing relations (4.18)–(4.20) are also valid for the constraint functions $^*H$, $^*H_a$ obtained by eliminating the λ multipliers from the constraint functions H, H_a. These are counterparts of the closing relations (III.12.45)–(III.12.47) in generalized Hamiltonian dynamics of tensor fields on a given background. The elimination process runs exactly as in Sec. III.12, so that we do not need to discuss it here.

The relations (4.14)–(4.16) ensure that the initial value constraints (3.21) will hold on a deformed embedding, if they hold on the original embedding together with the λ equations $\partial H/\partial\lambda^{a}=0$. This is a necessary consistency test for Hamiltonian geometrodynamics with tensor sources in first-order form. Similarly, the closing relations (4.18)—(4.20) ensure that the initial value constraints (2.15) in pure geometrodynamics are preserved in course of the dynamical evolution.

5. ELIMINATION OF λ MULTIPLIERS

Our main result about the Hamiltonian structure of geometrodynamics with tensor sources was the additive form (3.9) of the total super-Hamiltonian accompanied by the additive form (3.6) of the gravitational momentum. These two implied that, however complicated the source-field super-Hamiltonian may be, the total super-Hamiltonian is always a quadratic function of the geometrodynamical momentum p^{ab}. Moreover, the coefficients of the quadratic form in p^{ab} are the components of DeWitt's supermetric (2.10), constructed entirely from the gravitational variables g_{ab}. Unfortunately, these statements are true only for sources described by the first-order formalism. This formalism is characterized by the presence of λ multipliers in addition to the dynamical degrees of freedom of the source field. At the end, we want to eliminate these nondynamical variables from the action. This can be done, but the detailed and possibly complicated structure of the field super-Hamiltonian then creeps into the p^{ab} dependence of the total super-Hamiltonian. The λ multipliers thus serve as a barrier which we put between the source-field variables and the geometrodynamical momentum. We must understand now the complications which arise when the barrier is put down.

For definiteness, consider again a covector field ϕ_{α}, $\lambda^{\alpha\beta}$. The $\lambda^{\perp\perp}$, $\lambda^{a\perp}$ projections become the field momenta π, π^{a}, but the $\lambda^{\perp b}$ and λ^{ab} projections remain in the hypersurface action (3.3) as Lagrange multipliers. They enter the action only through the translation part H_{t} of the super-Hamiltonian, while the tilt super-Hamiltonian $H_{\not t}$ and the supermomentum H_{a} are λ independent.

To eliminate the λ multipliers, we return to λ equations (3.20). Generically, these equations determine $\lambda^{\perp a}$, λ^{ab} as functions of the dynamical field variables $\phi_{\perp}$, $\pi^{\perp}$, ϕ_{a}, π^{a} and the geometrodynamical variables g_{ab}, p^{ab}. Substituting these functions into the super-Hamiltonian $H_{\not t}$, we get the modified super-Hamiltonian $^{*}H$ which correctly generates the dynamics of the canonical variables $\phi_{\perp}$, $\pi^{\perp}$, ϕ_{a}, π^{a}, g_{ab}, p^{ab} (cf. Sec. III.10). Because $H_{\not t}$ does not depend on λ, we need to discuss only what the elimination of multipliers does to the structure of $^{*}H_{t}$.

1. At a very general level, describe $\underline{\phi}$ by a Lagrangian potential Λ which is not necessarily a quadratic function of $\overline{\lambda}$. This means that the field equations for $\underline{\phi}$ propagating on a given background are nonlinear and the field super-Hamiltonian H^{ϕ}_{t} is not a quadratic function of the field momentum π. Let the field be derivatively coupled to gravity. While P^{cd} is a linear function of the multipliers $\lambda^{\perp b}$ and λ^{ab}, the expressions $\partial\Lambda/\partial\lambda^{ab}$ are definitely nonlinear in $\lambda^{\perp b}$ and λ^{ab}. The multipliers determined from Eqs. (3.20) will thus be some nonlinear functions of the geometrodynamical momentum p^{ab}. When we substitute them into H_{t}, $^{*}H_{t}$ will cease to be a quadratic function of the geometrodynamical momentum. This shows that the nonquadratic structure of the field potential Λ induces through the derivative coupling a nonquadratic structure of the super-Hamiltonian $^{*}H$ in the geometrodynamical momentum.

2. Traditionally, one works only with Lagrangian potentials Λ which are quadratic functions of $\overline{\lambda}$. Even more specifically, one restricts Λ to quadratic forms of both $\overline{\lambda}$ and $\underline{\phi}$, which ensures that the equations for $\underline{\phi}$ propagating on a given background $\underline{g}$ are linear and homogeneous. What can be said about the structure of $^{*}H$ after such fields are derivatively coupled to gravity?

First, it is obvious that the derivatives $\partial\Lambda/\partial\lambda^{\perp b}$ and $\partial\Lambda/\partial\lambda^{ab}$ are linear both in the multipliers and in the field momenta $\pi^{\perp}$, π^{a}. The expression P^{cd} is also a linear function of the multipliers and of the field momenta, so that Eqs. (3.20) determine the multipliers as some linear (though nonhomogeneous) functions of the field momenta and the geometrodynamical momentum p^{ab}. Notice, however, that the coefficients of the momenta p^{ab}, $\pi^{\perp}$, π^{a} in these linear functions will depend on the field coordinates $\phi_{\perp}$, ϕ_{a}, because $\phi_{\perp}$, ϕ_{a} are present in the $p^{cd}-P^{cd}$ terms in Eqs. (3.20).

With our choice of Λ, the field super-Hamiltonian H^{ϕ}_{t} is a quadratic function of the field momenta and the multipliers. After the multipliers are eliminated from H^{ϕ}_{t}, it becomes a quadratic function of the field momenta and the geometrodynamical momentum, with cross terms arising between $\pi^{\perp}$, π^{a}, and p^{ab}, and the coefficients depending not only on the metric g_{ab}, but also on the field coordinates $\phi_{\perp}$, ϕ_{a}. The same thing happens to the gravitational super-Hamiltonian H^{g}, since P^{ab} becomes a linear nonhomogeneous function of the momenta $\pi^{\perp}$, π^{a}, p^{ab}, with coefficients depending on $\phi_{\perp}$, ϕ_{a}, g_{ab} after the elimination of multipliers.

Briefly, denoting by $P_{A}=\{p^{ab},\pi^{\perp},\pi^{a}\}$ the collection of momenta, $^{*}H_{t}$ has the general form

$$^{*}H_{t}={}^{*}G^{AB}P_{A}P_{B}+{}^{*}A^{B}P_{B}+{}^{*}V, \qquad (5.1)$$

in which the coefficients $^{*}G^{AB}$, $^{*}A^{B}$, and $^{*}V$ depend on the metric g_{ab}, the field coordinates $\phi_{\perp}$, ϕ_{a}, and the space derivatives of these variables. In particular, the supermetric $^{*}G^{AB}$ does not have the block form [cf. Eq. (5.3)] which is characteristic for the fields with nonderivative gravitational coupling. The "Riemannian structure" which $^{*}G^{AB}$ imposes on the total configuration space $Q^{A}=\{g_{ab},\phi_{\perp},\phi_{A}\}$ does not thus induce a $\phi_{\perp}$, ϕ_{a}-independent "Riemannian structure" in the configuration space Riem$(\mathfrak{m})$ of the graviational variables. This seems to indicate that, when gravity is derivatively coupled to sources, not only the source-field variables, but the gravitational variables as well, may propagate off the light cones determined by the spacetime metric $\underline{g}$.

One can trace these complications back to the spacetime form (3.2) of the Einstein law of gravitation. To identify the hyperbolic operator which governs the propagation of the gravitational variables, one tries to isolate in Einstein's law all second-order derivatives of the metric with respect to time. It seems to have been largely overlooked in the literature on the Cauchy problem that such derivatives can occur not only in the Einstein tensor $G^{\alpha\beta}$, but also in the symmetrical energy—momentum tensor $T^{\alpha\beta}$. The canonical energy—momentum tensor $\Theta^{\alpha\beta}$, to be sure, depends only on the first derivatives of the metric, but the Belinfante—Rosenfeld spin energy—momentun tensor $S^{\alpha\beta\gamma}{}_{;\gamma}$ contains, in general, the second derivatives of the metric, because the spin tensor $S^{\alpha\beta\gamma}$ itself contains the first derivatives (See Sec. III.3 for the definitions of these objects from the field Lagrangian; in the second-order formalism, $\bar{\lambda}$ variables become expressed through the first covariant derivatives of the $\underline{\phi}$ variables and contain thus the first derivatives of the metric.) Moreover, from the construction of $S^{\alpha\beta\gamma}$ it is obvious that the second derivatives of the metric have the fields $\underline{\phi}$ in front of them as coefficients. Including these terms into the hyperbolic operator, it becomes obvious why the concept of gravitational supermetric independent of the source fields breaks down.

One can get the explicit form of the coefficients $*G^{AB}$, $*A^B$, $*V$ in concrete cases. One of the simplest examples is the covector field described by the Lagrangian potential $\Lambda = -\frac{1}{2}\lambda^{\alpha\beta}\lambda_{\alpha\beta}$, which was studied in Sec. III.10. One can see in detail that the supermetric does not have the block form, but the expressions are fairly complicated and we will not write them down in detail.

3. The whole situation vastly simplifies for fields with nonderivative gravitational coupling. As discussed in Sec. III.11, $P^{ab}=0$ for such fields and the extrinsic curvature drops out of the hypersurface Lagrangian. In a covector case, $\phi_\perp$ and $\lambda^{[ab]}$ together play the role of Lagrange multipliers. Varying them, we get Eqs. (III.11.21), (II.11.22),

$$g^{-1/2}\frac{\partial H}{\partial\lambda^{[ab]}} = -\phi_{[a,b]} + \frac{\partial\Lambda}{\partial\lambda^{[ab]}} = 0,$$

$$\frac{\partial H}{\partial\phi_\perp} = \epsilon\pi^a{}_{|a} + \frac{\partial\underline{\Lambda}}{\partial\phi} = 0. \tag{5.2}$$

Generically, these equations determine the multipliers $\phi_\perp$, $\lambda^{[ab]}$ in terms of g_{ab} and the dynamical field variables ϕ_a, π^a. However, the gravitational momentum $p^{ab} = \pi^{ab}$ does not enter into this solution. The modified super-Hamiltonian $*H_\perp$ thus has the form

$$*H_\perp = H^g(g_{ab}, \pi^{ab}) + *H^\phi_\perp(\phi_a, \pi^a). \tag{5.3}$$

Here, H^g is the standard super-Hamiltonian (2.10) for the free gravitational field. There are no cross terms between the gravitational momentum π^{ab} and the field momentum π^a. The gravitational super-Hamiltonian H^g is a quadratic form of π^{ab}, with DeWitt's supermetric constructed entirely out of the metric g_{ab}. The configuration space of the gravitational variables thus has a "Riemannian structure" independent of the sources. This holds even if the super-Hamiltonian of the source itself is nonquadratic in the field momenta. When Λ is taken as a quadratic form of the field variables $\lambda^{[\alpha\beta]}$, ϕ_α, the field super-Hamiltonian $*H^\phi_\perp$ also becomes quadratic in the field momentum. The coefficients of this quadratic function, however, depend on g_{ab}, so that the configuration space of the field variables ϕ_a never acquires a "Riemannian structure" of its own, independent of the metric. This means, of course, that the field variables always propagate on top of geometry.

Our conclusion is that one can build a consistent Dirac-ADM scheme for geometry derivatively coupled to general tensor sources described by a second-order formalism. However, quite apart from the usual difficulties with pure spin and positive definiteness of the field energy, other unpalatable features of the derivative gravitational coupling appear in the hypersurface formulation. They are connected with the loss of DeWitt's "Riemannian structure" in the space of gravitational coordinates, and disappear only for fields with nonderivative gravitational coupling.

ACKNOWLEDGMENT

This work was completed during my stay at the Niels Bohr Institute in the spring term 1976. My warm thanks are due to the Institute for its kind hospitality.

[1] K. Kuchař, J. Math. Phys. 17, 777 (1976).
[2] K. Kuchař, J. Math. Phys. 17, 792 (1976).
[3] K. Kuchař, J. Math. Phys. 17, 801 (1976).
[4] R. Arnowitt, S. Deser, and C.W. Misner, Phys. Rev. 116, 1322 (1959); "The Dynamics of General Relativity," in: *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962); P.A.M. Dirac, Can. J. Math. 3, 1 (1951).
[5] D.G. Boulware and S. Deser, J. Math. Phys. 8, 1468 (1962).
[6] A. Fisher and J. Marsden, 1976 Essay for the Gravity Research Foundation, Gloucester, Massachusetts.
[7] Y. Bruhat, "The Cauchy Problem," in: *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[8] D. Hilbert, Mitt. Nachr. Ges. Wiss. Göttingen, 395 (1915).
[9] P.A.M. Dirac, Proc. Roy. Soc. (London) A246, 333 (1958). For the ADM treatment, see the second reference in Ref. 4 and the original papers quoted there.
[10] K. Kuchǎr, J. Math. Phys. 13, 768 (1972).

Product integrals and the Schrödinger equation

John D. Dollard and Charles N. Friedman

Department of Mathematics, University of Texas, Austin, Texas 78712

(Received 7 October 1976; revised manuscript received 1 November 1976)

A brief introduction to product integration is given. The theory developed is used to give a simple and rigorous analysis of the asymptotic behavior ($r \to \infty$) of positive-energy solutions of the radial Schrödinger equation. Absence of positive-energy bound states is proved for various classes of potentials. It is shown that $E=1$ is the *only* positive energy for which the Wigner–von Neumann potential can have a positive-energy bound state. The results proved imply (as will be shown in a later publication) existence of the Møller wave matrices for the potential $V(R)=(\sin r)/r$ and various related potentials. A brief discussion is given to justify the WKB approximation which gives the wavefunction asymptotically for large positive values of the energy E.

INTRODUCTION

The purpose of this paper is to give a derivation, using product integrals, of some properties of the radial Schrödinger equation. Most of the results proved are already known in one form or another, but new results are proved about potentials of the Wigner–von Newmann type[1] and some associated potentials. The use of the product integral allows a swift presentation with a single underlying idea and very simple proofs. In the first section we review the concept of product integration and some facts about product integrals.

I. PRODUCT INTEGRALS

Notation: $\mathbb{R}$ and $\mathbb{C}$ denote the real and complex numbers respectively. $\mathbb{C}_n$ is the set of $n\times 1$ matrices with entries in $\mathbb{C}$, and $\mathbb{C}_{n\times n}$ is the set of $n\times n$ matrices with entries in $\mathbb{C}$. If

$$\varphi \in \mathbb{C}_n \quad \text{with } \varphi = \begin{pmatrix} c_1 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{pmatrix},$$

we write

$$\|\varphi\| = \left(\sum_{i=1}^{n} |c_i|^2\right)^{1/2}. \tag{1}$$

For $B \in \mathbb{C}_{n\times n}$ we put

$$\|B\| = \sup_{\substack{\varphi\in\mathbb{C}_n \\ \|\varphi\|=1}} \|B\varphi\| \tag{2}$$

and

$$\exp B = \sum_{n=0}^{\infty} \frac{B^n}{n!} \tag{3}$$

We take for granted various standard properties of $\mathbb{C}_n$ and $\mathbb{C}_{n\times n}$. All statements about matrix-valued functions (concerning continuity, differentiability, etc.) are to be understood entrywise.

If $A:[a,b]\to\mathbb{C}_{n\times n}$ is a function, the product integral of A over $[a,b]$ can be defined if A is (entrywise) Lebesgue integrable.[2,7] In this work we consider only functions A which are continuous, since we shall be concerned with differential equations of the type

$$\frac{dU}{dx}(x) = A(x)U(x), \quad U(a)=U_0, \tag{4}$$

where U_0 belongs either to $\mathbb{C}_n$ or $\mathbb{C}_{n\times n}$. In such equations, it is natural to assume that A is continuous. We now sketch the definition of the product integral for a continuous function A.[3] As motivation, we consider the problem of determining the value $U(b)$ from the initial value problem (4). [We note that $U(b)$ is uniquely determined by (4) since by standard arguments the solution of (4) is unique.] If A were a constant function, the answer would be

$$U(b) = \exp[A(b-a)]U_0 \tag{5}$$

since the solution $U(x)$ of (4) is clearly $\exp[A(x-a)]$. If A is not constant, a plausible method of constructing an approximate value for $U(b)$ is the following: Let $P=\{s_0, s_1, \ldots, s_n\}$ be a partition of $[a,b]$, with $\Delta s_i = s_i - s_{i-1}$. Let $\mu(P)$ denote the mesh of P (length of its longest subinterval). Now A is continuous. Thus if $\mu(P)$ is small, all the values taken by A in any subinterval $[s_{i-1}, s_i]$ will be close together. Given the value $U(a) = U(x_0)$, we can compute an approximate value for $U(x_1)$ by assuming that, in the interval $[s_0, s_1]$, A takes the constant value $A(s_1)$. The result [analogous to (5)] is

$$U(s_1) \approx \exp[A(s_1)\Delta s_1]\, U(s_0). \tag{6}$$

An approximate value for $U(s_2)$ is found by assuming that, in $[s_1, s_2]$, A takes the constant value $A(s_2)$:

$$U(s_2) \approx \exp[A(s_2)\Delta s_2]U(s_1) \approx \exp[A(s_2)\Delta s_2] \times \exp[A(s_1)\Delta s_1]U(s_0). \tag{7}$$

Continuing this process, we arrive at an approximate expression for $U(b)$:

$$U(b) = U(s_n) \approx \left\{\prod_{i=1}^{n} \exp[A(s_i)\Delta s_i]\right\} U(s_0). \tag{8}$$

It should be noted that, in the product on the right-hand side of (8), the order of the factors is important, because the various $A(s_i)$ in general will not commute. By construction, in this product smaller values of s occur further to the right. We shall refer to such a product as a *Riemann product* for A.

Now, if the mesh of P is very small, it is reasonable to suppose that the approximation made above is quite good. One is then led to conjecture that

$$U(b)=\lim_{\mu(P)\to 0}\prod_{i=1}^{n}\exp[A(s_i)\Delta s_i]U(s_0). \tag{8'}$$

This is indeed correct.[3] In fact, the limit of the Riemann product exists independent of the initial condition U_0, and defines the *product integral* of A over $[a,b]$, denoted $\Pi_a^b\exp[A(s)\,ds]$:

Definition 1: Let A: $[a,b]\to C_{n\times n}$ be continuous. With notation as above, the product integral of A over $[a,b]$ is defined by

$$\prod_a^b\exp[A(s)\,ds]=\lim_{\mu(P)\to 0}\prod_{i=1}^{n}\exp[A(s_i)\Delta s_i]. \tag{9}$$

Thus Eq. (8) becomes

$$U(b)=\prod_a^b\exp[A(s)\,ds]U(a). \tag{10}$$

We shall not prove existence of the product integral here. (The interested reader can consult Ref. 3.) We now discuss some facts about product integrals. First, if $x,y\in[a,b]$ with $y<x$, then $\Pi_y^x\exp[A(s)\,ds]$ can be defined by partitioning $[y,x]$ as we previously partitioned $[a,b]$. If U is the solution of (4), we find the equation analogous to (10):

$$U(x)=\prod_y^x\exp[A(s)\,ds]U(y). \tag{11}$$

Equation (11) states that $\Pi_y^x\exp[A(s)\,ds]$ is the propagator for Eq. (4). Certain properties of $\Pi_y^x\exp[A(s)\,ds]$ follow immediately from this fact, namely:

Property 1: Let $A:[a,b]\to\mathbb{C}_{n\times n}$ be continuous. If $x,y\in[a,b]$ with $x>y$, then

$$\frac{d}{dx}\prod_y^x\exp[A(s)\,ds]=A(x)\prod_y^x\exp[A(s)\,ds]. \tag{12}$$

Property 2: Let $A:[a,b]\to\mathbb{C}_{n\times n}$ be continuous. If $x,y,z\in[a,b]$ with $x>y>z$, then

$$\prod_z^x\exp[A(s)\,ds]=\prod_y^x\exp[A(s)\,ds]\prod_z^y\exp[A(s)\,ds]. \tag{13}$$

These properties can also be obtained directly from the definition of the product integral. Equations (12) and (13) correspond to the following properties of the usual Riemann integral:

$$\frac{d}{dx}\int_y^x A(s)\,ds=A(x) \tag{14}$$

and

$$\int_z^x A(s)\,ds=\int_y^x A(s)\,ds+\int_z^y A(s)\,ds. \tag{15}$$

One useful facet of the theory of product integration is the strong formal analogy between this theory and the usual theory of Riemann integration, as just illustrated. Careful reasoning based on this analogy leads to conjectures about properties of product integrals, which usually turn out to be true. One must only remember that, in making this analogy, sums in ordinary integration go over to products in the present theory. Likewise the additive neutral element 0 goes over to the multiplicative neutral element I (the identity matrix) and the additive inverse $-B$ goes over to the multiplicative inverse B^{-1}.

Property 3: Let $A:[a,b]\to\mathbb{C}_{n\times n}$ be continuous, and let $x,y\in[a,b]$ with $y<x$. Then $\Pi_y^x\exp[A(s)\,ds]$ is nonsingular.

Proof: Using familiar rules of matrix calculation, we compute the determinant of $\Pi_y^x\exp[A(s)\,ds]$:

Letting P denote a partition of $[y,x]$, we have

$$\begin{aligned}\det\prod_y^x\exp[A(s)\,ds]&=\lim_{\mu(P)\to 0}\det\prod_{i=1}^{n}\exp[A(s_i)\Delta s_i]\\&=\lim_{\mu(P)\to 0}\prod_{i=1}^{n}\det\exp[A(s_i)\Delta s_i]\\&=\lim_{\mu(P)\to 0}\prod_{i=1}^{n}\exp[\mathrm{tr}A(s_i)\Delta s_i]\\&=\lim_{\mu(P)\to 0}\exp\left(\sum_{i=1}^{n}\mathrm{tr}A(s_i)\Delta s_i\right)\\&=\exp[\textstyle\int_y^x\mathrm{tr}A(s)\,ds]\neq 0.\end{aligned} \tag{16}$$

This proves the result.

Property 3 allows us to make the following definition, analogous to that in the ordinary theory of integration:

Definition 2: Let $A:[a,b]\to\mathbb{C}_{n\times n}$ be continuous and let $x,y\in[a,b]$ with $x\le y$. We define

$$\prod_x^x\exp[A(s)\,ds]=I \tag{17}$$

and

$$\prod_y^x\exp[A(s)\,ds]=\left(\prod_x^y\exp[A(s)\,ds]\right)^{-1}. \tag{18}$$

Using Definition 2, it is easy to show that

Property 4: Properties 1, 2, and 3 are valid for any $x,y,z\in[a,b]$.

If the family $\{A(s)\,|\,s\in[a,b]\}$ is *commutative* $\{A(s)A(s')=A(s')A(s)$ for all $s,s'\in[a,b]\}$, then the product integral of A can be evaluated explicity:

Property 5: Suppose A is continuous on $[a,b]$ and $\{A(s)\,|\,s\in[a,b]\}$ is a commutative family. Let $x,y\in[a,b]$. Then

$$\prod_y^x\exp[A(s)\,ds]=\exp[\textstyle\int_y^x A(s)\,ds]. \tag{19}$$

Proof: We give the proof for the case $y<x$. (The general result follows immediately.) Letting P denote a partition of $[y,x]$ and using the assumed commutativity, we have

$$\begin{aligned}\prod_y^x\exp[A(s)\,ds]&=\lim_{\mu(P)\to 0}\prod_{i=1}^{n}\exp[A(s_i)\Delta s_i]\\&=\lim_{\mu(P)\to 0}\exp\left(\sum_{i=1}^{n}A(s_i)\Delta s_i\right)\\&=\exp[\textstyle\int_y^x A(s)\,ds].\end{aligned} \tag{20}$$

This proves the result.

In our later work it will be useful to have certain bounds on product integrals, which we now obtain:

Property 6: Let A be continuous on $[a,b]$. Let $x,y\in[a,b]$ with $y\le x$. Then

$$\|\prod_y^x \exp[A(s)\,ds]\| \leqslant \exp[\textstyle\int_y^x \|A(s)\|ds]. \tag{21}$$

Proof: If $y=x$, Eq. (21) is obvious. If $y<x$, let P denote a partition of $[y,x]$. Then

$$\begin{aligned}\|\prod_y^x \exp[A(s)\,ds]\| &= \lim_{\mu(P)\to 0} \|\prod_{i=1}^n \exp[A(s_i)\Delta s_i]\| \\ &\leqslant \lim_{\mu(P)\to 0} \prod_{i=1}^n \|\exp[A(s_i)\Delta s_i]\| \\ &\leqslant \lim_{\mu(P)\to 0} \prod_{i=1}^n \exp[\|A(s_i)\|\Delta s_i] \\ &= \lim_{\mu(P)\to 0} \exp\left(\sum_{i=1}^n \|A(s_i)\|\Delta s_i\right) \\ &= \exp[\textstyle\int_y^x \|A(s)\|ds]. \end{aligned} \tag{22}$$

This proves Property 6.

Property 7: Let A be continuous on $[a,b]$. Let $x,y \in [a,b]$ with $y \leqslant x$. Then

$$\|\prod_y^x \exp[A(s)\,ds]-I\| \leqslant \exp[\textstyle\int_y^x \|A(s)\|ds]-1. \tag{23}$$

Proof: Because of Eq. (12) and the fact that $\prod_y^y \exp[A(s)\,ds]$ is the identity, we have

$$\prod_y^x \exp[A(s)\,ds] = I + \int_y^x A(s) \prod_y^s \exp[A(u)\,du]\,ds. \tag{24}$$

Using the bound from Property 6, we find

$$\begin{aligned}\|\prod_y^x \exp[A(s)\,ds]-I\| &\leqslant \int_y^x \|A(s)\| \exp[\textstyle\int_y^s \|A(u)\|\,du]\,ds \\ &= \exp[\textstyle\int_y^x \|A(u)\|\,du]-1,\end{aligned} \tag{25}$$

where the last inequality is elementary. This proves Property 7.

There are certain very useful rules for manipulating product integrals which will be the basis for our later analysis of the Schrödinger equation. These are

Property 8 (*the sum rule*): Let A and B be continuous on $[a,b]$, and let $x,y \in [a,b]$. Let

$$P(x) = \prod_y^x \exp[A(s)\,ds]. \tag{26}$$

Then

$$\prod_y^x \exp[A(s)+B(s)]\,ds = P(x) \prod_y^x \exp[P^{-1}(s)B(s)P(s)\,ds]. \tag{27}$$

Proof: The two sides of (27) agree when $x=y$, and (considered as functions of x) they satisfy the same differential equation, as a brief computation shows. Indeed, if $F(x,y)$ denotes either side of (27), then

$$\frac{dF}{dx}(x,y) = [A(x)+B(x)]F(x,y), \quad F(y,y)=I. \tag{28}$$

The uniqueness theorem for the initial value problem (28) shows that the two sides of (27) are equal, proving the sum rule.

It is sometimes useful to be able to make a similarity transformation on the integrand of a product integral. If T is a constant nonsingular matrix, then directly from the definition of the product integral and the formula

$$T^{-1}e^B T = \exp(T^{-1}BT) \tag{29}$$

one finds

$$T^{-1}\prod_y^x \exp[A(s)\,ds]T = \prod_y^x \exp[T^{-1}A(s)T\,ds] \tag{30}$$

(exercise for the reader). We shall also have occasion to use a "similarity transformation" in which T is not a constant:

Property 9 (*the similarity rule*): Suppose that $T:[a,b] \to \mathbb{C}_{n\times n}$ has a continuous derivative T'. Also suppose that $T(x)$ is nonsingular for all $x \in [a,b]$. Let A be continuous on $[a,b]$. Then for $x,y \in [a,b]$ we have

$$\begin{aligned} &T^{-1}(x)\prod_y^x \exp[A(s)\,ds]T(y) \\ &\quad= \prod_y^x \exp[\{T^{-1}(s)A(s)T(s) - T^{-1}(s)T'(s)\}\,ds]. \end{aligned} \tag{31}$$

Proof: Multiply both sides of (31) on the left by $T(x)$. The two sides of the resulting equation agree when $x=y$ and satisfy the same differential equation. As in the proof of the sum rule, we find that the two sides agree for all $x \in [a,b]$, completing the proof.

In studying asymptotic behavior of solutions of differential equations of the type (4), we will need some information on *improper* product integrals:

Definition 3: Let $A:[a,\infty) \to \mathbb{C}_{n\times n}$ be continuous. We define the improper product integral $\prod_a^\infty \exp[A(s)\,ds]$ by

$$\prod_a^\infty \exp[A(s)\,ds] = \lim_{x\to\infty} \prod_a^x \exp[A(s)\,ds] \tag{32}$$

provided the indicated limit exists.

An improper product integral can be singular. [Take $A(s)=-I$ for all s. By Property 5 we have

$$\prod_a^x \exp[A(s)\,ds] = \exp[-I(x-a)] \xrightarrow[x\to\infty]{} 0.] \tag{33}$$

We are interested in conditions under which $\prod_a^\infty \times \exp[A(s)\,ds]$ exists and is nonsingular. Under these conditions, the solution $U(x)$ of (4) will have a nonzero limit as $x\to\infty$ if $U(a)\neq 0$.

Property 10: Suppose that $A:[a,\infty) \to \mathbb{C}_{n\times n}$ is continuous and that $A \in L^1(a,\infty)$, i.e.,

$$\int_a^\infty \|A(s)\|ds < \infty. \tag{34}$$

Then $\prod_a^\infty \exp[A(s)\,ds]$ exists and is nonsingular.

Proof: Let $x,y \in [a,\infty]$ with $y \leqslant x$. Using the multiplicative property (Property 2) and the bounds from Properties 6 and 7, we have

$$\begin{aligned} &\|\prod_a^x \exp[A(s)\,ds] - \prod_a^y \exp[A(s)\,ds]\| \\ &\quad= \left\|\left(\prod_y^x \exp[A(s)\,ds]-I\right)\prod_a^y \exp[A(s)\,ds]\right\| \\ &\quad\leqslant \|\prod_y^x \exp[A(s)\,ds]-I\|\,\|\prod_a^y \exp[A(s)\,ds]\| \\ &\quad\leqslant \{\exp[\textstyle\int_y^x \|A(s)\|\,ds]-1\}\exp[\textstyle\int_a^u \|A(s)\|\,ds]. \end{aligned} \tag{35}$$

Because of (34) the second factor on the right-hand side of (35) is bounded. Also because of (34), the integral $\int_y^x \|A(s)\|\, ds$ can be made arbitrarily small by taking x and y large enough, and this shows that for large enough x and y the first factor on the right-hand side of (35) will be as small as desired. Thus $\Pi_a^x \times \exp[A(s)\, ds]$ has the Cauchy property and hence converges as $x \to \infty$. To see that the limit is nonsingular, we compute

$$\begin{aligned} \det \prod_a^\infty \exp[A(s)\, ds] &= \lim_{x\to\infty} \det \prod_a^x \exp[A(s)\, ds] \\ &= \lim_{x\to\infty} \exp\left[\int_a^x \operatorname{tr} A(s)\, ds\right] \\ &= \exp\left[\int_a^\infty \operatorname{tr} A(s)\, ds\right] \neq 0, \end{aligned} \tag{36}$$

where existence of $\int_a^\infty \operatorname{tr} A(s)\, ds$ is implied by (34). This proves Property 10.

Corollary: Suppose that A and B are continuous functions from $[a, \infty)$ to $\mathbb{C}_{n\times n}$. Suppose that $\Pi_a^\infty \exp[A(s)\, ds]$ exists and is nonsingular and that $B \in L^1(a, \infty)$. Then $\Pi_a^\infty \exp[\{A(s) + B(s)\}\, ds]$ exists and is nonsingular.

Proof: Let

$$P(x) = \prod_a^x \exp[A(s)\, ds]. \tag{37}$$

By the sum rule, we have

$$\prod_a^x \exp[\{A(s) + B(s)\}\, ds = P(x) \prod_a^x \exp[P^{-1}(s) B(s) P(s)\, ds]. \tag{38}$$

By hypothesis, the first factor on the right-hand side of (38) has a nonsingular limit as $x \to \infty$. If the same can be shown for the second factor, the corollary will be proved. By Property 10, the second factor has a nonsingular limit if $P^{-1}BP \in L^1(a, \infty)$. Since $B \in L^1(a, \infty)$, we will have $P^{-1}BP \in L^1(a, \infty)$ if P and P^{-1} are bounded on $[a, \infty)$. But this is an easy consequence of the fact that $P(x)$ is continuous (even differentiable) on (a, ∞) and has a nonsingular limit as $x \to \infty$. Thus the corollary is proved.

In the analysis of potentials of the Wigner—von Neumann type, we shall need the following more delicate convergence theorem[2]:

Property 11: Suppose that $A : [a, \infty) \to \mathbb{C}_{n\times n}$ is continuous Suppose that the *improper* integral

$$\operatorname{imp} \int_a^\infty A(s)\, ds = \lim_{b\to\infty} \int_a^b A(s)\, ds \tag{39}$$

exists. [This can occur through cancellations even if $A \notin L^1(a, \infty)$.] Write

$$H(x) = \operatorname{imp} \int_x^\infty A(s)\, ds. \tag{40}$$

Suppose that $HA \in L^1(a, \infty)$. I.e., suppose that

$$\int_a^\infty \|H(s)A(s)\|\, ds < \infty. \tag{41}$$

Then $\Pi_a^\infty \exp[A(s)\, ds]$ exists and is nonsingular.

Proof: By the multiplicative property (Property 2) it is clearly enough to show that $\Pi_b^\infty \exp[A(s)\, ds]$ exists and is nonsingular for some $b \geq a$. By inspection, $H(x) \to 0$ as $x \to \infty$, so we may choose b so large that $\|H(x)\| \leq \frac{1}{2}$ for $x \geq b$. Then

$$\|[I + H(x)]^{-1}\| \leq 2 \quad \text{for } x \geq b. \tag{42}$$

Let

$$P(x) = \prod_b^x \exp[A(s)\, ds] \tag{43}$$

and

$$Q(x) = [I + H(x)]P(x). \tag{44}$$

Since $I + H(x) \to I$ as $x \to \infty$, existence and nonsingularity of a limit for $P(x)$ is equivalent to the same properties for $Q(x)$. Now since $H' = -A$ we have

$$\begin{aligned} Q'(x) &= -A(x)P(x) + (I + H(x))A(x)P(x) \\ &= H(x)A(x)P(x) \\ &\equiv C(x)Q(x), \end{aligned} \tag{45}$$

where

$$C(x) = H(x)A(x)[I + H(x)]^{-1}. \tag{46}$$

From (45) and (11) we obtain

$$Q(x) = \prod_b^x \exp[C(s)\, ds] Q(b). \tag{47}$$

Now by (41) and (42) it is clear that $C \in L^1(b, \infty)$. Thus $\Pi_b^x \exp[C(s)\, ds]$ has a nonsingular limit Π_+ as $x \to \infty$. Then $Q(x)$ converges to $\Pi_+ Q(b)$. $Q(b)$ is nonsingular by (44) and nonsingularity of $[I + H(b)]$ and $P(b)$. So $\Pi_+ Q(b)$ is nonsingular, proving Property 11. Property 11 can be generalized in various ways, as we remark in a subsequent paper. (In particular, it is enough that the commutator $[H, A]$ belong to L^1.) However, we shall not need these generalizations here.

II. THE RADIAL SCHRÖDINGER EQUATION

In this section, we use product integrals to study the asymptotic behavior of stationary-state positive-energy solutions of the radial Schrödinger equation. We adopt units in which $\hbar = 1$ and the mass m is $\frac{1}{2}$. We begin by writing down the three-dimensional Schrödinger equation with a central potential. In our units, this equation has the form

$$(-\Delta + V)\psi = E\psi \quad (E > 0). \tag{48}$$

In this equation, Δ is Laplace's operator and V is the operation of multiplication by a real-valued function $V(r)$, where r denotes the distance from the origin. In order to guarantee that $-\Delta + V$ is self-adjoint on an appropriate domain in $L^2(\mathbb{R}^3)$, it is necessary to make further assumptions on V. A typical assumption is that V is a Kato potential (i.e., $V = V_1 + V_2$, where V_1 is square-integrable over $\mathbb{R}^3$ and V_2 is bounded on $\mathbb{R}^3$). This assumption guarantees the desired self-adjointness.[4] However, once we have passed to the radial Schrödinger equation, as we shall do below, all the statements we shall prove hold independent of the assumption that V is a Kato potential, and we shall therefore suppress this assumption in stating them. When we later use our results on the radial Schrödinger equation to make statements about the nonexistence of positive-energy bound states for the Hamiltonian $-\Delta + V$, we will enforce the requirement that V be a Kato potential. In order to use the theory of product integration developed above, we assume that $V(r)$ is continuous on $(0, \infty)$. (This allows bad behavior at the origin.)

Separating radial and angular variables in (48) as usual, we obtain for each nonnegative integer l the radial Schrödinger equation with angular momentum l:

$$R''(r) + (2/r)R'(r) = \{V(r) + [l(l+1)/r^2] - E\}R(r) \quad (r > 0). \tag{49}$$

If we make the usual substitution

$$R(r) = \chi(r)/r \tag{50}$$

and set

$$W(r) = V(r) + l(l+1)/r^2, \tag{51}$$

Eq. (49) becomes

$$\chi''(r) = (W(r) - E)\,\chi(r) \quad (r > 0). \tag{52}$$

Clearly the point $r = 0$ is a singular point of either of the Eqs. (49) or (52). To achieve boundedness of R near 0, one must [in view of (50)] choose a solution of (52) which approaches zero as $r \to 0$. However, we shall not be much concerned with this condition, because our interest is to give the asymptotic form of *any* solution of (52). Introducing the matrix function

$$\varphi(r) = \begin{pmatrix} \chi(r) \\ \chi'(r) \end{pmatrix}, \tag{53}$$

Eq. (52) can be rewritten as

$$\varphi'(r) = A(r)\varphi(r) \tag{54}$$

with

$$A(r) = \begin{pmatrix} 0 & 1 \\ W(r) - E & 0 \end{pmatrix}. \tag{55}$$

By our work in the last section, the unique solution of (54) on $(0, \infty)$ is

$$\varphi(r) = \prod_a^r \exp[A(s)\,ds]\varphi(a) \quad (a > 0). \tag{56}$$

The asymptotic form of φ (and hence of χ) can now be studied using (56). If the potential V has "short range," then one expects the solutions of (54) to behave asymptotically like plane waves. That is, letting k denote the positive square root of E, the solutions should behave like linear combinations of $\exp(ikx)$ and $\exp(-ikx)$. Our first result shows that this is the case if $V \in L^1(a, \infty)$ for some $a > 0$.

To begin our analysis, we write

$$A(r) = A_0 + A_1(r), \tag{57}$$

where

$$A_0 = \begin{pmatrix} 0 & 1 \\ -E & 0 \end{pmatrix} \tag{58}$$

and

$$A_1(r) = \begin{pmatrix} 0 & 0 \\ W(r) & 0 \end{pmatrix}. \tag{59}$$

If the potential $V(r)$ is "small" at large r, then $W(r)$ will also be small and the leading term of $A(r)$ will be the constant term A_0. The characteristic equation

$$\det(\lambda - A_0) = 0 \tag{60}$$

shows that A_0 has the eigenvalues $\pm ik$, where k is the positive square root of E. Letting

$$M = \begin{pmatrix} 1 & 1 \\ ik & -ik \end{pmatrix}, \tag{61}$$

the matrix M can be used to bring A_0 to diagonal form:

$$M^{-1}A_0M = \begin{pmatrix} ik & 0 \\ 0 & -ik \end{pmatrix} \equiv B_0. \tag{62}$$

Since $A_1(r)$ is "small" compared to A_0 when r is large, the matrix $B(r) = M^{-1}A(r)M$ should then be approximately diagonal. A simple computation yields

$$B(r) = M^{-1}A(r)M = \begin{pmatrix} ik & 0 \\ 0 & -ik \end{pmatrix} + \frac{iW(r)}{2k}\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \equiv B_0 + B_1(r). \tag{63}$$

It is something of an advantage to be able to deal with diagonal or approximately diagonal matrices. To exploit (63), we use (30) to find

$$M^{-1}\prod_a^r \exp[A(s)\,ds]M = \prod_a^r \exp[M^{-1}A(s)M\,ds] = \prod_a^r \exp[B(s)\,ds] \tag{64}$$

or

$$\prod_a^r \exp[A(s)\,ds] = M\prod_a^r \exp[B(s)\,ds]M^{-1}. \tag{65}$$

To analyze the asymptotic behavior of the left-hand side of (65), we thus need only study the product integral of B. In our first applications, we will use the sum rule to "pull out" the first term in (63).

Theorem 1: Suppose that V is continuous on $(0, \infty)$ and that $V \in L^1(a, \infty)$, i.e.,

$$\int_a^\infty |V(s)|\,ds < \infty. \tag{66}$$

Let φ be as in (56) with $\varphi(a) \neq 0$. Then there are constant 2×1 matrices C_+ and C_- whose (1, 1) entries are not both zero, such that

$$\varphi(r) \approx C_+\exp(ikr) + C_-\exp(-ikr), \quad r \to \infty, \tag{67}$$

where the notation $\approx$ of (67) means

$$\lim_{r\to\infty} \|\varphi(r) - C_+\exp(ikr) - C_-\exp(-ikr)\| = 0.$$

Remark: Since the (1, 1) entries of C_+ and C_- are not both zero, the last equation allows us to conclude for the wavefunction $\chi(r)$ of (53) that

$$\lim_{r\to\infty} |\chi(r) - d_+\exp(ikr) - d_-\exp(-ikr)| = 0, \tag{68}$$

where the coefficients d_+ and d_- are not both zero.

Proof of Theorem 1: With notation as above we have, using the sum rule,

$$\prod_a^r \exp[B(s)\,ds] = \prod_a^r \exp[\{B_0 + B_1(s)\}\,ds]$$

$$= P_0(r) \prod_a^r \exp[\hat{B}_1(s)\,ds], \tag{69}$$

where

$$P_0(r) = \prod_a^r \exp(B_0\,ds) \tag{70}$$

$$\hat{B}_1(s) = P_0^{-1}(s) B_1(s) P_0(s). \tag{71}$$

Since the family $\{B_0 | s \in [a, \infty)\}$ is obviously commutative, Property 5 yields

$$P_0(r) = \exp\left(\int_a^r B_0\,ds\right) = \exp[B_0(r-a)]$$

$$= \begin{pmatrix} \exp[ik(r-a)] & 0 \\ 0 & \exp[-ik(r-a)] \end{pmatrix}. \tag{72}$$

It is then simple to compute $\hat{B}_1(s)$:

$$\hat{B}_1(s) = \frac{iW(s)}{2k} \begin{pmatrix} -1 & -\exp[-2ik(s-a)] \\ \exp[2ik(s-a)] & 1 \end{pmatrix}. \tag{73}$$

Now because of (66) and the fact that $l(l+1)/s^2$ clearly belongs to $L^1(a, \infty)$, we have $W \in L^1(a, \infty)$. The matrix in (72) has norm 2, so that

$$\|\hat{B}_1(s)\| = |W(s)|/k \tag{74}$$

and it follows that $\hat{B}_1 \in L^1(a, \infty)$. Thus by Property 10 the improper product integral

$$\hat{\Pi}_1 = \prod_a^\infty \exp[\hat{B}_1(s)\,ds] \tag{75}$$

exists and is nonsingular. From (69) and the fact that $P_0(r)$ is unitary, we then have

$$\left\|\prod_a^r \exp[B(s)\,ds] - P_0(r)\hat{\Pi}_1\right\|$$

$$= \left\|P_0(r)\left\{\prod_a^r \exp[\hat{B}_1(s)\,ds] - \hat{\Pi}_1\right\}\right\|$$

$$= \left\|\prod_a^r \exp[\hat{B}_1(s)\,ds] - \hat{\Pi}_1\right\| \underset{r\to\infty}{\longrightarrow} 0 \tag{76}$$

or

$$\prod_a^r \exp[B(s)\,ds] \approx P_0(r)\hat{\Pi}_1. \tag{77}$$

Thus by (65) we have

$$\prod_a^r \exp[A(s)\,ds] \approx M P_0(r) \hat{\Pi}_1 M^{-1} \tag{78}$$

and

$$\varphi(r) = \prod_a^r \exp[A(s)\,ds]\varphi(a)$$

$$\approx M \begin{pmatrix} \exp[ik(r-a)] & 0 \\ 0 & \exp[-ik(r-a)] \end{pmatrix} \hat{\Pi}_1 M^{-1}\varphi(a). \tag{79}$$

Writing $\binom{p}{q} = \hat{\Pi}_1 M^{-1}\varphi(a)$, the facts that $\varphi(a) \neq 0$ and $\hat{\Pi}_1 M^{-1}$ is nonsingular imply that not both p and q equal zero. Then (79) yields

$$\varphi(r) \approx M \begin{pmatrix} \exp[ik(r-a)]p \\ \exp[-ik(r-a)]q \end{pmatrix}$$

$$= \begin{pmatrix} p \\ ikp \end{pmatrix} \exp[ik(r-a)] + \begin{pmatrix} q \\ -ikq \end{pmatrix} \exp[-ik(r-a)].$$

This last equation is clearly the same as (67), and by inspection the (1, 1) entries of C_+ and C_- are not both zero.

We remark that Theorem 1 covers any continuous potential $V(r)$ with $V \in L^2(\mathbb{R}^3)$ [this implies $V \in L^1(a, \infty)$] and any potential of the form

$$V(r) = 1/r^\beta, \quad \beta > 1. \tag{80}$$

It does not, however, cover the Coulomb potential

$$V(r) = \lambda/r \quad (\lambda \text{ real}). \tag{81}$$

We shall analyze this and similar potentials later. For the present we turn to potentials of the form

$$V(r) = \lambda(\sin\mu r^\alpha)/r^\beta \tag{82}$$

where λ is a real number and μ, α, β satisfy the conditions

$$\mu > 0, \quad \alpha > 0, \quad \beta > 0, \quad \alpha + 2\beta > 2. \tag{83}$$

[Note that the potential $\lambda(\sin\mu r)/r$ satisfies these conditions.] We remark that if (83) holds, then $\alpha + \beta > \alpha/2 + \beta > 1$. If V is given by (83), then integrating by parts we find

$$\int_r^R V(s)\,ds = \frac{-\lambda\cos\mu s^\alpha}{\mu\alpha s^{\beta+\alpha-1}}\Bigg|_r^R$$

$$- \frac{\lambda(\beta+\alpha-1)}{\mu\alpha}\int_r^R \frac{\cos\mu s^\alpha}{s^{\beta+\alpha}}\,ds. \tag{84}$$

From this we see that the following improper integral exists:

$$\operatorname{imp}\int_r^\infty V(s)\,ds = \frac{\lambda\cos\mu r^\alpha}{\mu\alpha r^{\beta+\alpha-1}} - \frac{\lambda(\beta+\alpha-1)}{\mu\alpha}$$

$$\times \int_r^\infty \frac{\cos\mu s^\alpha}{s^{\beta+\alpha}}\,ds. \tag{85}$$

A simple estimate of the integral in (85) now yields

$$\operatorname{imp}\int_r^\infty V(s)\,ds = O(1/r^{\beta+\alpha-1}), \quad r \to \infty. \tag{86}$$

We can now prove

Theorem 2: Suppose that $V(r)$ is given by (82) and that (83) holds. Then, if $\varphi(r)$ is as in (56) with $\varphi(a) \neq 0$, the conclusion (67) of Theorem 1 holds, *except* that if $\alpha = 1$ then the conclusion of Theorem 1 may not hold for the single value $k = \mu/2$.

Proof: For simplicity we will deal only with the case $\alpha = 1$, for which an exceptional value of k arises. (The argument for $\alpha \neq 1$ is entirely analogous.) We note that with $\alpha = 1$, Eq. (83) simply states that $\beta > \frac{1}{2}$. Our goal once more is to prove existence of the improper product integral $\hat{\Pi}_1$ of (75). We remark that writing $W(s) = V(s) + l(l+1)/s^2$ in the expression (73) for $\hat{B}_1(s)$, the term involving $l(l+1)/s^2$ belongs to $L^1(a, \infty)$, so that by the corollary to Property 10 we can prove existence and nonsingularity of $\hat{\Pi}_1$ by proving existence and nonsingularity of the improper product integral of the term in $\hat{B}_1(s)$ involving $V(s)$ alone. In other words, writing

$$C_1(s) = i\,\frac{V(s)}{2k}\begin{pmatrix} -1 & -\exp[-2ik(s-a)] \\ \exp[2ik(s-a)] & 1 \end{pmatrix}, \tag{87}$$

we wish to prove existence and nonsingularity of

$$\Pi_1 = \prod_a^{\infty} \exp[C_1(s)\,ds]. \tag{88}$$

According to Property 11, it is sufficient for this purpose to verify that the improper integral

$$H_1(r) = \int_r^{\infty} C_1(s)\,ds \tag{89}$$

exists, and that

$$H_1 C_1 \in L^1(a, \infty). \tag{90}$$

Examining (87), we see that the (1, 1) and (2, 2) entries of $H_1(r)$ involve the improper integral of V, which [by (86) with $\alpha = 1$] is of order $1/r^{\beta}$. We now consider the (2, 1) entry: We have

$$\lambda \int_r^R \exp(2iks)\frac{\sin\mu s}{s^{\beta}}\,ds$$
$$= \frac{\lambda}{2i}\int_r^R \frac{\exp(2iks)}{s^{\beta}}(\exp(i\mu s) - \exp(-i\mu s))\,ds. \tag{91}$$

Thus we must examine integrals of the form

$$I_{\pm}(r, R) = \int_r^R \frac{\exp[i(2k \pm \mu)s]}{s^{\beta}}\,ds. \tag{92}$$

This is the point at which the value $k = \mu/2$ is seen to be exceptional. If $k = \mu/2$, then $I_+(r, R)$ will converge as $R \to \infty$ (this will be apparent shortly) while if $\beta \leqslant 1$ the integral $I_-(r, R)$ obviously will *not* converge as $R \to \infty$. Thus the limit as $R \to \infty$ of the integral in (91) will not exist. (The reader can convince himself that this difficulty would not arise for $\alpha \neq 1$.) We now proceed under the assumption that $k \neq \mu/2$. Then, integrating by parts, we find

$$I_{\pm}(r, R) = \frac{\exp[i(2k \pm \mu)s]}{i(2k \pm \mu)s^{\beta}}\bigg|_r^R + \frac{\beta}{i(2k \pm \mu)}$$
$$\times \int_r^R \frac{\exp[i(2k \pm \mu)s]}{s^{\beta+1}}\,ds. \tag{93}$$

From (93) it is clear that the limits

$$I_{\pm}(r) = \lim_{R \to \infty} I_{\pm}(r, R) \tag{94}$$

exist, and an elementary estimate of the integral in (93) yields

$$I_{\pm}(r) = O(1/r^{\beta}), \tag{95}$$

Equation (95) shows that the (2, 1) entry of $H_1(r)$ is $O(1/r^{\beta})$. Estimating the (1, 2) entry in the same way and combining the above results, we obtain

$$\|H_1(r)\| = O(1/r^{\beta}). \tag{96}$$

Since clearly

$$\|C_1(r)\| = O(1/r^{\beta}) \tag{97}$$

and $\beta > \frac{1}{2}$, we have $H_1 C_1 \in L^1(a, \infty)$ and we are done.

Corollary: Suppose that

$$V(r) = \lambda(\sin\mu r^{\alpha})/r^{\beta} + V_1(r), \tag{98}$$

where μ, α, β are as in (83) and $V_1 \in L^1(a, \infty)$. Then the conclusion of Theorem 2 holds.

Proof: This is a simple application of the corollary of Property 10.

Theorems 1 and 2 can be used to prove the nonexistence of positive energy bound states for potentials which satisfy their hypotheses and which are also Kato potentials. (For other references on positive energy bounds states, see Refs. 5, 6 and references therein.) If the Hamiltonian

$$H = -\hbar^2\Delta/2m + V \tag{99}$$

had a positive energy bound state Ψ, then there would be a bound state Ψ_l with the same energy and definite angular momentum l. It is necessary to make a brief argument to pass from the statement that Ψ_l is an eigenfunction of the abstract operator H to the statement that the corresponding function $\varphi_l = \binom{\chi_l}{\chi_l'}$ is given by the product integral (56), but this can be done (for Kato potentials satisfying the hypotheses of Theorems 1 or 2) using the detailed description of the domain of H and the theory of integral equations. Taking this result for granted, if Ψ_l is a positive energy bound state, then $\chi_l(r) = r\Psi_l(r)$ has for large r the asymptotic form given by (68), and from this it is clear that $\Psi_l(r)$ cannot be square integrable over $\mathbb{R}^3$, a contradiction. Thus we can conclude the absence of positive-energy bound states for the potentials of Theorems 1 and 2 except for the exceptional case $\alpha = 1$, $k = \mu/2$ in Theorem 2, for which the asymptotic form was not established.

Using the corollary to Theorem 2, we can say the following: If V has the form (114) with $\alpha = 1$ and $V_1 \in L^1$, then V cannot have a positive-energy bound state except possibly for $k = \mu/2$ or $E = k^2 = (\mu/2)^2$. Now the Wigner–von Neumann potential[1] mentioned earlier is a Kato potential and has the asymptotic form[5]

$$V(r) = -8(\sin 2r)/r + O(1/r^2), \quad r \to \infty. \tag{100}$$

This potential is known to have a bound state at $E = 1$. Clearly the potential also has the form (98), with $V_1 \in L^1$, $\alpha = 1$. Since $\mu = 2$ for the potential (100), this example shows that it is actually possible for a bound state to occur at $E = (\mu/2)^2$. Our analysis above further shows that the energy $E = 1$ is the *only* positive energy for which the Wigner–von Neumann potential has a bound state. This improves Simon's result[5] that this potential has no bound states for $E > 16$.

We now study cases in which an "anomalous" behavior of the wavefunction arises. We return to equation (63) for $B(s)$. As illustrated in Theorem 2, the portion of $B(s)$ containing the factor $l(l+1)/s^2$ does not affect the character of the asymptotic behavior of $\Pi_a^r \exp[B(s)\,ds]$.

In the present discussion we set $l=0$ for simplicity. (All the results to be derived hold equally for $l\neq 0$.) Writing $V(s)$ instead of $W(s)$ in (63), we now make a different decomposition of B, namely

$$B(s)=C_1(s)+C_2(s) \tag{101}$$

with

$$C_1(s)=i\left(k-\frac{V(s)}{2k}\right)\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix} \tag{102}$$

and

$$C_2(s)=\frac{iV(s)}{2k}\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}. \tag{103}$$

As before, we use the sum rule to find

$$\prod_a^r \exp[B(s)\,ds]=\prod_a^r \exp[\{C_1(s)+C_2(s)\}ds]$$
$$=Q_1(r)\prod_a^r \exp[\hat{C}_2(s)\,ds] \tag{104}$$

with

$$Q_1(r)=\prod_a^r \exp[C_1(s)\,ds] \tag{105}$$

and

$$\hat{C}_2(s)=Q_1^{-1}(s)C_2(s)Q_1(s). \tag{106}$$

Now the family $\{C_1(s)\,|\,s\in[a,\infty)\}$ is commutative. Hence, setting

$$\theta(k,r)=\int_a^r\left(k-\frac{V(s)}{2k}\right)ds=k(r-a)-\frac{1}{2k}\int_a^r V(s)\,ds, \tag{107}$$

we have

$$Q_1(r)=\exp\left(\textstyle\int_a^r C_1(s)\,ds\right)=\begin{pmatrix}\exp[i\theta(k,r)] & 0\\ 0 & \exp[-i\theta(k,r)]\end{pmatrix}. \tag{108}$$

We then find

$$\hat{C}_2(s)=\frac{iV(s)}{2k}\begin{pmatrix}0 & -\exp[-2i\theta(k,s)]\\ \exp[2i\theta(k,s)] & 0\end{pmatrix}. \tag{109}$$

If it is possible to prove that the improper product integral

$$\hat{\Pi}_2=\prod_a^\infty \exp[\hat{C}_2(s)\,ds] \tag{110}$$

exists, then (104) will yield

$$\prod_a^r \exp[B(s)\,ds]\approx Q_1(r)\hat{\Pi}_2=\begin{pmatrix}\exp[ik(r-a)-(i/2k)\int_a^r V(s)\,ds] & 0\\ 0 & \exp[-ik(r-a)+(i/2k)\int_a^r V(s)\,ds]\end{pmatrix}\hat{\Pi}_2. \tag{111}$$

Let $\varphi(r)$ be as in (56). If the integral $\int_a^r V(s)\,ds$ does not converge as $r\to\infty$, then (111) implies an asymptotic form for $\varphi(r)$ consisting of the "usual" asymptotic form modified by the "anomalous" factors $\exp[\pm(i/2k)\int_a^r \times V(s)\,ds]$. This is most familiar for the Coulomb potential (81), for which the anomalous factors have the form $\exp[\pm(i\lambda/2k)\log(r/a)]$. (Of course, if $\lim_{r\to\infty}\int_a^r V(s) \times ds$ *does* exist, (111) implies the usual asymptotic form.)

Existence and nonsingularity of the improper product integral $\hat{\Pi}_2$ can be proved under suitable hypotheses on V using Property 11. Setting

$$\hat{H}_2(r)=\mathrm{imp}\int_r^\infty \hat{C}_2(s)\,ds \tag{112}$$

the problem is to show that $\hat{H}_2$ is defined and $\hat{H}_2\hat{C}_2 \in L^1(a,\infty)$. As an illustration we prove

Theorem 3: Let $V(r)=\lambda/r^\beta$ with λ real, $\beta>\frac{1}{2}$. Let $\varphi(r)$ be as in (56) with $\varphi(a)\neq 0$. Then

$$\varphi(r)\approx C_+\exp[ikr-(i/2k)\int_a^r V(s)\,ds]$$
$$+C_-\exp[-ikr+(i/2k)\int_a^r V(s)\,ds], \tag{113}$$

where C_+ and C_- are constant 2×1 matrices, not both of whose (1, 1) entries are zero.

Proof: We need only prove existence and nonsingularity of $\hat{\Pi}_2$. (See the prooof of Theorem 1.) We estimate a matrix element of $\hat{H}_2(r)$:

$$\int_r^R \frac{\exp[2i\theta(k,s)]}{s^\beta}ds=\left.\frac{\exp[2i\theta(k,s)]}{2is^\beta\theta'(k,s)}\right|_r^R$$
$$-\frac{1}{2i}\int_r^R \exp[2i\theta(k,s)]$$
$$\times\frac{d}{ds}\left(\frac{1}{s^\beta\theta'(k,s)}\right)ds. \tag{114}$$

Now

$$s^\beta\theta'(k,s)=s^\beta(k-1/2ks^\beta)=ks^\beta-1/2k. \tag{115}$$

Using (115), one easily sees that the limit as $R\to\infty$ can be taken in (114) and that the resulting quantity is $O(1/r^\beta)$. Arguing similarly for the other nonzero matrix element of $\hat{H}_2(r)$, we have

$$\|\hat{H}_2(r)\|=O(1/r^\beta). \tag{116}$$

Since $\hat{C}_2(r)$ is also $O(1/r^\beta)$ and $\beta>\frac{1}{2}$, we have $\hat{H}_2\hat{C}_2 \in L^1(a,\infty)$, finishing the proof.

Theorem 3 excludes from consideration potentials of the type

$$V(r)=\lambda/r^\beta,\quad \lambda \text{ real},\quad 0<\beta\leq\tfrac{1}{2}. \tag{117}$$

For such potentials and a variety of others, we can show that the wave-functions have the asymptotic form suggested by "WKB intuition." The following theorem covers any potential with a continuous derivative integrable over (b_0,∞) for some $b_0>0$, hence in particular any potential of the type (117). The proof of this theorem is a bit different from the proofs above. For this reason we shall prove the theorem with $l\neq 0$ to convince the reader that again the term $l(l+1)/r^2$ makes no difference.

Theorem 4: Suppose that V is continuous on $(0, \infty)$, V is continuously differentiable on $[b_0, \infty)$ for some $b_0 > 0$, and that

$$\int_{b_0}^{\infty} |V'(s)|\, ds < \infty. \tag{118}$$

Let

$$V_0 = \lim_{r\to\infty} V(r) \tag{119}$$

[this limit exists by (118)]. Let $\varphi(r)$ be any solution of (54) which is not identically zero. Then for $E > V_0$ we have

$$\varphi(r) \approx C_+ \exp[i \int_{b_0}^{r} \sqrt{E - W(s)}\, ds] + C_- \exp[-i \int_{b_0}^{r} \sqrt{E - W(s)}\, ds], \tag{120}$$

where $C_\pm$ are constant 2×1 matrices, not both of whose (1.1) entries are zero, and W is given by (51).

Proof: Replacing V by $V - V_0$ we may assume $V_0 = 0$, and we must then prove (120) for $E > 0$. With $A(r)$ as in (55), the solution of (54) is given by

$$\varphi(r) = \prod_{b}^{r} \exp[A(s)\, ds]\, \varphi(b), \tag{121}$$

where b is any number in $(0, \infty)$. The function φ is not indentically zero if and only if $\varphi(b)$ is nonzero. Since $E > 0$ and $W(r) \to 0$ as $r \to \infty$, we may assume that $b \geq b_0$ is large enough so that

$$E - W(r) \geq \alpha > 0 \quad \text{for } r \in [b, \infty). \tag{122}$$

Thus $[E - W(r)]^{-1}$ is bounded on $[b, \infty)$. For $r \in [b, \infty)$ we define

$$\lambda(r) = \sqrt{E - W(r)} \quad \text{(positive square root)}. \tag{123}$$

We now study the representation (121) for φ. This time, instead of making the decomposition (57) of $A(r)$, we diagonalize $A(r)$ directly. The characteristic equation for $A(r)$ yields the eigenvalues $\pm i\lambda(r)$, where $\lambda(r)$ is given by (123). Setting

$$\Lambda(r) = \begin{pmatrix} i\lambda(r) & 0 \\ 0 & -i\lambda(r) \end{pmatrix} \tag{124}$$

and

$$M(r) = \begin{pmatrix} 1 & 1 \\ i\lambda(r) & -i\lambda(r) \end{pmatrix}, \tag{125}$$

we have

$$A(r) = M(r)\Lambda(r)M^{-1}(r). \tag{126}$$

The similarity rule (applied with $T = M$) now yields

$$\prod_{b}^{r} \exp[A(s)\, ds] = M(r)H(r)M^{-1}(b) \tag{127}$$

with

$$H(r) = \prod_{b}^{r} \exp[\{M^{-1}(s)A(s)M(s) - M^{-1}(s)M'(s)\}\, ds] = \prod_{b}^{r} \exp[\{\Lambda(s) - M^{-1}(s)M'(s)\}\, ds]. \tag{128}$$

Equation (127) should be compared with Eq. (65), in which M does not depend on r. Examining (127), we note that $M(r)$ converges as $r \to \infty$ to the constant nonsingular matrix M of (61) [since $W(r) \to 0$ as as $r \to \infty$]. Thus, to analyze the asymptotic behavior of the right-hand side of (127), we need only study the behavior of $H(r)$. Let us write

$$L(r) = \prod_{b}^{r} \exp[\Lambda(s)\, ds]. \tag{129}$$

Since $\{\Lambda(r) \mid r \in [b, \infty)\}$ is commutative, we have

$$L(r) = \exp[\int_b^r \Lambda(s)\, ds] = \begin{pmatrix} \exp[i\sigma(r)] & 0 \\ 0 & \exp[-i\sigma(r)] \end{pmatrix}, \tag{130}$$

where

$$\sigma(r) = \int_b^r \lambda(s)\, ds = \int_b^r \sqrt{E - W(s)}\, ds. \tag{131}$$

Note that $L(r)$ contains exactly the exponentials we need in (120). Using (128) and the sum rule, we find

$$H(r) = L(r) \prod_{b}^{r} \exp[N(s)\, ds] \tag{132}$$

with

$$N(s) = -L^{-1}(s)M^{-1}(s)M'(s)L(s). \tag{133}$$

If it can be shown that $\prod_b^r \exp[N(s)\, ds]$ converges as $r \to \infty$ to a nonsingular limit $\hat{\Pi}_+$, then combining our previous remarks we will have

$$\prod_{a}^{r} \exp[A(s)\, ds] \approx ML(r)\hat{\Pi}_+ M^{-1}(b), \quad r \to \infty \tag{134}$$

and because of the form of $L(r)$, this will prove (120). Convergence of $\prod_b^r \exp[N(s)\, ds]$ will follow if we can prove $N \in L^1(b, \infty)$. Since $L(r)$ is unitary, we have

$$\|N(s)\| = \|M^{-1}(s)M'(s)\|. \tag{135}$$

A brief calculation yields

$$M^{-1}(s)M'(s) = \frac{\lambda'(s)}{2\lambda(s)} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = -\frac{1}{4} \frac{W'(s)}{E - W(s)} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{136}$$

Then, since the matrix on the right-hand side of (136) has norm 2, we find for $s \geq b$

$$\|M^{-1}(s)M'(s)\| = \frac{1}{2} \frac{|W'(s)|}{|E - W(s)|} \leq \frac{1}{2\alpha} |W'(s)|, \tag{137}$$

where we have used (122). Now

$$W'(s) = V'(s) - 2l(l+1)/s^3 \tag{138}$$

so that $|W'(s)|$ is integrable over $[b, \infty)$. Thus $N \in L^1(b, \infty)$, and we are done.

Theorems 3 and 4 can be used in the same way as Theorems 1 and 2, to deduce the nonexistence of bound states with positive energy or energy greater than V_0. We shall not elaborate on this fact.

III. ASYMPTOTICS OF SOLUTIONS OF THE SCHRÖDINGER EQUATION FOR LARGE E

In this section we shall indicate briefly how to obtain the asymptotic form of solutions of the Schrödinger equation for large positive values of the energy E. The results are those suggested by the WKB method (or quasiclassical approximation) and may be derived without product integrals; however, the use of product integration yields the results very quickly and provides explicit error estimates.

For simplicity we consider the one-dimensional Schrödinger equation

$$-\frac{d^2y}{dx^2}+V(x)y=Ey \tag{139}$$

on some interval of the real numbers, $\mathbf{R}$. We assume $V(x)$ is real-valued and continuously differentiable and that $E-V(x)>0$. We write (139) as a system:

$$\begin{pmatrix} y(x) \\ y'(x) \end{pmatrix}'=\begin{pmatrix} 0 & 1 \\ V(x)-E & 0 \end{pmatrix}\begin{pmatrix} y(x) \\ y'(x) \end{pmatrix}. \tag{140}$$

The solution is given by the product integral

$$\begin{pmatrix} y(x) \\ y'(x) \end{pmatrix}=\prod_{x_0}^{x}\exp[A_1(s)\,ds]\cdot\begin{pmatrix} y(x_0) \\ y'(x_0) \end{pmatrix},$$

$$A_1(s)=\begin{pmatrix} 0 & 1 \\ V(s)-E & 0 \end{pmatrix}. \tag{141}$$

The method of proof of Theorem 4 of the previous section shows that this last equation can be written as

$$\begin{aligned}\begin{pmatrix} y(x) \\ y'(x) \end{pmatrix}=&\begin{pmatrix} 1 & 1 \\ ik(x) & -ik(x) \end{pmatrix}\\ &\times\begin{pmatrix} \exp[iK(x)] & 0 \\ 0 & \exp[-iK(x)] \end{pmatrix}\prod_{x_0}^{x}\exp[A_2(s)\,ds]\\ &\times\begin{pmatrix} \frac{1}{2} & 1/2ik(x_0) \\ \frac{1}{2} & -1/2ik(x_0) \end{pmatrix}\begin{pmatrix} y(x_0) \\ y'(x_0) \end{pmatrix},\end{aligned} \tag{142}$$

where $k(x)=\sqrt{E-V(x)}$ is the positive square root,

$$K(x)=\int_{x_0}^{x}k(s)\,ds,$$

$$A_2(s)=\frac{V'(s)}{4(V(s)-E)}\begin{pmatrix} -1 & \exp[-2iK(s)] \\ \exp[2iK(s)] & -1 \end{pmatrix}.$$

From the estimate (23) of Sec. I applied to $\Pi_{x_0}^{x}\times\exp[A_2(s)\,ds]$ together with the fact that

$$\left\|\begin{matrix} 1 & \exp(i\theta) \\ \exp(-i\theta) & 1 \end{matrix}\right\|=2 \quad\text{for real } \theta,$$

we have

$$\prod_{x_0}^{x}\exp[A_2(s)\,ds]=I+R(x,x_0,E)$$

$$\text{where } \|R(x,x_0,E)\|\leqslant\exp\left(\frac{1}{2}\int_{x_0}^{x}\frac{|V'(s)|\,ds}{E-V(s)}\right)-1. \tag{143}$$

For example, in any region where V is uniformly bounded and V' is integrable, $\|R(x,x_0,E)\|=O(1/E)$ uniformly in x. If we denote a solution of (139) by $y(x,E)$, then (142), (143) give after some matrix multiplications

$$\begin{aligned}y(x,E)=&\,y(x_0)\cos\left[\int_{x_0}^{x}\sqrt{E-V(s)}\,ds\right]+[y'(x_0)/\sqrt{E-V(x_0)}]\\ &\times\sin\left[\int_{x_0}^{x}\sqrt{E-V(s)}\,ds\right]+r(x,x_0,E)\end{aligned} \tag{144}$$

with

$$|r(x,x_0,E)|\leqslant\left(|y(x_0)|+\frac{|y'(x_0)|}{\sqrt{E-V(x_0)}}\right)\times\left[\exp\left(\tfrac{1}{2}\int_{x_0}^{x}\frac{(V'(s))\,ds}{E-V(s)}\right)-1\right].$$

We note that in any region where V is uniformly bounded,

$$\int_{x_0}^{x}\sqrt{E-V(s)}\,ds=\sqrt{E}\int_{x_0}^{x}\sqrt{1-V(s)/E}\,ds$$

$$=\sqrt{E}\,(x-x_0)+O(1/\sqrt{E})$$

uniformly in x. Hence, if V is uniformly bounded and V' integrable, we have

$$\begin{aligned}y(x,E)=&\,y(x_0)\cos[\sqrt{E}\,(x-x_0)]+[y'(x_0)/\sqrt{E}\,]\\ &\times\sin[\sqrt{E}\,(x-x_0)]+O(1/\sqrt{E})\end{aligned} \tag{145}$$

uniformly in x. Equations (144) and (145) constitute the WKB approximation giving the wavefunction asymptotically for large $E>0$.

Formulas similar to (144) may be derived in a similar manner for many other examples of equations or systems of equations involving a parameter.

ACKNOWLEDGMENT

We would like to thank Barry Simon for an interesting discussion.

[1] J. von Neumann and E. P. Wigner, Z. Physik 30, 465—67 (1929). For a correction of the algebraic errors in this paper, see Ref. 5.

[2] J.D. Dollard and C.N. Friedman, "On Strong Product Integration" (submitted to J. Funct. Anal.).

[3] V. Volterra and B. Hostinsky, *Operations Infinitesimales Lineaires* (Gauthier-Villars, Paris, 1938).

[4] T. Kato, Trans. Am. Math. Soc. 70, 195—211 (1951).

[5] B. Simon, Comm. Pure Appl. Math. 22, 531—38 (1969).

[6] J. Weidmann, Math. Z. 98, 268—302 (1967).

[7] J. D. Dollard and C. N. Friedman, monograph on product integration (Addison-Wesley, Reading, Mass.), to appear.

On the theory of time-dependent linear canonical transformations as applied to Hamiltonians of the harmonic oscillator type

P. G. L. Leach

Departments of Mathematics, La Trobe University, Bundoora, 3083, Australia

(Received 2 August 1976; revised manuscript received 1 November 1976)

Writing the canonical variables ($\mathbf{q}^T$, $\mathbf{p}^T$) as (ω^T), we develop a method for transforming the time-dependent Hamiltonian $H = A_{\mu\nu}(t)\omega^\mu\omega^\nu + B_\mu\omega^\nu + C(t)$ to the time-independent form $\overline{H} = (1/2)\delta_{\mu\nu}\overline{\omega}^\mu\overline{\omega}^\nu$ using the linear transformation $\overline{\omega}^\mu = s^\mu{}_\nu(t)\omega^\nu + r^\mu(t)$. Differential equations are obtained for the parameters $s^\mu{}_\nu$ and r^μ. The transformed Hamiltonian enables the construction of an invariant I and an invariant matrix $[I^{\mu\nu}]$. These invariants apply to both the classical and quantum mechanical problems. The invariant I has the dynamical symmetry group SU(n), and this characterizes all systems with Hamiltonians of the form of H.

1. INTRODUCTION

Time-dependent canonical transformations have not attracted much detailed attention in the past. Even in the better treatises on classical mechanics, the explicit implications of time dependence are not explored. Although Whittaker[1] allows for the presence of time in his theory, not one of his many examples involves a time-dependent transformation. One of the best of the newer texts, that by Sudarshan and Makunda[2] develops conditions which all canonical transformations must satisfy, but does not provide particular examples. Pars[3] does remark that a two-dimensional linear transformation with time-dependent coefficients will be canonical under certain constraints.

Linear canonical transformations have been used in quantum mechanics by Moshinsky and others[4,5] for problems concerning the harmonic oscillator. More recently, Günther and Leach,[6] in their discussion of invariants for time-dependent oscillators, made use of a simple time-dependent linear canonical transformation to provide an interpretation of the Lewis invariant.[7-10] They did not provide a general discussion of such transformation. A subsequent paper by Leach[11] provides a partial discussion of the theory in the course of showing that the Lewis invariant is a particular case of a more general invariant.

In this paper we develop the theory of time-dependent linear canonical transformations as it impinges upon Hamiltonians of harmonic oscillator type. We show how we may find invariants for time-dependent Hamiltonians of (oscillator type) using the transformations and give explicit formulas for the determination of their coefficients. These results apply equally well to classical and quantum mechanics.

The existence of such invariants has a twofold application. As Lewis and Riesenfeld[9] have shown, it leads to the solution of the time-dependent Schrödinger equation. From these invariants, the generators of dynamical symmetry groups may be obtained. Fradkin[12,13] has demonstrated this for the three-dimensional time-independent isotropic harmonic oscillator while Günther and Leach[6] have extended his work to the corresponding time-dependent problem. We shall show that the Lie group SU(n) characterizes most n-dimensional harmonic oscillator-type Hamiltonians, usually as a noninvariance group. We hesitate to call it a characteristic noninvariance symmetry group in view of the use of that term in a different context by Mukunda *et al.*[14] It is possible that this result may be of advantage in the discussion of the evolution of coherent states for oscillator systems.[15,16] Finally we present a particularly simple expression from which the generators of the symmetry group may be obtained.

2. CONDITIONS FOR CANONICITY

The theory which we develop is rendered more simply in the formalism which is becoming increasingly popular. We write the canonical variables (q_i, p_i) as

$$\begin{aligned} q_i &= \omega^\mu, \quad i=1,n, \quad \mu=1,n, \\ p_i &= \omega^\mu, \quad i=1,n, \quad \mu=n+1,2n \end{aligned} \tag{2.1}$$

so that Hamilton's equations of motion are

$$\overset{\circ}{\omega}{}^\mu = \epsilon^{\mu\nu}\frac{\partial H}{\partial \omega^\nu}, \tag{2.2}$$

in which $[\epsilon^{\mu\nu}]$ is the symplectic matrix of order $2n$. Its inverse is written as $[\epsilon_{\mu\nu}]$. Thus

$$[\epsilon^{\mu\nu}] = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}, \quad [\epsilon_{\mu\nu}] = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}. \tag{2.3}$$

The general linear transformation from coordinates ω^μ to $\overline{\omega}^\mu$ is

$$\overline{\omega}^\mu = s^\mu{}_\nu\omega^\nu + r^\mu, \tag{2.4}$$

in which the parameters are real and may be time-dependent. We take the transformation to be measure-preserving. Thus the determinant of the coefficient matrix $[\mathbf{s}^\mu{}_\nu]$ is unity and the inverse exists. The transformation is canonical provided the Poisson bracket condition

$$[\overline{\omega}^\mu, \overline{\omega}^\nu]_{\mathrm{PB}\omega} = \epsilon^{\mu\nu} \longleftrightarrow \epsilon^{\alpha\beta}s^\mu{}_\alpha s^\nu{}_\beta = \epsilon^{\mu\nu}, \tag{2.5}$$

is satisfied. Writing

$$[s^\nu{}_\mu] = S, \quad [\epsilon^{\mu\nu}] = \epsilon, \tag{2.6}$$

in matrix notation (2.5) is

$$S\epsilon S^T = \epsilon, \tag{2.7}$$

i. e., the transformation matrix S is a member of a $2n$-dimensional real symplectic group. Writing S in block form

$$S=\begin{pmatrix} S_1 & S_2 \\ S_3 & S_4 \end{pmatrix}, \tag{2.8}$$

(2.8) indicates that

$$\begin{aligned} S_1S_2{}^T &= S_2S_1{}^T, \\ S_3S_4{}^T &= S_4S_3{}^T, \\ S_1S_4{}^T - S_2S_3{}^T &= I_n, \end{aligned} \tag{2.9}$$

which is exactly the same set of conditions[4] as for time-independent transformations.

The transformation (2.4) may be canonical without being symplectic since the latter requires the biquadratic form $\omega^\mu\epsilon_{\mu\nu}\omega'^\nu$ to be invariant under the transformation. In the new coordinate system, the biquadratic form is

$$\begin{aligned}\bar\omega^\nu\epsilon_{\mu\nu}\bar\omega'^\nu = {} & \omega^\sigma s^\mu{}_\sigma\epsilon_{\mu\nu}s^\nu{}_\rho\omega'^\rho + r^\mu\epsilon_{\mu\nu}r^\nu + s^\mu{}_\sigma\epsilon_{\mu\nu}r^\nu\omega^\sigma \\ & + s^\nu{}_\rho\epsilon_{\mu\nu}r^\mu\omega'^\rho. \end{aligned} \tag{2.10}$$

Since the coordinate-free term is identically zero, (2.10) is the same as $\omega^\mu\epsilon_{\mu\nu}\omega'^\nu$ provided

$$s^\mu{}_\sigma\epsilon_{\mu\nu}s^\nu{}_\rho = \epsilon_{\sigma\rho}, \tag{2.11}$$

$$\begin{aligned} & s^\mu{}_\sigma\epsilon_{\mu\nu}r^\nu = 0, \\ & s^\nu{}_\rho\epsilon_{\mu\nu}r^\mu = 0, \\ & \longleftrightarrow S\epsilon S^T = \epsilon, \;\; S^T\epsilon^T\mathbf{r} = 0. \end{aligned} \tag{2.12}$$

The first of (2.12) is just (2.7). Since the determinants of S and ϵ are unity, the second of (2.12) requires $\mathbf{r}$ to be zero. Thus a nonhomogeneous linear transformation cannot be symplectic. However, it is canonical provided the coefficient matrix S is symplectic.

3. THE TRANSFORMATION MATRICES

Under a linear transformation from canonical coordinates ω^μ to $\bar\omega^\mu$, the general oscillator-type Hamiltonian

$$H = A_{\mu\nu}(t)\omega^\mu\omega^\nu + B_\mu(t)\omega^\mu + C(t), \tag{3.1}$$

is transformed to

$$\bar H = \bar A_{\mu\nu}(t)\,\bar\omega^\mu\bar\omega^\nu + \bar B_\mu(t)\bar\omega^\mu + \bar C(t), \tag{3.2}$$

where the matrices $A_{\mu\nu}(t)$ and $\bar A_{\mu\nu}(t)$ are real, symmetric, and of the same rank. All coefficients are at least continuous over the interval of time which is of interest. The time evolution of $\bar\omega^\mu$ may be studied through either

$$\dot{\bar\omega}^\mu = \epsilon^{\mu\nu}\frac{\partial\bar H}{\partial\bar\omega^\nu}, \tag{3.3}$$

or

$$\dot{\bar\omega}^\mu = \frac{d}{dt}(s^\mu{}_\nu\omega^\nu + r^\nu). \tag{3.4}$$

Equating (3.3) to (3.4), substituting for H, $\bar H$, and $\bar\omega^\mu$, and separating coordinate-free and coordinate-dependent parts, we obtain for the $(2n)^2$ $s^\mu{}_\nu$ $(2n)^2$ first order linear differential equations

$$\mathring{s}^\mu{}_\nu = 2\epsilon^{\mu\sigma}\bar A_{\sigma\rho}s^\rho{}_\nu - 2s^\mu{}_\sigma\epsilon^{\sigma\rho}A_{\rho\nu}, \tag{3.5}$$

and, for the $2n$ r^μ, $2n$ equations

$$\mathring{r}^\mu - 2\epsilon^{\mu\nu}\bar A_{\nu\sigma}r^\sigma = \epsilon^{\mu\nu}\bar B_\nu - s^\mu{}_\nu\epsilon^{\mu\sigma}B_\sigma. \tag{3.6}$$

Following Ince,[17] provided the coefficients and independent terms are continuous over some closed interval $[t_1, t_2]$, the sets of equations (3.5) and (3.6) will have solution sets which are continuous and unique in the interval (t_1, t_2) and which assume given values for some $t_0 \in (t_1, t_2)$. Since these given values are left for us to assign, we shall use them, in the case of S, to ensure that the requirements of (2.9) and det.S being unity are satisfied. For $\mathbf{r}$ the choice is a matter of convenience rather than requirement.

We observe that the functions $C(t)$ and $\bar C(t)$ do not occur in (3.5) and (3.6). This reflects the fact that Hamilton's equations are not affected by the addition of an arbitrary function of time to the Hamiltonian. In particular, $\bar C(t)$ may be set at zero in $\bar{\boldsymbol{H}}$. As simplicity of form is a hand-maiden of successful analysis, we take the simplest form possible for $\bar H$ by setting

$$\bar A_{\mu\nu}(t) = \tfrac{1}{2}\delta_{\mu\nu}, \quad \bar B_\mu(t) = 0, \quad \bar C(t) = 0. \tag{3.7}$$

Equations (3.5) and (3.6) become

$$\dot{s}^\mu{}_\nu = \epsilon^{\mu\sigma}\delta_{\sigma\rho}s^\rho{}_\nu - 2s^\mu{}_\sigma\epsilon^{\sigma\rho}A_{\rho\nu}, \tag{3.8}$$

$$\mathring{r}^\mu - \epsilon^{\mu\nu}\delta_{\nu\sigma}r^\sigma = -s^\mu{}_\nu\epsilon^{\nu\sigma}B_\sigma. \tag{3.9}$$

If H contains no terms linear in ω^μ, (3.9) has the simple solution set

$$r^\mu = 0, \quad \mu = 1, 2n. \tag{3.10}$$

If the B_μ are not all zero, the solution set of (3.9) depends upon that of (3.8) due to the presence of the $s^\mu{}_\nu$ in (3.9).

4. INVARIANTS

We have seen that a transformation

$$\bar\omega^\mu = s^\mu{}_\nu\omega^\nu + r^\mu, \tag{4.1}$$

exists which transforms the time-dependent Hamiltonian

$$H = A_{\mu\nu}(t)\omega^\mu\omega^\nu + B_\mu(t)\omega^\mu + C(t), \tag{4.2}$$

to a time-independent form

$$\bar H = \tfrac{1}{2}\bar\omega^\mu\delta_{\mu\nu}\bar\omega^\nu. \tag{4.3}$$

In the coordinate system $\{\bar\omega^\mu\}$, $\bar H$ is an invariant. In terms of the original coordinate system $\{\omega^\mu\}$,

$$\bar H = \tfrac{1}{2}(s^\mu{}_\sigma\omega^\sigma + r^\mu)\delta_{\mu\nu}(s^\nu{}_\rho\omega^\rho + r^\nu). \tag{4.4}$$

Direct calculation shows that

$$\dot{\bar H} = [\bar H, H]_{\mathrm{PB}_\omega} + \frac{\partial\bar H}{\partial t} = 0, \tag{4.5}$$

when (3.8) and (3.9) are taken into account. Thus (4.4) is an exact invariant of the motion described by H. In matrix form, this invariant, which is now written as I, is

$$I = \tfrac{1}{2}(S\omega + \mathbf{r})^T(S\omega + \mathbf{r}), \tag{4.6}$$

where S and r are solutions of the linear differential equations

$$\overset{\circ}{S} = \epsilon S - 2S\epsilon A, \tag{4.7}$$

$$\overset{\circ}{\mathbf{r}} = \epsilon\mathbf{r} - S\epsilon\mathbf{B}, \tag{4.8}$$

and A and $\mathbf{B}$ are the matrix coefficients in H.

The existence of the scalar invariant I suggests the construction of a symmetric invariant $2n \times 2n$ matrix

$$[I^{\mu\nu}] = \tfrac{1}{2}\overline{\omega}\,\overline{\omega}^T + \tfrac{1}{2}(\epsilon\overline{\omega})(\epsilon\overline{\omega})^T, \tag{4.9}$$

$$= \tfrac{1}{2}(S\omega + \mathbf{r})(S\omega + \mathbf{r})^T + \tfrac{1}{2}\epsilon(S\omega + \mathbf{r})(S\omega + \mathbf{r})^T\epsilon^T. \tag{4.10}$$

The invariance of $[I^{\mu\nu}]$ in terms of the original coordinates is easily checked directly using (4.7), (4.8) and

$$\dot{\omega} = \epsilon(2A\omega + \mathbf{B}). \tag{4.11}$$

We note that the invariant I is half the trace of $[I^{\mu\nu}]$. Disregarding the significance of upper and lower suffices, the matrix $[I^{\mu\nu}]$ may be written in block form as

$$[I^{\mu\nu}] = \begin{bmatrix} [I_{ij}] & [L_{ij}] \\ -[L_{ij}] & [I_{ij}] \end{bmatrix}, \tag{4.12}$$

in which

$$I_{ij} = \tfrac{1}{2}(Q_iQ_j + P_iP_j), \qquad L_{ij} = \tfrac{1}{2}(Q_iP_j - P_iQ_j), \tag{4.13}$$

where

$$\overline{\omega}^T = (\mathbf{Q}^T, \mathbf{P}^T). \tag{4.14}$$

It is obvious that we may identify I_{ij} with Fradkin's invariant tensor[12] and L_{ij} with the angular momentum tensor for the coordinate system $\{\overline{\omega}^\mu\}$.

There are $n^2 - n$ distinct off-diagonal terms in $[I^{\mu\nu}]$ and $n - 1$ linear combinations of the diagonal terms which are linearly independent of I. Each element of $[I^{\mu\nu}]$ has zero Poisson bracket with I. Thus there are $n^2 - 1$ separate invariants. As has already been described adequately in the literature (e.g., Refs. 6 and 12), linearly independent combinations of these invariants provide a suitable basis for the representation of the Lie algebra SU(n). Hence the invariant I possesses the dynamical symmetry group SU(n). As $[I, H]_{\mathrm{PB}}$ is nonzero for oscillators other than the n-dimensional, time-independent harmonic oscillator, SU(n) is generally a noninvariance dynamical symmetry group for H.

We conclude that Hamiltonians of the type

$$H = A_{\mu\nu}(t)\omega^\mu\omega^\nu + B_\mu(t)\omega^\mu + C(t), \tag{4.2}$$

are characterized by the dynamical symmetry group SU(n). Following Mukunda *et al.*,[14] the invariant I will also possess a characteristic noninvariance dynamical symmetry group which is SU($n+1$) (compact) or SU(n, 1) (noncompact). Consequently this same symmetry group may be associated with the Hamiltonian (4.2).

5. QUANTUM INVARIANTS

For the quantum mechanical problem corresponding to the Hamiltonian (4.2), an invariant operator and an invariant matrix operator are defined in the same way as the classical invariant and matrix except that now the ω^μ are operators which satisfy the commutation relations

$$[\omega^\mu, \omega^\nu] = i\hbar\epsilon^{\mu\nu}. \tag{5.1}$$

In both cases it may be verified readily that

$$\dot{O} = \frac{1}{i\hbar}[O, H] + \frac{\partial O}{\partial t} = 0, \tag{5.2}$$

where O is the operator I or $[I^{\mu\nu}]$. Since each element $I^{\mu\nu}$ commutes with I, we may use those elements to construct the generators of the symmetry group in the usual way (e.g., Günther and Leach[6]). The remarks about the characteristic noninvariance symmetry group in Sec. 4 also apply here.

6. CONCLUSION

The existence of a quantum mechanical invariant for oscillator-type systems does more than provide a symmetry group for such systems. As Lewis and Riesenfeld[9] have shown, the invariant enables the solution of the time-dependent Schrödinger equation. This solution may be obtained more readily using the transformations discussed here by expressing the invariant more simply in a different coordinate system. Recently Wolf[18] discussed the solution of the time-independent Schrödinger equation,

$$H\psi = \lambda\psi, \tag{6.1}$$

$$H = \alpha p^2 + \beta q^2 + \gamma(qp + pq) + \delta p + \epsilon q + \xi, \tag{6.2}$$

where the coefficients are constant, using linear canonical transformations. In general, the wavefunctions in different coordinate systems are related by an integral transform. In the special case where the transformation has the form

$$\omega = \begin{bmatrix} a & 0 \\ c & d \end{bmatrix}\omega + \mathbf{r}, \tag{6.3}$$

the relationship is a phase change with possibly a scaling term independent of the coordinates.

Wolf's results may be extended to Hamiltonians of the form considered in this paper. In a subsequent paper we shall demonstrate that the wavefunctions for all time-dependent oscillators are related by a phase and scaling term to the wavefunction for

$$H = \tfrac{1}{2}\rho^{-2}\omega^T\omega, \tag{6.4}$$

where $\rho(t)$ is any solution of a second order differential equation of the type

$$\ddot{\rho} + f(t)\overset{\circ}{\rho} + g(t)\rho = c\rho^{-3}. \tag{6.5}$$

More generally the wavefunction is related to that of the time-independent Hamiltonian

$$H = \tfrac{1}{2}\omega^T\omega, \tag{6.6}$$

only by an integral transform. The expression for the kernel is the subject of current investigation.

In this paper we have confined our attention to the theoretical aspects of linear transformations and oscillator-type Hamiltonians. A simple example has been given by Leach.[11] In the subsequent paper preferred to above we shall describe more complex examples.

Note added in proof: That the transformation matrix S is symplectic for all time provided it is symplectic at $t=t_0$ is easily shown. Suppose $S=[s^{\mu}{}_{\nu}]$ is the solution of (3.8). Then $M=SJS^T$ satisfies the equation $\dot{M}=\epsilon M-M\epsilon$. This represents a set of linear order differential equations with an equilibrium point at $M=\epsilon$. Suppose $S(t_0)$ is chosen in such a way that $M(t_0)=\epsilon$. This is possible since there are $(2n)^2$ arbitrary constants in $S(t)$. Then $M(t)=\epsilon$ and so $S(t)$ is symplectic for all t. I am indebted to Dr. W. Sarlet of the Rijksuniversiteit-Gent (private communication) for this note.

ACKNOWLEDGMENT

We thank the referee for his constructive remarks and for providing the opportunity to rewrite Sec. 6 in the light of our more recent work.

[1] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge University, Cambridge, 1937), Chap. 9.

[2] E. C. G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974), pp. 35—44.

[3] L. A. Pars, *A Treatise on Analytical Dynamics* (Heinemann, London, 1965), p. 500.

[4] M. Moshinsky and C. Quesne, J. Math. Phys. **12**, 1772 (1971).

[5] M. Moshinsky, T. H. Seligman, and K. B. Wolf, J. Math. Phys. **13**, 901 (1972).

[6] N. J. Günther and P. G. L. Leach, "Generalized Invariants for the Time-Dependent Harmonic Oscillator" (to be published in J. Math. Phys.).

[7] H. R. Lewis, Jr., Phys. Rev. Lett. **18**, 510, 636 (1967).

[8] H. R. Lewis, Jr., J. Math. Phys. **9**, 1976 (1968).

[9] H. R. Lewis, Jr., and W. B. Riesenfeld, J. Math. Phys. **10**, 1458 (1969).

[10] H. R. Lewis, Jr., Phys. Rev. **172**, 1313 (1968). Although it is not stated as such, the writer in fact uses a transformation of the type whose theory is developed here.

[11] P. G. L. Leach, "On a Generalization of the Lewis Invariant for the Time-Dependent Harmonic Oscillator," La Trobe University Departments of Mathematics preprint (unnumbered).

[12] D. M. Fradkin, Am. J. Phys. **33**, 207 (1965).

[13] D. M. Fradkin, Prog. Theoret. Phys. **37**, 798 (1967).

[14] N. Mukunda, L. O'Raifeartaigh, and E. C. G. Sudarshan, Phys. Rev. Lett. **15**, 1041 (1965).

[15] C. L. Mehta and E. C. G. Sudarshan, Phys. Lett. **28**, 574 (1964).

[16] C. L. Mehta, P. Chand, E. C. G. Sudarshan, and R. Vedam, Phys. Rev. **157**, 1198 (1967).

[17] E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956), pp. 71, 72.

[18] K. B. Wolf, J. Math. Phys. **17**, 601 (1976).

Derivation of quantum mechanics from stochastic electrodynamics

L. de la Peña-Auerbach* and A. M. Cetto

Instituto de Física, UNAM, Apartado Postal 20—364, México 20, D. F., Mexico

(Received 2 August 1976)

From the equation of motion for a radiating charged particle embedded in the zero-point radiation field we construct a stochastic Liouville equation which serves to derive, by a smoothing process, a Fokker–Planck-type equation with infinite memory. We show that an exact alternative form of this phase-space equation is the Schrödinger equation in configuration space, with radiative corrections. In the asymptotic, radiationless limit (when the radiative corrections become negligible), the phase-space density reduces to Wigner's distribution, thus confirming Weyl's rule of correspondence. We briefly discuss several other implications of stochastic electrodynamics which are relevant for quantum theory in general.

INTRODUCTION

Random[1,2] or stochastic electrodynamics[3-5] (SED) is the theory of motion of charged particles in the presence of the electromagnetic vacuum, considered as a random radiation field at a given temperature, for most purposes taken as zero. This theory has been used to predict several phenomena which are usually considered to have a strictly quantum theoretical nature, such as, e.g., the van der Waals forces between polarizable particles[6] or between macroscopic conductors,[7] or the quantal behavior of simple systems such as the harmonic oscillator, including the radiative shift of the energy levels[3,4,8,9] and the lifetime of its excited states,[10] etc.

On the conceptual side, SED offers an elementary and coherent explanation of certain basic features of quantum mechanics (QM), such as the random behavior of matter, as due to its interaction with the stochastic field, or the stability of atomic systems, as a result of the eventual balance between the energy loss by radiation and the energy pickup from the stochastic field. Moreover, SED offers a perspective for answering questions of a physical nature, which cannot even be posed within the frame of the usual QM, concerning, for example, the dynamics of the transition to a state of equilibrium, or the connection between the wave properties of matter and the statistical properties of the background field.[11]

The conceptual picture offered by SED and the concrete results obtained up to now strongly suggest the possibility of considering it as an alternative to QM or, more correctly, to (nonrelativistic) quantum electrodynamics (QED); this has stimulated various attempts to establish the connection with quantum theory on firm grounds (see, e.g., Refs. 2, 8, 12, and 13), some of them very interesting and suggestive indeed. However, these attempts are not always as general and conclusive as we would like them to be, due, perhaps above all, to mathematical difficulties. Yet SED has now reached a stage of development where it seems possible to use it as the basis for the development of a fundamental theory of QM; this is the motivation for the present work.

The structure of the paper is as follows: In Sec. I we define the SED problem; here we introduce the fundamental postulate about the existence of the zero-point radiation field with appropriately defined statistical properties. It is through these properties that Planck's constant enters into the description, since the spectral density assumed is just that of the photon vacuum of QED.

In Sec. II we derive the corresponding Fokker–Planck-type equation for the distribution in phase space, on the basis of a stochastic Liouville equation; the process turns out to be definitely non-Markoffian in phase space. The transition to configuration space is performed in Sec. III, where an infinite hierarchy of flow equations is obtained. Through an appropriate change of variables $(x, p \to z_+, z_-)$ containing a free parameter β, the first two equations can be combined and integrated to give a "perturbed" Schrödinger equation in terms of β (Sec. IV), the perturbation being a result of the coupling with the rest of the hierarchy. When a state of equilibrium with the background field is reached, at which we may assume that the coupling becomes negligibly small, we get just Schrödinger's equation. In this quantum mechanical régime, where the z_+ and z_- spaces separate, the mean energy attains an extremum value (Sec. V). The description reduces in this limit to that given by the (phenomenological) theory of stochastic QM, developed in previous work.[14,15] The corrections to the Schrödinger equation in the general case represent the (nonrelativistic) radiative corrections, which are not explicitly calculated here.

In Sec. VI we show that the phase-space distribution coincides in the quantum-mechanical régime with the Wigner distribution. This means that the average values of physical quantities given by SED in the equilibrium limit coincide with the corresponding quantum mechanical quantities, if these are calculated using Weyl's rule of correspondence, since the Wigner distribution is uniquely determined by this rule.[16] Before the quantum mechanical régime is attained (i.e., at short times), the SED system can of course violate the predictions of QM; this applies, in particular, to the Heisenberg inequalities.

Section VII is dedicated to the calculation of the parameter β contained in the new variables z_+, z_-. A straightforward calculation on the basis of the Langevin equation for the harmonic oscillator gives the expected value $\beta = \frac{1}{2}\hbar$.

Finally, in Section VIII we discuss some of the relevant results and their implications concerning the connection between SED and QM. It is clear that many important aspects of the problem are barely touched upon—or even ignored; we hope that this paper serves to stimulate interest in their study.

I. THE SED SYSTEM

The subject of SED is a nonrelativistic charged particle acted on by three forces: an external force, the radiative reaction force, and the force due to the electric component of the stochastic background field. The corresponding equation of motion is the so-called Langevin equation (in one dimension, for simplicity):

$$m\ddot{x} = f + m\tau\dddot{x} + eE\,, \quad \tau = 2e^2/3mc^3\,, \tag{1}$$

where m is the mass and e the charge of the particle. The external force f is assumed conservative. In the dipole approximation—which is consistent with this nonrelativistic treatment—the random electric field E is a function of time only. The statistical properties of this field, at temperature $T = 0$, are assumed to be the following[1,4,5,9]:

(i) E is a stationary Gaussian process with zero average,

(ii) its spectral energy density is

$$\rho(\omega) = \hbar\omega^3/2\pi^2c^3 \tag{2}$$

which means that the autocorrelation function of its Fourier transform

$$\tilde{E}(\omega) \equiv \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} E(t)e^{i\omega t}dt$$

is

$$\langle \tilde{E}(\omega)\tilde{E}^*(\omega')\rangle_E = \frac{2\hbar|\omega|^3}{3c^3}\delta(\omega - \omega')\,, \tag{3}$$

where $\langle\,\rangle_E$ denotes the average over all samples of E. Equation (2) implies that the energy per normal mode of the field is $\frac{1}{2}\hbar\omega$. At temperatures $T > 0$ the energy per normal mode follows Planck's law and Eqs. (2) and (3) would have to be modified accordingly.

The assumption on the Gaussian character of the probability distribution of the field amplitudes, is introduced as the simplest and most plausible one. Equation (2) for the energy spectral density of the field has been derived[1,17] under the assumption that the random electromagnetic field has a Lorentz-invariant spectrum. It is possible to show[5] that the random field with properties (i) and (ii) above, closely resembles the free radiation field of quantum electrodynamics—which is, of course, most desirable. It should be noted, however, that Eq. (2) does not represent a true spectral density, because it is not integrable; this gives rise to ultraviolet divergencies in SED, of a nature similar to those encountered in QED; clearly, this problem demands a careful study. However, since the major results presented here are not affected by such difficulties, we leave the discussion of the problem for a forthcoming paper.

Another problem—connected to the former through the fluctuation-dissipation theorem[18]—is the approximate nature of the damping term $m\tau\dddot{x}$ which gives rise to runaway solutions to Eq. (1). These unphysical solutions can, however, be eliminated by the simple expedient of rewriting Eq. (1) in such a form as to guarantee that the final acceleration of the particle is zero,

$$m\ddot{x} = \frac{1}{\tau}\int_t^\infty e^{(t-t')/\tau}[f(x(t')) + eE(t')]\,dt'\,.$$

The solutions of this equation are the physical solutions of Eq. (1), as is easily seen by differentiating it with respect to t. Now it can be approximated by a differential equation, by developing $f(x(t'))$ around $x(t)$ on account of the smallness of τ (for an electron, $\tau \sim 10^{-22}$ s). Keeping only the first two terms of the Taylor series, we get

$$m\ddot{x} = f(x) + \tau\dot{x}f'(x) + eE_m\,, \tag{4}$$

where

$$E_m \equiv \frac{1}{\tau}\int_t^\infty e^{(t-t')/\tau}E(t')\,dt'$$

has the modified spectral density

$$\rho_m(\omega) = \frac{\hbar\omega^3}{2\pi^2c^3(1+\tau^2\omega^2)}\,. \tag{5}$$

For a force linear in x, Eq. (4) can be readily solved. Indeed, it has been shown that the harmonic oscillator of SED acquires, at times $t \gg (\tau\omega^2)^{-1}$, precisely the properties of the quantum oscillator[1,4,9]; the equilibrium mixture at temperatures greater than 0,[1,9] the Lamb shift,[3,4,8,9] the (finite) mass renormalization,[9] and the mean lifetime of the excited states[10] are correctly predicted by the theory.

There is no general method, however, for solving Eq. (4) with an arbitrary force. We shall therefore construct a statistical description in terms of an equation of the Fokker-Planck type for the probability density.

II. A FOKKER-PLANCK-TYPE EQUATION IN PHASE SPACE

To construct this equation we proceed as follows. Equation (4) may be written as the system

$$m\dot{x} = p\,, \quad \dot{p} = f + \tau f'p/m + eE_m\,. \tag{6}$$

This set of equations defines a stochastic dynamical process in phase space. For each particular realization of E_m, the density of points in phase space—which we call $R(x,p,t)$—satisfies an equation of continuity,

$$\frac{\partial R}{\partial t} + \frac{\partial}{\partial x}(\dot{x}R) + \frac{\partial}{\partial p}(\dot{p}R) = 0\,, \tag{7}$$

which can be cast in the form of a stochastic Liouville equation, using Eqs. (6),

$$\frac{\partial R}{\partial t} + \hat{L}R = -e\frac{\partial}{\partial p}(E_mR)\,, \tag{8}$$

where $\hat{L}$ is the nonrandom Liouville operator with radiation force,

$$\hat{L} = \frac{1}{m}\frac{\partial}{\partial x}p + \frac{\partial}{\partial p}\left(f + \tau f'\frac{p}{m}\right)\,. \tag{9}$$

We are not interested in the motion of a single system but in the average motion of an ensemble, the elements of which correspond to the different E_m; we shall therefore take an average over all realizations of E_m. For this purpose it is convenient to introduce the averaging or smoothing operator $\hat{P}$ such that for any phase function $A(x,p,t)$, $\hat{P}A \equiv \langle A\rangle_E$ ($\hat{P}$ is a projector, since $\hat{P}^2=\hat{P}$). A can then be decomposed into the sum of its average, $\hat{P}A$, and its random part, $(1-\hat{P})A$. For the stochastic field E_m, in particular, we have

$$\hat{P}E_m(t_1)E_m(t_2)\cdots E_m(t_{2n+1})\hat{P}A=0\,,$$

$$\hat{P}E_m(t_1)E_m(t_2)\cdots E_m(t_{2n})\hat{P}A \tag{10}$$

$$=\sum_{\text{all pairs}}\langle E_m(t_i)E_m(t_j)\rangle_E\cdots\langle E_m(t_r)E_m(t_s)\rangle_E\langle A\rangle_E\,,$$

where $n=0,1,2,\cdots$, for all A.

Let us now separate the density function R into its average and stochastic parts, respectively,

$$Q=\hat{P}R\,,\quad \delta Q=(1-\hat{P})R \tag{11}$$

and introduce them into Eq. (8),

$$\frac{\partial}{\partial t}(Q+\delta Q)+\hat{L}(Q+\delta Q)=-e\frac{\partial}{\partial p}E_m(Q+\delta Q)\,. \tag{12}$$

Applying first $\hat{P}$ and then $(1-\hat{P})$ to this equation, we get

$$\frac{\partial Q}{\partial t}+\hat{L}Q=-e\frac{\partial}{\partial p}\hat{P}E_m\delta Q\,, \tag{13}$$

$$\frac{\partial}{\partial t}\delta Q+\hat{L}\delta Q=-e\frac{\partial}{\partial p}E_mQ-e\frac{\partial}{\partial p}(1-\hat{P})E_m\delta Q\,. \tag{14}$$

Now we want to eliminate δQ. With this purpose we introduce the inverse of the operator $\partial/\partial t+\hat{L}$, which we call $\hat{G}$; then for any phase function A we can write

$$\hat{G}A(x,p,t)\equiv\left(\frac{\partial}{\partial t}+\hat{L}\right)^{-1}A(x,p,t)=\int_0^t e^{-\hat{L}(t-t')}A(x,p,t')\,dt'\,. \tag{15}$$

Now we invert Eqs. (13) and (14), to get

$$Q=-e\hat{G}\frac{\partial}{\partial p}\hat{P}E_m\delta Q\,, \tag{16}$$

$$\delta Q=-e\hat{G}\frac{\partial}{\partial p}E_mQ-e\hat{G}\frac{\partial}{\partial p}(1-\hat{P})E_m\delta Q \tag{17}$$

and by combining these we obtain the equation

$$Q=e^2\hat{G}\frac{\partial}{\partial p}\hat{P}E_m\left[1+e\hat{G}\frac{\partial}{\partial p}(1-\hat{P})E_m\right]^{-1}\hat{G}\frac{\partial}{\partial p}E_mQ$$

which can be rewritten as

$$\frac{\partial Q}{\partial t}+\hat{L}Q=e^2\frac{\partial}{\partial p}\hat{P}E_m\left[1+e\hat{G}\frac{\partial}{\partial p}(1-\hat{P})E_m\right]^{-1}\hat{G}\frac{\partial}{\partial p}E_mQ\,. \tag{18}$$

This is not a convenient expression, since the operator $\hat{G}$ and the stochastic field E_m appear in the denominator; but it can be formally expanded in series, yielding

$$\frac{\partial Q}{\partial t}+\hat{L}Q=e^2\frac{\partial}{\partial p}\hat{P}E_m\sum_{k=0}^{\infty}\left[-e\hat{G}\frac{\partial}{\partial p}(1-\hat{P})E_m\right]^k\hat{G}\frac{\partial}{\partial p}E_mQ$$

$$=-e\frac{\partial}{\partial p}\hat{P}E_m\sum_{k=1}^{\infty}\left[-e\hat{G}\frac{\partial}{\partial p}(1-\hat{P})E_m\right]^kQ\,. \tag{19}$$

To obtain the second equality we have written the last factor E_mQ in the equivalent form $(1-\hat{P})E_mQ$.

The method used to go from the random equation (8) to the nonrandom equation (18) is called the method of smoothing and has been introduced independently by a number of authors for different problems; for a detailed exposition see Ref. 19.

On account of the first of Eqs. (10) and of the nonrandom character of $\hat{G}$, all terms with even k on the right side of Eq. (19) vanish and we are left with

$$\frac{\partial Q}{\partial t}+\hat{L}Q=e\frac{\partial}{\partial p}\hat{P}E_m\sum_{k=0}^{\infty}\left[e\hat{G}\frac{\partial}{\partial p}(1-\hat{P})E_m\right]^{2k+1}Q\,. \tag{20}$$

This equation has still an infinite number of terms which represent the (averaged) effects of the multiple scattering by the vacuum field, which means that it includes the (nonrelativistic, all order) radiative corrections of QED. Although for QM it is sufficient to keep the first term ($k=0$), we shall keep them all and write the rhs of Eq. (20), for short, as

$$e^2\frac{\partial}{\partial p}\hat{P}E_m\hat{G}\frac{\partial}{\partial p}E_m\sum_{k=0}^{\infty}\left[e\hat{G}\frac{\partial}{\partial p}(1-\hat{P})E_m\right]^{2k}Q$$

$$\equiv\frac{\partial}{\partial p}\sum_{k=0}^{\infty}\hat{D}_{2k}\frac{\partial Q}{\partial p}=\frac{\partial}{\partial p}\hat{D}\frac{\partial}{\partial p}Q\,. \tag{21}$$

$\hat{D}$ is a complicated integrodifferential operator, but it has the valuable property of commuting with the operator $\partial/\partial p$. To prove this we take the first term of the series,

$$D_0\frac{\partial Q}{\partial p}=e^2\hat{P}E_m\hat{G}\frac{\partial}{\partial p}E_mQ\,; \tag{22}$$

next we replace the operator $\hat{G}$ of Eq. (15) by a kernel, thus writing

$$\hat{G}A(x,p,t)=\int_0^t dt'\iint dx'dp'\mathcal{G}(x-x',p-p',t-t')A(x',p',t')\,, \tag{23}$$

where $\mathcal{G}$ is a retarded Green function,

$$\left(\frac{\partial}{\partial t}+\hat{L}\right)\mathcal{G}=0\,,\quad \mathcal{G}(x,p,0)=\delta(x)\delta(p)\,. \tag{24}$$

Equation (23) is easily verified by applying to it the operator $\partial/\partial t+\hat{L}$ and using Eqs. (24). For $A=\partial(E_mQ)/\partial p$ we obtain from Eqs. (22) and (23),

$$\hat{D}_0\frac{\partial Q}{\partial p}=e^2\hat{P}E_m(t)\int_0^t dt'\iint dx'dp'\mathcal{G}(x-x',p-p',t-t')$$

$$\times E_m(t')\frac{\partial}{\partial p'}Q(x',p',t')$$

$$=e^2\frac{\partial}{\partial p}\int_0^t dt'\,\hat{P}E_m(t)E_m(t')$$

$$\times\iint dx'dp'\mathcal{G}(x-x',p-p',t-t')Q(x',p',t')$$

$$=\frac{\partial}{\partial p}\hat{D}_0Q\,, \tag{25}$$

where we have performed an integration by parts. Thus we see that the first term of the series turns out to be a derivative with respect to p. Inspection of Eq. (21)

shows that the same is true for the remaining terms of the series; we may therefore write

$$\hat{D}\frac{\partial}{\partial p}=\frac{\partial}{\partial p}\hat{D}\,. \tag{26}$$

Using this property we may cast the differential equation for the phase-space distribution, Eq. (20), into the form

$$\frac{\partial Q}{\partial t}+\frac{p}{m}\frac{\partial Q}{\partial x}+\frac{\partial}{\partial p}\left(f+\frac{\tau}{m}f'p\right)Q=\frac{\partial^2}{\partial p^2}\hat{D}Q\,, \tag{27}$$

where $\hat{D}$ is a linear operator involving a time integration from 0 to t, as seen from Eq. (25).

We have thus arrived at a Fokker–Planck-type equain phase space. A remarkable feature of this equation is its non-Markoffian character: the change of Q at time t depends on the previous values of Q, on account of the right-hand side, which involves an integration over the past. This non-Markoffian character is a direct consequence of the finite duration of the field autocorrelation, implied in the frequency-dependent spectrum, Eq. (3). This is a most important point: We are dealing with a non-Markoffian process in phase space, and the ensuing configuration-space statistical description will clearly be non-Markoffian as well. This seems to contradict previous work by Nelson and others,[14,15] according to which QM may be understood as the result of a Markoff process in configuration space. Close inspection reveals, however, that the phenomenological stochastic theories of QM always contain, in some form or another, a dynamical postulate which gives rise to a nonclassical (i.e., different from Brownian-motion) behavior of the system; this point is only too seldom made explicit. The price we pay for the apparent simplicity of the phenomenological Markoffian approaches to QM are the conceptual difficulties arising from their nonclassical dynamical content (see, e.g., Refs. 20 and 21), which have gone so far as to lead to the assertion that the relationship between stochastic processes and QM is purely formal and devoid of any physical meaning.[22,23] These difficulties are lifted, at least in principle, as soon as we recognize that the underlying stochastic process of QM is much more complex than was previously assumed.

Equation (27) offers the possibility—in principle, at least—to study QM in phase space. This would not be the first phase-space description of QM ever made: Since the early work of Wigner,[24] various authors have proposed different phase-space descriptions which, through the use of different correspondence rules, can be shown to be formally equivalent to ordinary QM (see, e.g., Refs. 16 and 25). We shall return to this point in Sec. VI. There are other approaches, notably that of Wiener and della Riccia,[26] based on a classical Liouville equation for a phase-space probability amplitude, and the SED approach based on the direct solution of the Langevin equation for the harmonic oscillator.[1,4,9] While the approach of Wigner, Moyal, etc., leads to a formal extension of usual QM, SED allows, in principle, for predictions outside the frame of usual QM (e.g., for short times).

Equation (27) does not seem much easier to solve than the original Langevin equation for an arbitrary force. A detailed phase-space description based on this equation would therefore involve serious mathematical difficulties, aside from those related to the eventual occurrence of infinities associated with the oversimplified form of $\rho(\omega)$. However, many properties of the system can be investigated without entering into such difficulties; in particular, we shall in the next section proceed to establish the connection with Schrödinger QM.

III. TRANSITION TO CONFIGURATION SPACE

A complete description of our system in configuration space can be obtained by simply multiplying Eq. (27) by p^n ($n=0,1,2,\cdots$) and integrating over the whole p-space. Or alternatively it can be obtained in terms of the characteristic function

$$\tilde{Q}(x,z,t)=\int Q(x,p,t)e^{ipz}\,dp \tag{28}$$

which is the generating function of the configuration-space-conditioned or local moments $\langle p^n\rangle_x\equiv\rho^{-1}\int p^nQ\,dp$, as follows from

$$\frac{\partial^n\tilde{Q}}{\partial z^n}=i^n\int p^nQe^{ipz}\,dp=i^n\rho\langle p^ne^{ipz}\rangle_x \tag{29}$$

by taking $z=0$,

$$\langle p^n\rangle_x=\frac{1}{\rho}\int p^nQ\,dp=(-i)^n\left(\frac{1}{\tilde{Q}}\frac{\partial^n\tilde{Q}}{\partial z^n}\right)_{z=0}\,. \tag{30}$$

Here we have introduced the density function in configuration space, which, according to Eq. (28), is

$$\rho(x,t)\equiv\int Q(x,p,t)\,dp=\tilde{Q}(x,0,t)\,. \tag{31}$$

From the Fourier transform of Eq. (27),

$$\frac{\partial\tilde{Q}}{\partial t}-\frac{i}{m}\frac{\partial^2\tilde{Q}}{\partial x\partial z}-izf\tilde{Q}-\frac{\tau}{m}f'z\frac{\partial\tilde{Q}}{\partial z}=-z^2(\hat{D}Q)^{\tilde{}} \tag{32}$$

and Eq. (29), we get

$$\frac{\partial}{\partial t}(\langle e^{ipz}\rangle_x\rho)+\frac{1}{m}\frac{\partial}{\partial x}(\langle pe^{ipz}\rangle_x\rho)-izf\langle e^{ipz}\rangle_x\rho$$

$$-\frac{i\tau}{m}f'z\langle pe^{ipz}\rangle_x\rho=-z^2(\hat{D}Q)^{\tilde{}} \tag{33}$$

where the tilde denotes Fourier transformation. This equation still contains all the phase-space information, as it is written in xz space. We can transfer this information to configuration space step by step, by developing e^{ipz} in Taylor's series around $z=0$ and separating the coefficients of z^n ($n=0,1,2,\cdots$). The first equations thus obtained are

$$\frac{\partial\rho}{\partial t}+\frac{1}{m}\frac{\partial}{\partial x}(\langle p\rangle_x\rho)=0\,, \tag{34}$$

$$\frac{\partial}{\partial t}(\langle p\rangle_x\rho)+\frac{1}{m}\frac{\partial}{\partial x}(\langle p^2\rangle_x\rho)-f\rho-\frac{\tau}{m}f'\langle p\rangle_x\rho=0\,, \tag{35}$$

$$\frac{\partial}{\partial t}(\langle p^2\rangle_x\rho)+\frac{1}{m}\frac{\partial}{\partial x}(\langle p^3\rangle_x\rho)-f\langle p\rangle_x\rho-\frac{\tau}{m}f'\langle p^2\rangle_x\rho=2(\hat{D}Q)^{\tilde{}}_{z=0}\,. \tag{36}$$

These equations describe the flow of matter, momentum and energy; higher powers of z would yield additional transport equations, each one containing a new local function $\langle p^n\rangle_x$. Consequently, the phase-space description is equivalent to an infinite hierarchy of transport equations (in configuration space). However, we are interested here in the connection with Schrödinger QM and to establish it, the first two equations are sufficient, as will be seen below.

The local moments appearing in Eqs. (34) and (35) may be rewritten, according to Eq. (30), as follows:

$$\langle p\rangle_x = -i\left(\frac{\partial}{\partial z}\ln\tilde{Q}\right)_{z=0} \tag{37}$$

and

$$\langle p^2\rangle_x = -\left(\frac{\partial}{\partial z}\ln\tilde{Q}\right)^2_{z=0} - \left(\frac{\partial^2}{\partial z^2}\ln\tilde{Q}\right)_{z=0}$$

$$= \langle p\rangle_x^2 - \left(\frac{\partial^2}{\partial z^2}\ln\tilde{Q}\right)_{z=0}. \tag{38}$$

The last expression is more conveniently written in terms of the variables[27]

$$z_+ = x+\beta z\,,\quad z_- = x-\beta z \tag{39}$$

where β is an arbitrary (real) constant whose value will be determined below. In fact, with $\partial_\pm \equiv \partial/\partial z_\pm$ Eq. (38) can be written

$$\langle p^2\rangle_x - \langle p\rangle_x^2 = -\beta^2[(\partial_+ - \partial_-)^2\ln\tilde{Q}]_{z=0}$$

$$= -\beta^2\left(\frac{\partial^2}{\partial x^2}\ln\tilde{Q}\right)_{z=0} + 4\beta^2(\partial_+\partial_-\ln\tilde{Q})_{z=0}. \tag{40}$$

Now we write $\tilde{Q}$ as the product of three functions, in the noncommittal form

$$\tilde{Q}(x,z,t) = q_+(z_+,t)q_-(z_-,t)q(z_+,z_-,t)\,, \tag{41}$$

where q is not further factorizable into z_+ and z_- functions. If we use this expression for $\tilde{Q}$ and recall that $\tilde{Q}(x,0,t)=\rho(x,t)$, we obtain from Eq. (40),

$$\langle p^2\rangle_x - \langle p\rangle_x^2 = -\beta^2\frac{\partial^2}{\partial x^2}\ln\rho + \sigma\,, \tag{42}$$

where

$$\sigma = 4\beta^2(\partial_+\partial_-\ln q)_{z=0}. \tag{43}$$

Observe that it is the function σ which connects Eq. (35) with the subsequent equations of the hierarchy. Now let us go back to the first two flow equations; for our present purposes it is convenient to write them in vector notation. With the usual expression for the flow velocity,

$$v_i = \langle p_i\rangle_x/m\,, \tag{44}$$

Eq. (34) assumes the form of the continuity equation (a sum over repeated indices is understood),

$$\frac{\partial\rho}{\partial t} + \partial_i(v_i\rho) = 0\,, \tag{45}$$

and Eq. (35) becomes

$$m\frac{\partial}{\partial t}(v_i\rho) + \frac{1}{m}\partial_j(\langle p_ip_j\rangle_x\rho) - f_i\rho - \tau(\partial_jf_i)v_j\rho = 0\,. \tag{46}$$

Since the vector form of Eq. (42) reads

$$\langle p_ip_j\rangle = m^2v_iv_j - \beta^2\partial_i\partial_j\ln\rho + \sigma_{ij}\,, \tag{47}$$

where

$$\sigma_{ij} = \sigma_{ji} = 2\beta^2[(\partial_{+_i}\partial_{-_j} + \partial_{+_j}\partial_{-_i})\ln q]_{z=0}\,, \tag{48}$$

Eq. (46) transforms into

$$m\frac{\partial}{\partial t}(v_i\rho) + m\partial_j(v_iv_j\rho) - \frac{\beta^2}{m}\partial_j(\rho\partial_i\partial_j\ln\rho) - f_i\rho$$

$$= -\frac{1}{m}\partial_j(\sigma_{ij}\rho) + \tau\rho v_j\partial_jf_i\,.$$

By further introducing Eq. (45) and the potential V associated with the external force $\vec{f}$, we obtain after some minor algebra,

$$m\frac{\partial v_i}{dt} + mv_j\partial_jv_i - \frac{\beta^2}{m}\partial_i[\tfrac{1}{2}(\partial_j\ln\rho)^2 + \partial_j^2\ln\rho] + \partial_iV$$

$$= -\frac{1}{m\rho}\partial_j(\rho\partial_{ij}) + \tau v_j\partial_jf_i \equiv F_i\,. \tag{49}$$

This equation has a close resemblance to the classical hydrodynamic equation for the momentum flow of a viscous fluid,[28] with a stress tensor $\rho\sigma_{ij}/m^2$. There is, however, an important difference, namely, the terms containing $\partial_j\ln\rho$. These are kinetic, nondissipative terms of stochastic origin,[29] which manifest the different nature of the SED system: it is obviously not a classical hydrodynamic system. Nevertheless, the transport equations may be useful (and in fact the particular form of them with $F_i=0$ has been used) to establish hydrodynamical analogs for QM.

It is also interesting to note that Eqs. (45) and (49) (with $F_i=0$) are precisely the fundamental equations of the stochastic theory of QM (SQM), the phenomenological counterpart of the present theory.[14,15] This theory offers thus a physical justification for the postulates of SQM and moreover, determines its free parameters from first principles, as will be seen below.

IV. THE SCHRÖDINGER EQUATION

We will now construct a first integral of Eq. (49). For this purpose, we write (without loss of generality) the flow velocity v and the right-hand member F in terms of new functions, as follows:

$$\vec{v} = \frac{1}{m}(2\beta\nabla S + \vec{B}),\quad \nabla\cdot\vec{B} = 0, \tag{50a}$$

$$\vec{F} = -\nabla\Phi + \vec{K},\quad \nabla\cdot\vec{K} = 0. \tag{50b}$$

With these expressions, Eq. (49) takes the form

$$\partial_i\left[2\beta\frac{\partial S}{\partial t} + \tfrac{1}{2}m\vec{v}^2 + \frac{\beta^2}{2m}\left(\frac{\partial_j\rho}{\rho}\right)^2 - \frac{\beta^2}{m}\frac{\partial_j^2\rho}{\rho} + V + \Phi\right]$$

$$= -\frac{\partial B_i}{\partial t} + v_j(\partial_iB_j - \partial_jB_i) + K_i, \tag{51}$$

which has as first integral the Hamilton–Jacobi equation

$$2\beta\frac{\partial S}{\partial t} + \frac{1}{2m}(2\beta\nabla S + \vec{B})^2 + \frac{\beta^2}{2m}\left(\frac{\nabla\rho}{\rho}\right)^2 - \frac{\beta^2}{m}\frac{\nabla^2\rho}{\rho} + V + \Phi = 0. \tag{52}$$

We have absorbed the integration constant into $\partial S/\partial t$ and have selected $\vec{B}$ so as to cancel the rhs of Eq. (51), i.e., $\vec{B}$ is a solution of the differential equation

$$\frac{\partial\vec{B}}{\partial t} - \vec{v}\times(\nabla\times\vec{B}) = \vec{K}. \tag{53}$$

Through the usual change of variable

$$\psi = (\rho)^{1/2} e^{iS} \tag{54}$$

Eq. (52) transforms into a complex equation for ψ and ψ^* which is trivially separable into the Schrödinger equation

$$2i\beta\frac{\partial\psi}{\partial t} = \frac{1}{2m}(-2i\beta\nabla + \vec{B})^2\psi + (V+\Phi)\psi \tag{55}$$

and its complex conjugate.

Observing that $\vec{B}$ and Φ appear as a vector and a scalar potential respectively, it is convenient to rewrite Eq. (55) as follows:

$$2i\beta\frac{\partial\psi}{\partial t} = \frac{1}{2m}\left(-2i\beta\nabla - \frac{e}{c}\vec{A}_r\right)^2\psi + (V + e\phi_r)\psi, \tag{56}$$

where

$$\vec{B} = -\frac{e}{c}\vec{A}_r, \quad \Phi = e\phi_r. \tag{57}$$

These potentials contain the nonrelativistic radiative corrections to the Schrödinger equation. From Eqs. (50*b*) and (53), they can be shown to satisfy the equation

$$\vec{F} = -e\nabla\phi_r - \frac{e}{c}\frac{\partial\vec{A}_r}{\partial t} + \frac{e}{c}\vec{v}\times(\nabla\times\vec{A}_r), \tag{58}$$

which assigns a meaning to $\vec{F}$ in Eq. (49). It is easy to show that the introduction of an external Lorentz force in the Langevin equation [or in Eq. (27)] results in the addition of the corresponding potentials to $\vec{A}_r$ and ϕ_r in the Schrödinger equation. For the calculation of the potentials we use Eq. (50b), from which we obtain

$$\Phi = \frac{1}{4\pi}\int\frac{\nabla\cdot\vec{F}(r')}{|r-r'|}d^3r', \quad K = \frac{1}{4\pi}\nabla\times\int\frac{\nabla\times\vec{F}(r')}{|r-r'|}d^3r' \tag{59}$$

where the components of $\vec{F}$ are given by Eq. (49),

$$F_i = -\frac{1}{m\rho}\partial_j(\rho\sigma_{ij}) - \tau\nu_j\partial_j\partial_i V. \tag{60}$$

To evaluate $\vec{F}$ explicitly we would need to know the function $\vec{\sigma}$, which in its turn requires the knowledge of the solution of the phase-space problem. Below we shall once more touch upon some aspects of the radiative corrections.

V. THE LIMIT OF QUANTUM MECHANICS

In the foregoing section we obtained for the SED system a Schrödinger equation containing a still undetermined parameter β and the effective electromagnetic potentials $\vec{A}_r$ and ϕ_r. For this equation to correspond to usual QM, we must have $\beta = \frac{1}{2}\hbar$, and $\vec{A}_r$ and ϕ_r must be zero.

Let us first examine the significance of this last condition being imposed on $\vec{A}_r$ and ϕ_r. From Eqs. (58) and (60) we see that this condition is met only if

$$\frac{1}{m\rho}\partial_j(\rho\sigma_{ij}) + \tau v_j\partial_j\partial_i V = 0. \tag{61}$$

We recall that usual QM is implicitly restricted to nonradiating systems, a restriction which is lifted only by explicitly introducing the interaction with the (quantized) vacuum field; the present theory, on the contrary, contains the radiation term as an essential ingredient. To obtain the radiationless limit implicit in QM we can simply take $\tau = 0$; but in dropping the term with τ we are accepting the existence of a stable physical solution, which is attained precisely thanks to the radiation force, that tends in the long run to compensate for the energy gained by the system due to the stochastic force. We must therefore be careful and take the limit $\tau\to 0$ only once the radiation has had the opportunity to stabilize the system, i.e., for long times. Some related points have been recently discussed by Boyer.[2]

With the above condition, Eq. (61) is satisfied only if $\sigma = 0$, which holds, in particular, if q does not depend on either z_+ or z_- and can therefore be taken as a constant [see Eq. (43)]; the Fourier transform of the phase-space distribution can be written in this case as the product of (time-dependent) functions of either z_+ or z_-, according to Eq. (41).

Clearly, a system subject to arbitrary initial conditions does not necessarily meet these two requirements at all times; but if it is to be correctly described by usual QM, it must evolve to a state of equilibrium with the stochastic background field, in which they apply at least approximately. Let us now prove that this "radiationless" "separable" limit described by ordinary QM (i.e., the quantum mechanical régime) is indeed a state of equilibrium.

According to Eq. (27), the ensemble mean of the kinetic and potential energies of the mechanical system evolves in the following way (recall that $\langle A\rangle \equiv \int AQ\,dp\,dx = \int\langle A\rangle_x \rho\,dx$):

$$\frac{d}{dt}\frac{\langle p^2\rangle}{2m} = \frac{1}{2m}\int p^2\frac{\partial Q}{\partial t}\,dp\,dx$$

$$= \frac{1}{m}\int\left[p\left(f + \frac{\tau}{m}f'p\right)Q + \hat{D}Q\right]dp\,dx,$$

$$\frac{d}{dt}\langle V\rangle = \int V\frac{\partial Q}{\partial t}\,dp\,dx = -\frac{1}{m}\int pfQ\,dp\,dx.$$

The time derivative of the mean energy is therefore

$$\frac{d\langle H\rangle}{dt} = \frac{1}{m}\int\left(\frac{\tau}{m}f'p^2Q + \hat{D}Q\right)dp\,dx$$

$$= \frac{\tau}{m}\langle f'p^2\rangle + \frac{1}{m}\langle\hat{D}\rangle. \tag{62}$$

It is clear from this expression that $\langle H\rangle$ may attain a constant value thanks to the balance between the energy radiated and the energy picked up from the stochastic field, as stated above. But since we are already dropping terms with τ in passing to the limit of usual QM, it only remains to prove that $\langle\hat{D}\rangle$ vanishes as well in this limit.

With this purpose, we use a result that will be proved in the next section, namely, that the phase-space distribution coincides in the quantum mechanical régime with the Wigner distribution; Moyal[16] has shown (on the basis of QM) that the Wigner distribution evolves according to the equation

$$\frac{\partial Q}{\partial t} = \sum_{n=0}^{\infty}\sum_{r=0}^{n} \frac{(-1)^n}{(n-r)!\,r!}\left(\frac{\partial}{\partial p}\right)^{n-r}\left(\frac{\partial}{\partial x}\right)^{r}[\alpha_{nr}Q], \qquad (63)$$

where[30]

$$\alpha_{2n+1,r} = (-1)^{n+r+1}\beta^{2n}\left(\frac{\partial}{\partial p}\right)^{r}\left(\frac{\partial}{\partial x}\right)^{2n+1-r} H,$$

$$\alpha_{2n,r} = 0,$$

and

$$H = \frac{p^2}{2m} + V(x).$$

For our purposes it is convenient to rewrite Eq. (63) in its equivalent form

$$\frac{\partial Q}{\partial t} + \frac{p}{m}\frac{\partial Q}{\partial x} + f\frac{\partial Q}{\partial p}$$
$$= -\sum_{n=1}^{\infty}\frac{(-1)^n\beta^{2n}}{(2n+1)!}\left(\frac{\partial}{\partial p}\right)^{2n+1}\left[Q\left(\frac{\partial}{\partial x}\right)^{2n} f\right]. \qquad (64)$$

By comparing Eq. (64) with Eq. (27) we see that in the quantum mechanical régime, $\hat{D}Q$ has the value

$$\hat{D}Q = -\frac{\partial}{\partial p}\sum_{n=1}^{\infty}\frac{(-1)^n\beta^{2n}}{(2n+1)!}\left(\frac{\partial}{\partial p}\right)^{2n-2}\left[Q\left(\frac{\partial}{\partial x}\right)^{2n} f\right], \qquad (65)$$

whence $\langle\hat{D}\rangle = \int \hat{D}Q\, dp\, dx = 0$. In the limit of QM, therefore, $d\langle H\rangle/dt = 0$ indeed.

The constant value of the average energy for any stationary state may be ascertained by looking for the extremum of $\langle H\rangle$ with respect to changes in the (normalized) density of particles, as is well known from elementary QM; this holds even if σ (now time independent) is different from zero. To see this, we recall that in a stationary state the flow velocity vanishes, and hence Eq. (42) reduces to

$$\langle p^2\rangle_x = -\beta^2\frac{d^2}{dx^2}\ln\rho + \sigma. \qquad (66)$$

The condition that the mean energy be an extremum,

$$\langle H\rangle = \int HQ\, dp\, dx = \int \rho\left(-\frac{\beta^2}{2m}\frac{d^2}{dx^2}\ln\rho + \frac{\sigma}{2m} + V\right)dx$$
$$= \int\left[\frac{2\beta^2}{m}\left(\frac{d\varphi}{dx}\right)^2 + \left(V + \frac{\sigma}{2m}\right)\varphi^2\right]dx = \text{extremum}, \qquad (67)$$

where $\varphi = (\rho)^{1/2}$, along with the normalization condition

$$\int \rho dx = \int \varphi^2 dx = 1,$$

constitute a variational problem, the solution of which is the Euler–Lagrange equation[31]

$$-\frac{2\beta^2}{m}\frac{d^2\varphi}{dx^2} + \left(V + \frac{\sigma}{2m}\right)\varphi = \langle H\rangle\,\varphi \qquad (68)$$

($\langle H\rangle$ plays the rôle of the Lagrange multiplier); but Eq. (68) is just the time-independent Schrödinger equation (below we show that $\beta = \frac{1}{2}\hbar$), with an additional term $(\sigma/2m)\varphi$ representing a possible remnant of the whole phase-space description. Though it appears as a potential energy, this term is actually of kinetic origin, as shown by Eq. (66).

Equation (68) offers an alternative way for the calculation of the effects of the radiative corrections on the energy of a stationary state; in particular, the first-order correction

$$\delta E = \langle\sigma\rangle/2m \qquad (69)$$

contains two contributions, namely, the mass renormalization due to the vacuum fluctuations, and the Lamb shift. The calculation of these radiative corrections in the general case is evidently a difficult task; however, a first-order calculation is easily performed for the harmonic oscillator on the basis of the Langevin equation, yielding for the mass renormalization[9]:

$$\delta m = 3m/8\pi\alpha \quad (\alpha = e^2/\hbar c)$$

and for the Lamb shift of the ground state[3,4,7,8]:

$$\delta E_L = \frac{\alpha\hbar^2\omega^2}{\pi mc^2}\ln\frac{3mc^2}{2\alpha\hbar\omega}.$$

The large (though finite) values obtained can be traced to the exaggerated contribution of $\rho(\omega)$ [see Eq. (2)] at high frequencies; if the usual nonrelativistic cutoff $\omega_c = mc^2/\hbar$ is introduced, $\delta m/m$ becomes of the order of the fine-structure constant α (but cut-off-dependent) and δE_L coincides with the value predicted by nonrelativistic QED with the same cutoff. For further details see Ref. 9.

We may easily understand why this calculation gives finite results: From Eq. (66) we see that to first order in perturbation theory, σ can be written in the form

$$\sigma = \langle p^2\rangle_x - \langle p^2\rangle_x^0, \qquad (70)$$

where $\langle p^2\rangle_x^0$ represents the unperturbed (i.e., $\sigma = 0$) local variance of p. Due to the long tail of $\langle E_m(t)E_m(t')\rangle_E$, both $\langle p^2\rangle_x^0$ and $\langle p^2\rangle_x$ contain infinite (unphysical) contributions; but their difference is finite.

VI. THE WIGNER DISTRIBUTION

In the foregoing section we saw that the characteristic function $\tilde{Q}$ takes on the simple form

$$\tilde{Q} = q_+(z_+, t)q_-(z_-, t)$$

in the quantum mechanical régime. From Eqs. (30), (31), (39), and (44) we see that ρ and v are given in this case by

$$\rho(x,t) = \tilde{Q}(x,0,t) = q_+(x,t)q_-(x,t) \qquad (71)$$

and

$$v(x,t) = -\frac{i}{m\rho}\left(\frac{\partial\tilde{Q}}{\partial z}\right)_{z=0} = -\frac{i\beta}{m}\frac{\partial}{\partial x}\ln\frac{q_+(x,t)}{q_-(x,t)}. \qquad (72)$$

By comparing with Eqs. (50a)—with $B = 0$, as corresponds to QM—and (54), we see that

$$q_+(x,t) = \psi(x,t), \quad q_-(x,t) = \psi^*(x,t), \qquad (73)$$

whence the phase-space distribution can be written [inverting Eq. (28)]:

$$Q(x,p,t) = \frac{1}{2\pi}\int \psi^*(x-\beta z,t)\psi(x+\beta z,t)e^{-ipz}dz. \qquad (74)$$

This is Wigner's distribution, if $\beta = \frac{1}{2}\hbar$. According to the present theory, this distribution is not expected to hold at all times, but is attained only as a result of a complicated evolution of the system towards the quantum-mechanical régime. As a consequence of this evolution, the phase-space description collapses in such a form that it becomes equivalent to the Schrödinger description, the whole phase-space description being no longer necessary. In other words, SED endows the Wigner description—which is usually presented as a possible phase-space *formalism* of ordinary QM (see, e.g., Ref. 25)—with a well-defined physical meaning: It is the phase-space description of the SED system in the quantum mechanical régime.

As is well known, the Schrödinger equation admits solutions which give negative values for the Wigner distribution[16,32]; for stationary states, for instance, only Gaussian configuration-space distributions give nonnegative phase-space distributions.[33] According to Moyal[16] and Marshall,[1] among others, if Q as given by Eq. (74) is to represent a real distribution, not merely a formal construct, only those quantum mechanical solutions which give nonnegative Q's are physically acceptable. This principle can be justified on the basis that the pure excited states of QM are not, strictly speaking, stationary states and therefore cannot be solutions of the whole phase-space theory, the truly stationary state being in general a quantum mechanical mixture. For example, the stationary state of the harmonic oscillator at temperature $T > 0$, as given by SED,[1,9] is the mixture of states with weights proportional to $\exp(-n\hbar\omega/kT)$, in agreement with statistical QM. Actually, since the Wigner distribution is in general only an approximate expression for the real phase-space distribution, this principle may be relaxed to some extent.

The present phase-space description provides a definite rule of correspondence between classical dynamical variables and quantum mechanical operators, since, as is well known, Wigner's distribution implies Weyl's rule[16,25]:

$$\exp(i\theta x + i\eta p) \rightarrow \exp(i\theta\hat{x} + i\eta\hat{p})$$

or equivalently[34]:

$$x^n p^m \rightarrow \frac{1}{2^n} \sum_{l=0}^{n} \binom{n}{l} \hat{x}^{n-l} \hat{p}^m \hat{x}^l \quad .$$

This means that the SED predictions in the quantum-mechanical régime coincide with those obtained from QM only if Weyl's rule is consistently applied; the same conclusion is contained in the work of Boyer[2] and Santos[12] since the symmetrization rule they propose is equivalent to Weyl's rule. However, the usual operator formalism of QM uses neither Weyl's nor any other correspondence rule; i.e., usual QM is incompatible with any phase-space description.[16,25] The reason behind this discrepancy is that the quantum mechanical dispersions and similar statistical quantities are calculated for variables which are already partly averaged; they are, in the language of analysis of variance, dispersions between classes.[35]

We thus see that ordinary QM is not a truly statistical theory, and hence, we certainly will find it impossible to adhere strictly to a statistical interpretation without at the same time violating some of its statements; we have just met with two instances of this problem. We are thus faced with the alternative: Either we accept a statistical phase-space description, implying the need to revise some of the usual tenets of QM, or we adhere to the usual definitions and hence give up any possibility of constructing a phase-space theory. As yet, the acceptance of a statistical theory with all its implications does not contradict any experimental fact, but differs from usual QM only in the interpretation of certain results.[36]

The present work reveals, moreover, several essential differences between a classical stochastic process, such as Brownian movement, and the stochastic process underlying QM. In the former, which is Markoff process, a state of local equilibrium is eventually attained in which the phase-space variables separate and equipartition of energy holds; in the latter, which is a process with long memory, it is not x and p that eventually separate, but z_+ and z_-, and the average energy attains an extremum value. Once the system has reached the quantum-mechanical régime, x and p remain forever inevitably correlated, $\langle(\Delta x)^2\rangle\langle(\Delta p)^2\rangle \geq \beta^2$. This inequality can of course be violated at short times, when the system is far from equilibrium.

It should be noted that we have assumed that the system may evolve towards equilibrium without specifying under what circumstances it does. Here we touch upon another aspect of the theory which requires careful attention, namely, the ergodic properties of the SED system; the study of some aspects of this important question has been successfully initiated by Claverie and Diner.[37]

VII. THE VALUE OF β

We have just seen that the variables z_+ and z_- define two disjoint subspaces in the equilibrium limit, which means that the parameter β is physically significant and ought therefore to be determined from physical considerations.

That β must be somehow related to $\hbar$ is clear from the fact that it is a measure of the dispersion produced on the dynamic variables by the random field [see, e.g., Eq. (42)], while $\hbar$ arises in SED as a measure of the dispersion of the random field itself [see, e.g., Eq. (3)].

Since the calculation of β for the general case is not a simple task, due to the complicated structure of the operator $\hat{D}$, we resort here directly to the Langevin equation, which contains the necessary information and is at least tractable in the linear case; we shall therefore prove the validity of the formula $\beta = \frac{1}{2}\hbar$ proposed above, for a harmonic oscillator.

The corresponding Langevin equation is [cf. Eq. (4)]

$$m\ddot{x} = -m\omega_0^2 x - m\tau\omega_0^2\dot{x} + eE_m(t) , \qquad (75)$$

where ω_0 is the natural frequency of the oscillator. Let us now average over the ensemble of realizations of E_m. Writing $x = x_c + x_s$, where $x_c \equiv \hat{P}x = \langle x\rangle_E$ and $x_s = (1-\hat{P})x$, we obtain from Eq. (75),

$$\ddot{x}_c = -\omega_0^2 x_c - \tau\omega_0^2 \dot{x}_c \tag{76}$$

and

$$\ddot{x}_s = -\omega_0^2 x_s - \tau\omega_0^2 \dot{x}_s + \frac{e}{m} E_m(t) . \tag{77}$$

The average and the purely random motions become uncoupled, thanks to the linear character of Eq. (75). The stationary solution of Eq. (77) is

$$x_s = \frac{e}{m(2\pi)^{1/2}} \int_{-\infty}^{\infty} \frac{\tilde{E}_m(\omega) e^{-i\omega t}}{\omega_0^2 - \omega^2 + i\tau\omega_0^2\omega}\, d\omega$$

and the dispersion of $\dot{x}_s$ is therefore [using Eq. (5)],

$$\begin{aligned}\langle \dot{x}_s^2 \rangle_E &= \lim_{\eta\to 0} \frac{\hbar\tau}{\pi m} \int_0^{\infty} \frac{\omega^5 e^{-\eta\omega} d\omega}{(1+\tau^2\omega^2)[(\omega_0^2-\omega^2)^2 + (\tau\omega_0^2\omega)^2]} \\ &= \frac{\hbar\omega_0}{2m} + O(\tau) .\end{aligned} \tag{78}$$

The introduction of the convergence factor is justified as follows. The integral expression for $\langle \dot{x}_s^2 \rangle_E$ is divergent, due to the "long tail" of the integrand; but it can be easily seen that the infinite contribution appears already in the transient part of the complete solution of Eq. (77), and hence must be subtracted from the steady-state solution due to its unphysical character; the result thus obtained is $\hbar\omega_0/2m$ (for details see Ref. 9). To simplify the calculation, we observe that the main contribution to the (regularized) integral comes from the sharp resonance at $\omega \simeq \omega_0$ and hence we may modify the shape of the tail without affecting the integral, as is confirmed by the essentially cut-off independent result, Eq. (78). Formally, the problem of evaluating $\langle \dot{x}_s^2 \rangle_E$ may be solved also by a mass renormalization, but the method referred to above (Ref. 8) seems more suggestive. No mathematical trick would of course be necessary if the spectral density of the vacuum field had the right form at high frequencies. This is the only instance in the present paper where we are directly faced with problems generated by the incorrect shape of the spectral density, but it already points to the relevance of this question for the convergence problem in quantum theory.

Since the stationary solution (corresponding to the ground state) of Eq. (76) is $x_c = 0$, $\dot{x}_c = 0$, the total momentum dispersion is given, in a first approximation, by

$$\langle p^2 \rangle = m^2 \langle \dot{x}_s^2 \rangle_E = m\hbar\omega_0/2 . \tag{79}$$

(A similar calculation gives $\langle x^2 \rangle = \hbar/2m\omega_0$, where the total average energy is $\langle H \rangle = \hbar\omega_0/2$, as expected, since the oscillator is in equilibrium with the vacuum field; of course, a more careful calculation yields in addition the Lamb shift correction.)

On the other hand, the dispersion of p for a stationary state ($v = 0$) in the quantum-mechanical régime ($\tau = 0$, $\sigma = 0$) is [cf. Eq. (66)]:

$$\langle p^2 \rangle = -\beta^2 \int \rho\, \partial_x^2 \ln\rho\, dx \ , \tag{80}$$

where ρ is the (integrable) solution of the stationary Schrődinger equation

$$-\frac{2\beta^2}{m} \frac{d^2(\rho)^{1/2}}{dx^2} + \tfrac{1}{2} m\omega_0^2 x^2 (\rho)^{1/2} = \tfrac{1}{2}\hbar\omega_0 (\rho)^{1/2}$$

namely,

$$\rho \sim \exp(-m\omega_0 x^2/\hbar) .$$

Introducing this ρ in Eq. (80) we obtain

$$\langle p^2 \rangle = 2m\omega_0\beta^2/\hbar \ ,$$

finally, comparing this with Eq. (79) we get

$$\beta = \tfrac{1}{2}\hbar \ . \tag{81}$$

Since the value obtained for β is independent of the specific properties of the oscillator, it seems reasonable to accept that it has the same value for any other type of force, at least up to terms not depending on τ.

VIII. FINAL REMARKS

The results presented in this paper show that SED contains the usual (nonrelativistic, spinless) QM as a particular, well defined limit case, obtained when a certain type of equilibrium in phase space is established and the system no longer radiates. Quantum mechanical equilibrium in phase-space differs from its classical counterpart, as discussed in Sec. VI.

The transition to configuration space justifies on physical grounds the use of the Wigner distribution in the limit of QM, and this, in its turn, gives us a reason for selecting the Weyl rule among the various correspondence rules proposed in the literature.

The present theory gives support to the interpretation of QM as a stochastic process, as proposed in previous work[14,15] and determines from first principles the values of the phenomenological parameters appearing therein. It offers, moreover, the possibility of analyzing the quantum mechanical process in phase space for short times, before the equilibrium state is reached—though the mathematical difficulties involved in such an analysis appear to be considerable; the process is far from being Markoffian, as is evident from the integro-differential Fokker–Planck-type equation (27). In the quantum-mechanical régime, the first two equations of the infinite hierarchy suffice to describe the process in configuration space, and these equations define a (nonclassical) Markoff process, as has been shown by the stochastic theory of QM.[14,15] Hence we conclude that for times long compared with an appropriately defined correlation time, the stochastic process may be approximated by a Markoff process in configuration space.

Since the interaction of the mechanical system with the zero-point radiation field is an essential part of the theory, it is possible, in principle, to obtain the radiative corrections to the energy and the lifetimes of excited states, without resorting to further postulates. For the same reason it should be possible to connect the theory with nonrelativistic QED. Besides the interpretative advantages of such an approach, it would probably throw light on some questions related to the need of renormalization; in fact, the present treatment has already allowed us to trace the origin of certain divergencies to the unphysical form of $\rho(\omega)$; it is clear from this that the problem of introducing an acceptable (presumably problem-dependent) form for the spectral density of the vacuum field deserves closer attention.

We have repeatedly referred to the problems generated by a spectrum of the form $\rho(\omega) \sim \omega^3$; let us now review some positive implications of it. As is well known, a uniform spectral density implies a field autocorrelation of the form $\langle E(t)E(t')\rangle_E = \text{const} \times \delta(t-t')$, and this yields a Fokker–Planck equation with no memory, as for Brownian motion [since the operator $\hat{D}$ reduces to a constant, as can be seen from Eqs. (24) and (25)]. The asymptotic solution of this classical Fokker–Planck equation is a Maxwellian distribution multiplied by a $\rho(x)$, which means that the correlation between x and p vanishes and σ as defined in Eq. (42) is essentially different from zero; hence a description in terms of a Schrödinger equation is spurious in this case. It is clear, then, that the spaces which separate under equilibrium are somehow determined by the structure of the spectral density; for QM, in particular, $\rho(\omega)$ must be frequency dependent. Whether $\rho \sim \omega^3$ is the only form leading to a natural description in terms of a linear wave equation is a question whose answer may eventually justify the wave properties of matter within the frame of stochastic theory. The frequency dependence of $\rho(\omega)$ is important also in connection with the atomic stability; in fact, it has been shown by qualitative arguments[2,21] that the hydrogenlike atom is a stable system if $\rho \sim \omega^3$ and the friction force is $\sim \dddot{x}$, while it is not stable for a white-noise spectrum (not even with a Brownian-type friction $\sim \dot{x}$). Moreover, the specific form proposed for $\rho(\omega)$, having been derived under the condition of Lorentz invariance, will allow the eventual passage to a relativistic theory without the need to abandon the main hypotheses of SED.

Finally we should remark that even though the SED system is a charged particle interacting with the random electromagnetic field, the charge does not appear in the final results concerning QM proper (i.e., the Schrödinger equation, the value of β, etc.); though it appears, of course, in the radiative corrections, through the fine-structure constant. We could therefore conceive of QM—along with Boyer[2]—as the mechanical limit of SED. The same mechanical equations would be obtained for a neutral particle, assuming that it possesses a fluctuating charge, provided the fluctuations are sufficiently rapid not to contradict the observed law of charge conservation. We may speculate then that the same mechanism is responsible for the quantum mechanical behavior of both charged and neutral particles. In fact, on the basis of the present results we could go even further and speculate that the mechanism responsible for the quantum mechanical behavior is of a more general nature, involving not only the electromagnetic radiation field—the *specific* effects of which are the radiative corrections—but possibly other stochastic fields as well; this would lead us to consider the stochastic theory as something much more fundamental than we have up to now.

ACKNOWLEDGMENTS

The authors would like to thank Professor T. A. Brody for many interesting discussions and comments and Dr. E. Santos for his comments in the early stage of development of the present work.

*Consultant, Instituto Nacional de Energía Nuclear, México.

[1] T. W. Marshall, Proc. R. Soc. **276A**, 475 (1963); Proc. Cambridge Philos. Soc. **61**, 537 (1965). {Among the earliest references on SED are: N. L. Kalitzin, Zh. Eksp. Teor. Fiz. **25**, 407 (1953); P. Braffort and C. Tzara, C. R. Acad. Sci. Paris **239**, 1779 (1954); A. A. Sokolov and V. S. Tumanov, Zh. Eksp. Teor. Fiz. **30**, 802 (1956) [Sov. Phys. JETP **3**, 958 (1956)].}

[2] T. H. Boyer, Phys. Rev. D **11**, 790, 809 (1975). The first paper constitutes a valuable review of certain topics of SED.

[3] P. Braffort, M. Surdin, and T. Taroni, C. R. Acad. Sci. Paris **261**, 4339 (1965); P. Braffort, C. R. Acad. Sci. Paris **270**, 12 (1970), and references therein.

[4] E. Santos, Nuovo Cimento B **19**, 57 (1974).

[5] E. Santos, Nuovo Cimento B **22**, 201 (1974).

[6] T. H. Boyer, Phys. Rev. A **5**, 1799 (1972); **6**, 314 (1972); **7** 1832 (1973); **9**, 2078 (1974).

[7] T. W. Marshall, Nuovo Cimento **38**, 206 (1965); T. H. Boyer, Phys. Rev. **174**, 1631 (1968).

[8] M. Surdin, Int. J. Theor. Phys. **4**, 117 (1971); Ann. Inst. Henri Poincaré **15**, 203 (1971); C. R. Acad. Sci. **278B**, 881 (1974).

[9] L. de la Peña and A. M. Cetto, Phys. Lett. A **47**, 183 (1974); Rev. Mex. Fís. **25**, 1 (1976).

[10] L. de la Peña and A. M. Cetto, preprint IFUNAM 75-21 (1975).

[11] Conjectures of this kind have been advanced by almost any author working on SED, as those cited in the above references, and also by other authors; see, e.g., A. F. Kracklauer, Scientia **109**, 111 (1974).

[12] E. Santos, J. Math. Phys. **15**, 1954 (1974); Ann. Fís. **71**, 329 (1975); Am. J. Phys. **44**, 278 (1976).

[13] L. de la Peña and A. M. Cetto, Phys. Lett. A **56**, 253 (1976).

[14] E. Nelson, Phys. Rev. **150**, 1079 (1966); L. de la Peña, Phys. Lett. A **27**, 594 (1968); J. Math. Phys. **10**, 1620 (1969). [The first stochastic theory of QM was proposed by I. Fényes, Z. Phys. **132**, 81 (1952).]

[15] L. de la Peña and A. M. Cetto, Found. Phys. **5**, 355 (1975).

[16] J. E. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).

[17] T. H. Boyer, Phys. Rev. **182**, 1374 (1969).

[18] H. B. Callen and T. A. Welton, Phys. Rev. **83**, 34 (1951).

[19] U. Frisch, in *Probabilistic Methods in Applied Mathematics*, edited by A. T. Bharucha-Reid (Academic, New York, 1968), Vol. I.

[20] P. Claverie and S. Diner, C. R. Acad. Sci. Ser. B **280**, 1 (1975).

[21] P. Claverie and S. Diner, in *Localization and Delocalization in Quantum Chemistry*, edited by O. Chalvet, R. Daudel, S. Diner, and J. P. Malrieu (Reidel, Dordrecht, 1976), Vol. II. This paper contains an enlightening discussion of some aspects of SQM and SED.

[22] J. G. Gilson, Proc. Cambridge Philos. Soc. **64**, 1061 (1968); F. G. Hall and R. E. Collins, J. Math. Phys. **12**, 100 (1971).

[23] A. F. Kracklauer, Phys. Rev. D **10**, 1358 (1974).

[24] E. P. Wigner, Phys. Rev. **40**, 749 (1932).

[25] L. Cohen, J. Math. Phys. **7**, 781 (1966); H. Margenau and L. Cohen, in *Quantum Theory and Reality*, edited by M. Bunge (Springer, Berlin, 1967).

[26] G. della Riccia and N. Wiener, J. Math. Phys. **7**, 1372 (1966).

[27] A similar procedure is used by Surdin, Ref. 8.

[28] See, e.g., K. Huang, *Statistical Mechanics* (Wiley, New York, 1963).

[29] L. de la Peña, R. M. Velasco, and A. M. Cetto, Rev. Mex. Fís. **19**, 193 (1970).

[30] We report a typographical error of sign in Moyal's Eq. (7.9).

[31] See, e.g., P. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953), Vol. I.

[32] E. P. Wigner, in *Perspectives in Quantum Theory*, edited by W. Yourgrau and A. Van der Merwe (MIT Press, Cambridge, Mass., 1971).

[33] C. Piquet, C. R. Acad. Sci. Ser. A **279**, 107 (1974).

[34] N. H. McCoy, Proc. Nat. Acad. Sci. **18**, 674 (1932).

[35] See, e.g., H. Cramér, *Mathematical Methods in Statistics*

(Princeton U.P., Princeton, New Jersey, 1951).

[36] One may find of course other points of discrepancy between SED and QM, which, being of a more interpretative and less formal nature, do not enter into our discussion; several such points of interest are discussed in the paper by Claverie published in the book of Ref. 21.

[37] P. Claverie and S. Diner, C. R. Acad. Sci. Ser B **277**, 579 (1973).

Impedance, zero energy wavefunction, and bound states[*]

André Martin

Cern, Geneva, Switzerland

Pierre C. Sabatier[†]

Département de Physique Mathématique, Université des Sciences & Techniques de Languedoc, 34060-Montpellier Cedex, France

(Received 17 January 1977)

We show that the presence, or absence, of bound states in the three-dimensional Schrödinger equation directly depends on the existence of zeros for a function which is a zero energy solution of the equation and which has the meaning of an impedance in a related equation. Several inequalities that are sufficient to prevent the existence of bound states are obtained from this remark. Some of them are new and bridge the gap between previous results.

Let $Z(x)$ be a positive function, with an absolutely continuous derivative for any real x, and going to a positive constant as $x \to \pm\infty$. It is well known that the operator $Z^{-2}(d/dx)Z^2(d/dx)$, which appears in the transmission line problem, has no "discrete spectrum," [negative eigenvalue, eigenfunction in $L^2(\mathrm{R})$].[1,2] $Z(x)$ inthis operator has the physical meaning of an impedance. Similarly, it has been shown[2,3] that the equation

$$\left[Z^{-2}\frac{d}{dr}Z^2\frac{d}{dr} - l(l+1)r^{-2}\right]\psi(r) = -E\psi(r) \tag{1}$$

has no solution such that $\psi(0) = 0$ and $\psi \in L_2(0,\infty)$, for any value of E. Equation (1) and the transmission line equation are equivalent to Schrödinger equations with the potential $Z^{-1}(d^2/dr^2)Z$, so that these results mean that such a potential has no bound state. In the radial case, this had also been shown directly.[4] A similar result has been proved in a radial coupled-channel case,[5] and the way it was proved (using Picard equations for the vectors made of the solutions and their gradients) can clearly be generalized to a many-channel case. Somewhat related results, concerning the equation $\mathrm{div}[(Z^2(\mathbf{r}) - E)\mathrm{grad}\psi] = 0$, were also obtained in plasma problems.[6] On the other hand, several inequalities that are sufficient to guarantee that a potential has no bound state are known in the literature.[7-9]

In the present paper, we show on the three-dimensional Schrödinger equation without spherical symmetry how the existence of a bound state is related to the impossibility of solving the equation

$$Z(\mathbf{r})\Delta[Z^{-1}(\mathbf{r})] = V(\mathbf{r}) \tag{2}$$

with a positive function $Z(\mathbf{r})$, and we show that some conditions for the absence of bound states are readily related to this property.

1. SUFFICIENT CONDITION FOR NO DISCRETE SPECTRUM OF (1) AND (2)

Let $Z(\mathbf{r})$ be a strictly positive real function, which is twice differentiable for any finite $\mathbf{r}$, and behaves, for $|\mathbf{r}| \to \infty$, in such a way that $Z(\mathbf{r})$ remains positive, Z, Z^{-1}, and $|\mathrm{grad}Z|$ are uniformly bounded, and ΔZ goes to zero more rapidly than $|\mathbf{r}|^{-2}$. Let $W(\mathbf{r})$ be a real nonnegative continuous function, going to zero more rapidly than $|\mathbf{r}|^{-2}$ for $|\mathbf{r}| \to \infty$. We claim that the equations

$$Z^2\mathrm{div}[Z^{-2}\mathrm{grad}\phi(\mathbf{r})] + (E - W)\phi(\mathbf{r}) = 0 \tag{3}$$

and

$$H\psi \equiv [-\Delta + (V + W)]\psi = E\psi \tag{4}$$

where

$$\psi(\mathbf{r}) = Z^{-1}\phi(\mathbf{r}) \tag{5}$$

and $V(\mathbf{r})$ is defined by (2), have no solution of negative energy in the set $\mathcal{E}$ of functions f such that f^2 and $|\mathrm{grad}f|^2$ belong to $L(\mathrm{R}_3)$.

Proof: The equations are shown to be formally equivalent by substituting (5) into (3) or (4). Since Z, Z^{-1}, and $|\mathrm{grad}Z|$ are uniformly bounded, ψ and $\mathrm{grad}\psi$ belong to $L_2(\mathrm{R}_3)$ if ϕ and $\mathrm{grad}\phi$ do, and conversely. Now consider the following equality, which readily follows from (3):

$$\begin{aligned} I &= E\iiint Z^{-2}\phi^2(\mathbf{r},E)d\mathbf{r} \\ &= \iiint Z^{-2}W(\mathbf{r})\phi^2(\mathbf{r},E)d\mathbf{r} \\ &\quad - \iiint \phi(\mathbf{r},E)\mathrm{div}[Z^{-2}\mathrm{grad}\phi(\mathbf{r},E)]d\mathbf{r}. \end{aligned} \tag{6}$$

Notice in passing that this equality expresses the energy conservation in the time-dependent form of (3). Elementary transformations, and the Gauss theorem, yield

$$\begin{aligned} I = &\iiint Z^{-2}W(\mathbf{r})\phi^2(\mathbf{r},E)d\mathbf{r} + \iiint Z^{-2}|\mathrm{grad}\phi|^2 d\mathbf{r} \\ &- \lim_{S\to\infty}\int\int_S Z^{-2}\phi\,\mathrm{grad}\phi.d\mathbf{S}. \end{aligned} \tag{7}$$

The limit in the last term is zero because $\phi|\mathrm{grad}\phi|$ belongs to $L(\mathrm{R}_3)$. Thus I reduces to the first two terms in the right-hand side of (7). These terms are positive or zero (0 iff $\phi \equiv 0$). I therefore is positive and thus E is also, which contradicts the assumption. QED
Notice that the result holds for Eq. (3) even if Z is unbounded above.

2. SUFFICIENT CONDITION FOR NO DISCRETE SPECTRUM OF (2)

Setting $F(\mathbf{r}) = [Z(\mathbf{r})]^{-1}$, we see that (2) is equivalent to the zero energy Schrödinger equation corresponding to the potential $V(\mathbf{r})$

$$\Delta F(\mathbf{r}) = V(\mathbf{r})F(\mathbf{r}). \tag{8}$$

Thus, for the Schrödinger equation (4), the result stated in Sec. 1 is equivalent to the nonexistence of a negative-energy bound state when a zero energy solution is strictly positive. We can directly show a more general result. If $V(\mathbf{r})$ is a continuous function, and goes to zero more rapidly that $|\mathbf{r}|^{-2}$ as $|\mathbf{r}| \to \infty$, it is well known that the wavefunction $\psi_0(\mathbf{r})$ that corresponds to the ground state has no zero, and decreases exponentially for large $|\mathbf{r}|$, together with its gradient. Thus we get from (4) and (8), using the Gauss theorem,

$$\begin{aligned}&\lim_{S\to\infty} \int\int_S (F\,\mathrm{grad}\psi_0 - \psi_0\mathrm{grad}F).d\mathrm{S}\\ &\quad = 0\\ &\quad = \int\int\int (E - W(\mathbf{r}))\psi_0(\mathbf{r})F(\mathbf{r})d\mathbf{r}\end{aligned} \tag{9}$$

and this is impossible for negative E and nonnegative W if there exists a solution F of (8), which keeps a constant sign, and remains bounded, or diverges less than exponentially as $|r| \to \infty$. Thus one gets here a nonnegativity condition instead of positivity; extremas, at which F and grad$F$ vanish, are allowed. But clearly, for a regular potential, these points cannot be too many (in particular, they cannot form a closed surface).

3. NECESSARY CONDITION FOR NO DISCRETE SPECTRUM OF (2)

We assume in the following that $V+W$ in Eq. (4) is continuous and goes to zero faster than $|\mathbf{r}|^{-2}$ as $|\mathbf{r}| \to \infty$. Now let us suppose we know a function $F(\mathbf{r})$ which

(a) is a continuously differentiable solution of Eq. (4) for the zero energy.

(b) is such that Fgrad$F$ is zero on a surface S enclosing a simply-connected finite volume $\mathcal{V}$.

(c) is such that grad$F\cdot$grad$F$ is bounded and strictly positive on a certain nonvanishing area $\Delta S \in S$.

Then we claim that H has indeed a negative energy bound state.

Proof: Since HF is zero, the integral on $\mathcal{V}$ of FHF is zero. Using (b), the Gauss theorem and some elementary vectorial algebra, we easily derive the equality

$$\int_{\mathcal{V}} FHF = 0 = \int_{\mathcal{V}} \mathrm{grad}F\,.\,\mathrm{grad}F d\mathbf{r} + \int_{\mathcal{V}} (W+V)F^2 d\mathbf{r}. \tag{10}$$

Now let us recall that H has a bound state if and only if[10] the minimum of $E(f) = \langle f|H|f\rangle / \langle f|f\rangle$ on a set of functions such that f and Hf belong to $L_2(\mathrm{R}_3)$ is negative. This minimum is then equal to the energy of the ground state. From (10) we know that if f is equal to

$$f = \overline{F} = \begin{cases} F(\mathbf{r}) & (\mathbf{r} \in \mathcal{V}),\\ 0 & (\mathbf{r} \notin \mathcal{V}),\end{cases} \tag{11}$$

then $E(f)$ is zero. From the assumptions (b) and (c) we know that F should be zero on ΔS. Since F is continuously differentiable, grad$F$ is a continuous function. Hence, there should exist in $\mathcal{V}$ a volume $\Delta\mathcal{V}$, containing ΔS, in which $|\mathrm{grad}F|$ is bounded below by a positive number, say, m, and bounded above by, say, M. Let dS be a part of ΔS and $d\mathcal{V}$ a part of $\Delta\mathcal{V}$ containing dS, and contained in a ball of diameter δ. Now F is zero in a part of $d\mathcal{V}$, and is differentiable, so that, for any point $\mathbf{d}$ of dS, and any point $\mathbf{r}$ of $d\mathcal{V}$, we can write

$$F(\mathbf{r}) = \int_{\mathbf{d}}^{\mathbf{r}} \mathrm{grad}F\,.d\mathbf{s}, \tag{12}$$

where the integral is taken on the straight line, of differential element $d\mathbf{s}$, going from $\mathbf{d}$ in dS to $\mathbf{r}$. Thus, the upper bound M of grad$F$ readily yields

$$|F(\mathbf{r})| < M\delta \quad (\mathbf{r} \in d\mathcal{V}). \tag{13}$$

Let us now use for f in the functional $E(f)$ a function F_δ which is equal to $\overline{F}$ for any $r \in \mathcal{V}$ except in $d\mathcal{V}$, where F_δ and its gradient should be taken equal to zero. We easily see that

$$\int_{\mathcal{V}} F_\delta H F_\delta = \frac{-\int_{d\mathcal{V}}\{|\mathrm{grad}F|^2 + (V+W)F^2\}d\mathbf{r}}{\int_{\mathcal{V}} F^2 d\mathbf{r} - \int_{d\mathcal{V}} F^2 d\mathbf{r}}. \tag{14}$$

The first term in the numerator is smaller than $-m^2\int_{d\mathcal{V}} d\mathbf{r}$ and the second one is smaller than $\delta^2 M^2 \int_{d\mathcal{V}} V(\mathbf{r})d\mathbf{r}$. The first term in the denominator does not depend on the size of $d\mathcal{V}$ but the second one is smaller than $\delta^2M^2 \int_{d\mathcal{V}} d\mathbf{r}$. We can always choose $d\mathcal{V}$, and thus δ, in such a way that the first terms in the numerator and denominator are dominant, so that $\int_{\mathcal{V}} F_\delta H F_\delta$ is strictly negative. This proves our point.

The result can be extended to infinite domains if ψ and gradψ are L_2 in the infinite directions. However, if $\mathcal{V}$ goes over to R_3, it is necessary to make the additional assumption that there is, at finite distance, at least a point at which ψ is zero and $|\mathrm{grad}\psi|$ is strictly positive and bounded.

One can also notice that in a situation where both F and grad$F$ are zero on the closed surface S, it is possible to continue F outside of $\mathcal{V}$ by zero, so that there would exist for any potential equal to $V+W$ inside $\mathcal{V}$, and arbitrarily continued outside of $\mathcal{V}$, a zero energy bound state. This is impossible.

4. APPLICATIONS

Since the presence, or absence, of bound states, is narrowly related to the sign of a continuous solution of Eq. (8), it is interesting at least for pedagogical reasons to derive a condition that is sufficient to guarantee a solution of constant sign. Now considering Eq. (8) as a Poisson equation, we see that the solution, if it exists, of the following integral equation:

$$F(\mathbf{r}) = 1 - (4\pi)^{-1} \int\int\int V(\rho)F(\rho)|\mathbf{r}-\rho|^{-1}d\rho \tag{15}$$

is also a solution of Eq. (8).

Since $W(\rho)$ is simply required to be nonnegative, we can always assume that $V(\rho)$ is everywhere nonpositive. Thus (15) has certainly a solution which is everywhere greater than or equal to one if the successive approximations algorithm corresponding to Eq. (15) converges.[11] It is equivalent to consider the equation

$$\begin{aligned}\delta F(\mathbf{r}) = {}&(4\pi)^{-1}\int\int\int |V(\rho)|\ |\mathbf{r}-\rho|^{-1}d\rho\\ &+ (4\pi)^{-1}\int\int\int |V(\rho)|\ |\mathbf{r}-\rho|^{-1}\delta F(\rho)d\rho.\end{aligned} \tag{16}$$

Recall that in a Banach space, β, a contracting mapping has one fixed point, which is given by the successive approximations algorithm, starting, for instance, at 0. Thus we only have to check that the first term in the right-hand side of (16) belongs to $\mathcal{B}$, and that the operator is a contraction. Throughout the four examples we give, we shall assume that $V(\mathbf{r})$ is continuous and goes to zero faster than $|\mathbf{r}|^{-2}$ as $|\mathbf{r}| \to \infty$.

A. 1st example

For a certain number $\epsilon \in]0,1[$, let $\mathcal{B}$ be the set of functions f such that $\int\int\int |\mathbf{r}|^{-2-\epsilon}|f(\mathbf{r})|\,d\mathbf{r}$ is finite. A simple calculation gives

$$\int |\mathbf{r}|^{-2-\epsilon}|\mathbf{r}-\rho|^{-1}d\mathbf{r} = 4\pi\,|\rho|^{-\epsilon}\left[\frac{1}{\epsilon}+\frac{1}{1-\epsilon}\right]. \tag{17}$$

Thus the first term in (16) belongs to $\mathcal{B}$ (it would suffice that $\int\int\int |V(\rho)|\,|\rho|^{-\epsilon}d\rho < \infty$). The contraction condition is

$$(4\pi)^{-1}\sup_{\rho}|V(\rho)|\;\;|\rho|^{2+\epsilon}\int|\mathbf{r}|^{-2-\epsilon}|\mathbf{r}-\rho|^{-1}d\mathbf{r}<1$$

or

$$\sup_{\rho}|\rho|^{2}|V(\rho)|<\left(\frac{1}{\epsilon}+\frac{1}{1-\epsilon}\right)^{-1}$$

or, for the best choice of ϵ, i.e., $\epsilon=\frac{1}{2}$,

$$\sup_{\rho}|\rho|^{2}|V(\rho)|<\tfrac{1}{4}.$$

Notice that in this inequality, which is well known, one can choose the center of coordinates in the most convenient way.

B. 2nd example

For the sake of simplicity, we assume here in addition that $|V(\mathbf{r})|$ decreases more rapidly than $|\mathbf{r}|^{-5/2-\epsilon}$ as $|\mathbf{r}|$ goes to ∞ ($\epsilon>0$), and that $|V(\mathbf{r})|$ is almost everywhere positive. We define $\mathcal{B}$ as the space of functions such that $\int\int\int|V(\mathbf{r})|\mathbf{f}^2(\mathbf{r})$ is finite. It is easy to check that the first term belongs to $\mathcal{B}$. The contraction condition is

$$(4\pi)^{-2}\int\int\int|V(\mathbf{r})|d\mathbf{r}\int|V(\rho)|\;\;|\mathbf{r}-\rho|^{-2}d\rho<1. \tag{18}$$

Using the Hardy–Littlewood–Sobolev inequality[12] we obtain as well

$$\int d\mathbf{r}\,|V(\mathbf{r})|^{3/2}<(32\pi)^{1/2}\sim 10.02. \tag{19}$$

Notice that both inequalities (17) and (18) do not depend on the center of coordinates and that the inequality given by Glaser *et al.* is better than (19) (it would lead to 12.82 in the right-hand side).

C. 3rd example

Let $\mathcal{B}$ be the space of continuous functions f, bounded at ∞, with the norm $\|f\|=\sup_{\rho}|f(\rho)|$. The contraction condition reads

$$(4\pi)^{-1}\sup_{\mathbf{r}}\int\int\int|V(\rho)|\;\;|\mathbf{r}-\rho|^{-1}d\rho<1 \tag{20}$$

and clearly this condition also guarantees that the first term belongs to $\mathcal{B}$. This condition, in which one should recall that $V(\rho)$ is the attractive part of the potential, was given by Hunziker[13] as a convergence condition for the Born series. It does not depend on the center of coordinates. In the case of spherically symmetric potential, it reduces to the Bargmann condition $\int_0^\infty \rho|V^-(\rho)|d\rho<1$, and thus is a best condition. We must realize however that condition (20) is more general than Bargmann's condition. Furthermore, condition (20), for the nonspherical case, is *better* than the one obtained by replacing $|V(\rho)|$ by a spherically symmetric upper bound.

D. 4th example

Let $\mathcal{B}$ be the space of functions of f such that, for a certain positive number ϵ, to be defined later, $|\mathbf{r}|^{\epsilon}f(\mathbf{r})$ is continuous on any compact set in R_3 and bounded at ∞, with the norm

$$\|f\|=\sup_{\rho}|\rho|^{\epsilon}|f(\rho)|.$$

$\mathcal{B}$ is complete, and the contraction condition reads

$$\sup_{\mathbf{r}}|\mathbf{r}|^{\epsilon}\int\int\int|V(\rho)|\;\;|\mathbf{r}-\rho|^{-1}|\rho|^{-\epsilon}d\rho<4\pi. \tag{21}$$

For $p>\frac{3}{2}$, it is possible to write the integrand in (20) as a product of two factors, the first one being $|V(\rho)|\times|\rho|^{2-3/p}$. Then, using Holder's inequality, one obtains results of the form given in Ref. (8), but these are not so good, and the method fails as $p\to\frac{3}{2}$. For $1\le p<\frac{3}{2}$, we take as a first factor $|V(\rho)|\;\;|\mathbf{r}-\rho|^{[2p-3-(p-1)s]/p}$. The Holder inequality then yields, instead of (20), the sufficient condition

$$\sup_{\mathbf{r}}\left\{\int\int\int|V(\rho)|^{p}|\mathbf{r}-\rho|^{2p-3-(p-1)s}|\rho|^{(p-1)s}d\rho\right\}$$
$$<(4\pi)^{p}\left\{\int\int\int\big|\mathbf{r}/|\mathbf{r}|-\mathbf{x}\big|^{-3+s}|\mathbf{x}|^{-s-\mu}d\mathbf{x}\right\}^{(1-p)}, \tag{22}$$

where $\mu=\epsilon p/(p-1)$, s and μ can be chosen arbitrarily in the set $s>0$, $(s+\mu)<3$, $\mu>0$. The right-hand side of (22) is equal to a number

$$(4\pi)^{-1}\left\{\frac{\Gamma(s+\mu-1)}{\Gamma(s-1)\Gamma(\mu-1)}\left[\tan\frac{\pi s}{2}\tan\frac{\pi\mu}{2}-1\right]\right\}^{p-1}. \tag{23}$$

The inequalities (22) obviously reduce to (20) for $p=1$. For $\epsilon>0$, and providing the potential is bounded at ∞ by $|\mathbf{r}|^{-2-\epsilon}$, the first term in the rhs of (16) belongs to $\mathcal{B}$. Besides, because of the continuity of V, it is easy to see that the solution of (16) which is obtained in $\mathcal{B}$ is actually continuous even at $|\mathbf{r}|=0$.

Our inequalities (22) and (23) *bridge the gap between* $p=1$ *and* $p=\frac{3}{2}$ in the inequalities of Ref. (8), since they hold here in the general case (no symmetry). In the referred paper, inequalities are given in this gap only in the special case of spherically symmetric potential. It is interesting to see that, in this last case and for the choice $s=2$, $\mu=\frac{1}{2}$ ($\mathbf{r}=0$) gives the supremum in (22), so that (22) reduces to

$$\int\int\int|V(\rho)|^{p}|\rho|^{2p-3}d\rho\le(4\pi)^{-1}4^{1-p}. \tag{24}$$

The inequality (24) is not optimal, as can be seen by comparing it to the corresponding one in Ref. (8). For the general case one may prefer choosing $s=1$, $\mu=1$, which gives

$$\sup_r \{ \} \leq (4\pi)^{-1}[\pi^2/4]^{1-p}. \qquad (25)$$

Clearly all these inequalities are simply examples of what can be obtained. Results of Secs. 1, 2, 3 can be extended, with some simple modifications, to spaces with dimension not equal to 3, but results of Sec. 4 cannot, because they are related to the sign of the Green's function in (15). In the one-dimensional case, one rather shows that a purely attractive potential indeed has a bound state.

ACKNOWLEDGMENTS

We are indebted to Dr. Chadan, Dr. Jaulent, and Dr. Miodek for interesting discussions and remarks.

*This work has been done as part of the program "Recherche Coopérative sur Programme no. 264: Etude interdisciplinaire des problèmes inverses."

†Physique Mathématique et Théorique, Equipe de recherche associée au CNRS no. 154.

[1] See R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1953), Chap. V, Sec. 3 (vibrating strings). In the case of transmission lines, the physical parameter is the square of the frequency. The stated result actually follows from the uniqueness proofs of D. S. Heim and C. B. Sharpe in IEEE Trans. Circuit Theory **14**, 394 (1967). It was stated, with a simple physical argument, by K. Aki and J.A. Ware, in J. Acoust. Soc. Am. **45**, 911 (1969).
See also I. Kay, in "Mathematics of Profile Inversion," NASA TMX-62, 150 (1972), p. 6.7.

[2] A discussion of the effects of positivity constraints appears in P. C. Sabatier, "On geophysical inverse problems and constraints," Sec. 2, to be published in the Z. Geophys.

[3] P. C. Sabatier, C. R. Acad. Sci. Paris Ser. B **278**, 545 (1974).

[4] K. Chadan and A. Montes, Phys. Rev. **164**, 1762 (1967), Sec. II.

[5] P. C. Sabatier, C. R. Acad. Sci. Paris Ser. B **278**, 603 (1974).

[6] E. M. Barston, Ann. Phys. **29**, 282 (1964).

[7] V. Bargmann, Rev. Mod. Phys. **21**, 488 (1949).

[8] V. Glaser and A. Martin, H. Grosse and W. Thirring, in *Studies in Mathematical Physics, Essays in Honor of V. Bargmann*, edited by E. H. Lieb, B. Simon, and A. S. Wightman (Princeton U.P., Princeton, N. J., 1976), p. 169 and ff.

[9] B. Simon, p. 305 and ff. in Ref. 8.

[10] R. Courant and D. Hilbert (in Ref. 1), Chap. VI, Sec. 1.

[11] That the absence of bound states is connected to the convergence of the Born series is of course not new, but usually, one needs the convergence for all values of the energy. Here, we stay at zero energy.

[12] See, for example, A. Martin, CERN Preprint Th 2085, especially the Appendix (published in Proceedings of RCP no. 25, Strasbourg, 1976).

[13] W. Hunziker, Helv. Phys. Acta **34**, 593 (1961).

Varieties of symmetry breaking in a class of gauge theories

Amir Schorr

International Centre for Theoretical Physics, Trieste, Italy

(Received 20 September 1976)

In unified gauge theories, the Higgs particles can interact in various ways. The problem of finding the symmetry-breaking directions can become very complicated in nontrivial cases, where the scalar fields have many interactions. A method is presented which predicts, in a simple way, the possible types of spontaneous symmetry breaking in a theory symmetric under the group $\mathrm{U}(N_1)\otimes\cdots\otimes\mathrm{U}(N_j)\otimes\mathrm{SU}(M_1)\otimes\cdots\otimes\mathrm{SU}(M_k)$. Within its framework it is possible to obtain results by drawing a new kind of graph. It is found that in such models, various phases (and hence phase transitions) are possible. There are distinct hierarchies in the symmetry breaking strengths and they are related.

I. INTRODUCTION

After it was proved that non-Abelian gauge theories can be renormalizable and that vector mesons can acquire mass[1] without losing their renormalizability, the way towards the unification of all interactions in a Yang—Mills type of theory was opened. Salam[2] and Weinberg[3] were the first, with a model unifying electromagnetic and weak interactions. The possibility of asymptotic freedom[4] encouraged others to include strong interactions in this program. Such unification could be based on one simple group[5] (and hence restricted to only one gauge coupling constant) or, alternatively, on a direct product of simple groups.[6] It seems to us that the latter alternative followed, in particular, by Pati and Salam[6] can be more predictive in the present domain of energies. Hence we present a technique for investigating the character of its Higgs potential.

In earlier work the Higgs particles were treated merely as an instrument for obtaining specific dynamics. The surviving "physical" scalars were expected to disappear when the theory became better developed. Now it seems that one may have to live with them. Usually one can arrange for some of the scalars to "disappear" by acquiring superheavy masses. Often, however, a number of relatively light (pseudo-Goldstone) scalars remain and must be taken seriously. (In such theories asymptotic freedom[7]) may have to be sacrificed.) A phenomenological investigation of the Salam—Weinberg scalar was made using principally its dilation character.[8] A lower limit on the scalar mass was established[9] by appealing to radiative mechanisms for symmetry restoration. Accepting their possible independent existence, one can give these scalars "color", by which we mean letting them feel more than one kind of interaction.

The purpose of this article is to present a mathematical technique for seeking out the minima of the classical potential, a quadratic polynomial in the scalar fields. This problem seemed to be a hard one to solve by conventional methods when the theory contains many kinds of Higgs fields, each one having more than one type of interaction. The article of Ling-Fong Li[10] (the only one which deals in general terms with this subject) treats some relatively simple cases.

In the technique which we present here, one finds, step by step, the structure of the symmetry breaking considering qualitative features of the interactions. Repulsive or attractive forces lead to different breakings, where the relations between the strengths of these forces determine the subgroups which remain unbroken. Our method uses in each step the consequence of the previous step. The technique proceeds by eliminating any variables which can be fixed, and hence need not be given any further consideration. In Secs. II and III we present the Higgs fields $\{A^{\delta}_{j,k}, A^{\delta *}_{j,k}\}^{\delta}_{j,k}$ and show that one needs to consider only the diagonal elements of $\{A^{\delta}A^{\delta *}\}^{\delta}$. In Sec. IV we give two conditions which allow us to use the original simple form of the potential, while obtaining results in a subspace which is obtained by fixing some of the variables. In this way we are not obliged to consider the complicated form of the potential in this subspace. In Secs. V and VII a graphical method of representing the problem is presented. These graphs are most crucial to our method. In many cases the drawing of such a graph is quite easy and can save one from complicated computations. From the topology of the final graph, the symmetry breaking is read instantly. Sections VI and VIII are dedicated to explaining the procedure, which is mainly based on finding the maximum space which has a positive definite metric, and hence contains a minimum point.

It appears that, contrary to the relatively boring situation produced by "grey" kinds of Higgs fields which are governed by a simple gauge symmetry, the colored Higgs theory based on $\mathrm{U}(N_1)\otimes\mathrm{U}(N_2)\otimes\cdots$, possesses a wide range of possibilities. The vacuum can be broken in various directions and with various strengths, and among the interesting consequences are the specific relations between the breakings of the different groups. We find that the transitions between different phases can be sharp (first order) or smooth (second order).

These classical results should be treated carefully as it was shown[9,11] that higher-order calculations including the interactions of the scalars with fermions and vectors, can completely change the structure of the symmetry breaking seen in the tree approximation. Moreover, it was shown that the symmetry restoration can be achieved by changing temperature,[12] density,[13] and external fields.[14] It seems to us that, taking into acccount all these effects, it might be constructive to present a theory where limited confinement can occur through phases which have the threshold of a first-order phase transition. The many-phase situation which arise in these "colored Higgs" models may lead to phenomena where different environments favor different phases.

Apart from that, we found that the usual hierarchy of symmetry-breaking can arise by two mechanisms. The first is characterized by degenerate minima of the classical potential, which leaves to higher-order contributions the decision as to the fine structure of the breaking. The second mechanism arises by the "coloredness" of the Higgs particles whereby a particular breaking in one group enforces a hiearchy of breakings in the others. The presentation of these features will occupy a series of articles. In the first one we present the general method of finding, in a simple way, the possible directions of the breakings. The second article will illustrate the method, describing in detail the situation in $U(N)\otimes U(N)\otimes U(N)$, and will include an examination of the symmetry breaking in the original Pati–Salam model. The third paper will represent physical features of these "colored" theories by presenting "improved" Pati–Salam models, and then investigate the possible hierarchy phase and phase transitions which can arise in these models.

II. DIAGONALIZATION

We are interested in a renormalizable invariant theory of the group $G=S^{N_1}\otimes S^{N_2}\otimes S^{N_3}\cdots\otimes S^{N_\kappa}$, where S^{N_δ} is either $SU(N_\delta)$ or $U(N_\delta)$. We have κ multiplets of Higgs–Kibble fields $A_\delta(\delta=1,\ldots,\kappa)$ where A_δ transforms under G as

$$\underset{(\kappa-\delta)}{(1,1,\ldots,1,}N_\delta,\overline{N}_{\delta-1},\underset{(\delta-2)}{1,\ldots,1,1)}$$

and $A_\delta^\dagger$ transforms correspondingly (Fig. G2-1). The Lagrangian has the conventional form[6] and contains the potential $V\{A^\delta,A^{\delta\dagger}\}$, which describes the interactions between these fields. We demand that V should be invariant under the whole group G and under the κ discrete transformations:

$$T^\delta: A_\delta=-A_\delta \quad (\delta=1,\ldots,\kappa)$$

$$T^\delta: A_j=A_j \quad (\delta\neq j). \tag{2.1}$$

Since we want our theory to be renormalizable, V includes only four field interactions and "mass terms"

$$V=\sum_{\delta=1}^{\kappa}\mu_\delta\,\mathrm{Tr}(A_\delta A_\delta^\dagger)+\sum_{\delta,\beta}^{\kappa}a_{\delta,\beta}\,\mathrm{Tr}(A_\delta A_\delta^\dagger)\,\mathrm{Tr}(A_\beta A_\beta^\dagger)$$

$$+\sum_{\delta=1}^{\kappa}[\alpha_\delta\,\mathrm{tr}(A_\delta A_\delta^\dagger A_\delta A_\delta^\dagger)+2\gamma_\delta^{\delta+1}\,\mathrm{Tr}(A_\delta A_\delta^\dagger A_{\delta+1}^\dagger A_{\delta+1})]. \tag{2.2}$$

It is clear that for $N_\delta=2,4$, other kinds of interactions might be added (e.g., like $\det A_\delta A_\beta$, etc.); but as we are not going to consider here such terms, it is demanded that the corresponding S^{N_δ} be $U(N_\delta)$ and not $SU(N_\delta)$, in order to avoid having such interactions appear as counterterms in high-order calculations.

Our aim is to find the minimum of the effective potential for the classical part of A_δ. We shall work at first only in the tree approximation and hopefully, get a few hints as to where and how higher order contributions can change the situation. The above potential should have a classical minimum for each of the Higgs fields, hence, we obtain the following relation (from now on A_δ is treated as the classical part of the Higgs fields) $(\delta=1,\ldots,\kappa)$:

$$\left.\frac{\partial V}{\partial(A_\delta^\dagger)^T}\right|_{C^\delta}=[\mu_\delta+2\sum_{\beta=1}^{\kappa}a_{\delta,\beta}\,\mathrm{Tr}(C_\beta C_\beta^\dagger)]\,C_\delta$$

$$+2\alpha_\delta C_\delta C_\delta^\dagger C_\delta+2\gamma_\delta^{\delta+1}C_{\delta+1}^\dagger C_\delta+2\gamma_{\delta-1}^{\delta}C_\delta C_{\delta-1}C_{\delta-1}^\dagger$$

$$=0,\quad \langle A_{j,m}^\delta\rangle=C_{j,m}^\delta. \tag{2.3}$$

The κ matrices $D_\delta=C_\delta C_\delta^\dagger$ are hermitian and transform under the S^{N_δ} group transformations only. Hence we can make κ independent gauge transformations to diagonalize simultaneously all the D_δ matrices. All the elements $D_{j,m}^\delta$ are real and non-negative, thus it makes sense to use the remaining gauge freedom to order their central diagonal,[15]

$$D_{1,1}^\delta\geq D_{2,2}^\delta\geq D_{3,3}^\delta\cdots\geq D_{N_\delta,N_\delta}^\delta\geq 0. \tag{2.4}$$

Our next step is to learn the structure of $G_\delta=C_\delta^\dagger C_\delta$. We multiply each of the κ equations (2.3) by the corresponding $C_\delta^\dagger$, and get for each δ

$$[\mu_\delta+2\sum_\beta a_{\delta,\beta}\rho_\beta]\,D_\delta+2\alpha_\delta D_\delta^2$$

$$+2\gamma_\delta^{\delta+1}G_{\delta+1}D_\delta+2\gamma_{\delta-1}^{\delta}C_\delta D_{\delta-1}C_\delta^\dagger=0\quad(\rho^\delta=\mathrm{tr}D_\delta=\mathrm{tr}G_\delta). \tag{2.5}$$

Taking the Hermitian conjugate of these equations, and then subtracting it from the original equation, we obtain

$$[D_\delta,G_{\delta+1}]=0 \qquad (\delta=1,\ldots,\kappa). \tag{2.6}$$

Hence, another set of κ independent gauge transformations which leaves all the D_δ invariant and diagonalizes all the matrices G_δ, can be carried out. This is a consequence of the structure of the matrices D_δ (2.4). Each D_δ is constructed from m_δ scalar matrices, e.g.,

$$D_\delta=\begin{bmatrix} D_1^\delta & & & & \\ & D_2^\delta & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & D_{m_\delta}^\delta \end{bmatrix},\quad D_j^\delta=d_j^\delta\cdot I.$$

Thus, from (2.6) one concludes that the nonzero elements of $G_{\delta+1}$ are concentrated in the equivalent diagonal

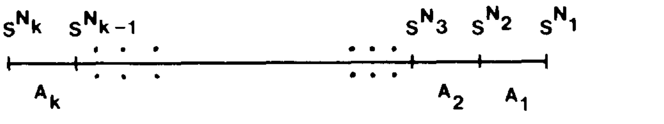


FIG. G2-1. Transformation properties of the Higgs fields.

matrices only,

$$\begin{pmatrix} G_1^{\delta+1} & & & & \\ & G_2^{\delta+1} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \; G_{m_\delta}^{\delta+1} \end{pmatrix}.$$

Hence we can diagonalize it by a gauge transformation ($G_j^{\delta+1}$ is Hermitian), while D_j^δ remains unchanged. From the knowledge that $C_\delta C_\delta^+$ and $C_\delta^+ C_\delta$ are diagonal, it follows that D_δ and G_δ have the same nonvanishing elements in their diagonal, except that the descending order of the elements in D_δ is lost in G_δ; this can be proven by multiplying $C_\delta^+ C_\delta = G_\delta$ by C_δ from the left which results in

$$D_\delta C_\delta = C_\delta G_\delta. \tag{2.7}$$

So for every $C_{j,k}^\delta \neq 0$ we obtain

$$D_{j,j}^\delta = G_{k,k}^\delta. \tag{2.8}$$

By choosing proper nonvanishing terms in C_δ, this correspondence proves to be one-to-one. It should be stated that the number n_δ of such nonvanishing elements is at most the lesser of N_δ and $N_{\delta-1}$ (for this purpose, we work in a κ modulo world; $N_{\kappa+1} = N_1$).

III. STRUCTURE IN SUBSPACES

The main problem begins only now, when one wants to find out from Eqs. (2.5) the set of $\{G_\delta, D_\delta\}$. But for every structure of G_δ we obtain a different kind of Eq. (2.5), each one giving us another extremum point, and we have no idea which is the real minimum. By "structure" we mean here those d_j^δ which are zero (where the case in which none of them vanished is trivial). So we must change our strategy and use the character of the potential V to find out in which form D_δ and G_δ would give us a global minimum.

The first row in the expression (2.2) for V is a function of $\rho_\delta = \mathrm{Tr}(A_\delta A_\delta^+)$ only, while the second one is dependent on the structure of A_δ. We shall indicate this second row by V_3. The minimum point of V is a global minimum point in an open space, and it should also be a real global minimum for any closed subspace which includes it. The subspace in which we are interested is defined by ($\delta = 1, \ldots, \kappa$)

$$\mathrm{Tr}(A_\delta A_\delta^+) = \langle \rho_\delta \rangle = \rho^\delta. \tag{3.1}$$

Until $\langle \rho \rangle$ is found we work in a subspace in which ρ is a constant. In this subspace, the first row of V in (2.2) is a constant, so we can concentrate on the second row V_3. Moreover, we limit our subspace to that in which the matrices $(A_\delta A_\delta^+)$ and $(A_\delta^+ A_\delta)$ are diagonal and (for each δ) comprised of the same nonvanishing elements *except for order*. We shall search for the minimum of V_3 in this subspace, as we already know that it includes the global minimum point which has this structure. As a first investigation, we examine the matrices $\langle A_\delta^+ A_\delta \rangle$ for the purpose of finding out their form as a function of the $\langle A_\delta A_\delta^+ \rangle$ matrices. We denote the diagonal elements $(A_\delta A_\delta^+)_{jj}$ by x_j^δ, and the whole diagonal of $(A_\delta A_\delta^+)$ by the vector $\mathbf{x}^\delta$, and similarly the diagonal of $A_\delta^+ A_\delta$ by $\mathbf{x}'^\delta$. In both there are n_δ nonzero components, and the vector $\mathbf{x}^\delta$ differs from $\mathbf{x}'^\delta$ in the order of its components and in the number of zeroes [according to the different dimensions of $A_\delta^+ A_\delta$ $(N_{\delta-1})$ and $A_\delta A_\delta^+ (N_\delta)$]. The second line of the potential can be written in this subspace as in (3.2), where from now on we shall seek the minimum of V_3 as a function of the $\mathbf{x}^\delta$, $\mathbf{x}'^\delta$,

$$V_3(\mathbf{x}^\delta, \mathbf{x}'^\delta) = \sum_j (\alpha_\delta \mathbf{x}^\delta \mathbf{x}^\delta + 2\gamma_\delta^{\delta+1} \mathbf{x}^\delta \mathbf{x}'^{\delta+1}). \tag{3.2}$$

We observe that the first term is independent of the "ordering" of $\mathbf{x}'^\delta$, so we consider for the moment the second term only.

We now take advantage in the ordered form (2.4) of the $\langle \mathbf{x}^\delta \rangle$ components, i.e.,

$$(\langle x_j^\delta \rangle) \geqslant (\langle x_{j+1}^\delta \rangle) \geqslant 0.$$

In this gauge it is clear that the $\langle \mathbf{x}'^\delta \rangle$ components, at the global minimum point of $V_3(\mathbf{x}^\delta, \mathbf{x}'^\delta)$, can be gauged to the form

$$(\langle x_{j+1}'^\delta \rangle) \geqslant (\langle x_j'^\delta \rangle) \geqslant 0 \tag{3.3}$$

if $\gamma_{\delta-1}^\delta < 0$. In such a case we denote each $\mathbf{x}'^\delta$ by $\mathbf{x}^{+\delta}$. In the opposite case, we have $\gamma_{\delta-1}^\delta > 0$ and hence

$$(\langle x_j'^\delta \rangle) \geqslant (\langle x_{j+1}'^\delta \rangle) \geqslant 0, \tag{3.4}$$

and we denote $\mathbf{x}'^\delta$ by $\mathbf{x}^{-\delta}$.[16] It should be noted that we are not interested at present in the degeneracy of the minimum of the potential. Rather, we shall be satisfied to find only one point in it. We shall limit our subspace again by demanding that all the nonzero components of the $\mathbf{x}^\delta$ concentrate on the lowest indexes, and the components of $\mathbf{x}'^\delta$ be equal to those of $\mathbf{x}^\delta$, except for the order, which is fixed according to the rules (3.3) and (3.4). The ordering of the vectors $\langle \mathbf{x}^\delta \rangle$ by a gauge transformation gives us the possibility of expressing the $\mathbf{x}'^\delta$ components in terms of the $\mathbf{x}^\delta$ variables which form the space R_N*,

$$\mathbf{x}^\delta = (x_1^\delta, x_2^\delta, \ldots, x_{n_\delta}^\delta, 0, 0, \ldots, 0)$$

$$\mathbf{x}'^\delta = \begin{matrix} (x_1^\delta, x_2^\delta, \ldots, x_{n_\delta}^\delta, 0, 0, \ldots, 0), & \gamma_{\delta-1}^\delta \leqslant 0, \\ (0, 0, \ldots, 0, x_{n_\delta}^\delta, x^\delta_{(n_\delta-1)}, \ldots, x_1), & \gamma_{\delta-1}^\delta > 0, \end{matrix}$$

$$(\delta = 1, \ldots, \kappa). \tag{3.5}$$

Since the global minimum point has this structure (2.4), we should find it as the minimum in this subspace.

The problem as stated now is to find the minimum point of V_3 inside the space R_N* of the free variables $\{x_j^\delta\}$ ($\delta = 1, \ldots, \kappa$) ($j = 1, \ldots, n_\delta$). The space R_N in which we are interested is the plane in R_N*, which is fixed by a constant ρ,

$$\sum_{j=1}^{n_\delta} x_j^\delta = \rho^\delta. \tag{3.6}$$

The minimum point $\langle \mathbf{x} \rangle$ should be inside the closed "pyramid" $R_{\bar{N}}$, defined by the "positivity condition" in R_N,

$$x_j^\delta \geqslant 0. \tag{3.7}$$

Symbolically this chain takes the form

$$R_N*\{\mathbf{x}^\delta\} \underset{(\sum \mathbf{x}=\rho)}{\supset} R_N \underset{(x\geqslant 0)}{\supset} R_{\bar{N}}. \tag{3.8}$$

$R_{\bar{N}}$ is a closed subspace, hence there is always a minimum point of V_3 in it. But, it might happen that the minimum point of V_3 in R_N is outside of $R_{\bar{N}}$, or there is no minimum point in R_N at all. In such cases $\langle\mathbf{x}\rangle_{\text{VEV}}$ is on the border of $R_{\bar{N}}$, which means that some of its components should vanish. As a result one can reduce the number of free variables by fixing some of the $\{x_j^\delta\}$ components to be zero. The dimensions of the three spaces R_N*, R_N, $R_{\bar{N}}$ as well as the value of some of the n_δ are reduced by this fixing. We call this most powerful technique, which we shall use repeatedly, "subspacing." The procedure of this algorithm will be the following. We begin by searching for the minimum point in the whole R_N. A necessary condition for having a minimum of V_3 inside a volume is that its second derivative matrix $\bar{V}$ should be positive definite[17] (notice that V_3 has a constant second derivative in the entire R_N) so, according to the positivity of $\bar{V}$, one finds out if there is a minimum point in R_N or not. In the case that such a minimum does not exist, we conclude that there is no minimum point inside the bounded pyramid $R_{\bar{N}}$ either. Hence we "subspace" as described above: fix $x_{j_0}^{\delta_0}$ to be zero and then look for a minimum in the $R_{(N-1)}$ space. (Simultaneously, n_{δ_0} is fixed to be the actual number of thenonvanishing component in $\mathbf{x}^{\delta_0}$.) We call this the NM or no-minimum case. In the case that a minimum exists in R_N, we shall examine if it satisfies conditions (3.7). If these conditions are satisfied then we have the desired minimum point in $R_{\bar{N}}$. But if (3.7) are violated, then the minimum point of R_N should be on its boundary and one should subspace as in the NM situation. We call that situation PM (pseudominimum). It should be clear that by transforming from R_N to $R_{(N-1)}$ we have the same problem as before and should search again for the position of the minimum in the same way, but with one variable less. This process is finite and in the extreme case the solution would be

$$x_j^\delta=\delta_{j,1}\rho^\delta \tag{3.9}$$

and none of the $\mathbf{x}^\delta$ components is a free variable any longer.

Let us have a break in our race to the minimum, and glance at what kinds of matrix $\langle A_\delta\rangle$ should be found there. There are two cases: for $\mathbf{x}^{+\delta}$ we have $\langle A^{+\delta}\rangle$ and for $\mathbf{x}^{-\delta}$ we have $\langle A^{-\delta}\rangle$, which predict the proper D^δ and G^δ

$$\langle A^{+\delta}\rangle=\left(\begin{array}{ccc|c} \sqrt{x^{+\delta}} & & 0 & \\ & \ddots & & 0 \\ 0 & & \sqrt{x^{+\delta}_{n\delta}} & \end{array}\right),\langle A^{-\delta}\rangle=\left(\begin{array}{c|ccc} & 0 & & \sqrt{x_1^{-\delta}} \\ 0 & & \cdot^{\cdot^{\cdot}} & \\ & \sqrt{x^{-\delta}_{n_\delta}} & & 0 \end{array}\right). \tag{3.10}$$

As we mentioned above for the $\langle\mathbf{x}^\delta\rangle$, the space of the $\langle A_\delta\rangle$ might be very large, but we are presently searching for only one point in it. One should notice that the ordering along the rows of $\langle A^{\pm\delta}\rangle$ is achieved by a gauge transformation but that the columns result from the structure of V_3.

IV. THE DOUBLET AND SINGLET CONDITIONS

The main technical problem in the above program is to find out the positivity of $\bar{V}$ which is the matrix of second derivatives of V_3 in R_N. The constant matrix $\bar{V}$ is not simple one, because of the "plane conditions" $[\sum_j x_j^\delta=\rho^\delta)$ and therefore we prefer to work in the R_N* space, where all the $\mathbf{x}^\delta$ components are free variables. A general vector in R_N, denoted by $\mathbf{x}$, is the direct sum of the vectors $\mathbf{x}^\delta$. The potential V_3 takes the form

$$V_3=\mathbf{x}V^*\mathbf{x}+\mathbf{ax}+c, \tag{4.1}$$

V^* plays the role of a constant "metric" in R_N* and is equal to the second derivative of V_3 in that space, c contains the self-interactions of the vectors $\mathbf{x}^\delta$ which are constant and hence already satisfy (3.9), and the $\mathbf{ax}$ terms are their interactions with all the other $\mathbf{x}^\delta$— the independent variables.

A vector $\mathbf{x}$ which satisfies the condition $\sum_{j=1}^{n\delta} x_j^\delta=0$ for every index δ will be indicated by $\mathbf{y}$. It is obvious that to have a minimum in R_N we need the following condition for any vector $\mathbf{y}$ in R_N*:

$$\mathbf{y}V^*\mathbf{y}\geqslant 0. \tag{4.2}$$

This means that in R_N, V_3 is positive definite and moving to any direction in the subspace R_N from the extremum point in it; the potential V_3 does not decrease. The condition (4.2) for the $\mathbf{x}$ vector is equivalent to the requirement that any submatrix of V^* which lies on the central diagonal should have positive determinants. We are not looking for a minimum in R_N* but in R_N, and hence have instead the condition (4.2) only for the $\mathbf{y}$ vectors. A necessary but not sufficient condition for (4.2) to be satisfied is the *"doublet condition"*:

"There is no minimum point in R_N, if V^* contains two identical central submatrices with negative determinants, for an equivalent set of variables."

Equivalent variables of the δ type are the components x_j^δ (with different values of j) of the same vector $\mathbf{x}^\delta$. The proof is straightforward. Assume that such a ma matrix v^l of dimension l and with negative determinants appears twice in V^* as v_1^l and v_2^l, and condition (4.2) is satisfied for all the vectors $\mathbf{y}$. The matrices v_1^l and v_2^l act on two equivalent spaces s_1^l and s_2^l, which are constructed from two equivalent sets $s_1=(x_{j_1}^{\delta_1}, x_{j_2}^{\delta_2}, \ldots, x_{j_l}^{\delta_l})$, $s_2=(x_{j_1'}^{\delta_1}, x_{j_2'}^{\delta_2}, \ldots, x_{j_l'}^{\delta_l})$. As a consequence of their nonpositivity there are two identical vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ in s_1 and in s_2, respectively, which satisfy $\mathbf{d}_1v_1\mathbf{d}_1=d_2v_2^ld_2<0$. Hence the vector $\mathbf{y}^d$ which is constructed from $\mathbf{d}_1$ and $-\mathbf{d}_2$: $\mathbf{y}^d=\mathrm{d}_1-\mathbf{d}_2$ (note that $\mathbf{y}^d$ is a $\mathbf{y}$ vector as a consequence of the equivalence between the sets of variables in s_1 and s_2, to which $\mathbf{d}_1$ and $\mathbf{d}_2$ belong) satisfies $\mathbf{y}^dV^*\mathbf{y}^d=d_1v_1^ld_1+d_2v_2^ld_2<0$, which contradicts our assumption that (4.2) is satisfied. Hence the "doublet condition" is necessary to satisfy the condition (4.2).

In the case that the "doublet condition" is not satisfied, one has the NM situation and should subspace. This fixing of some variables to be constants changes the structure of V^* (which is the second derivative matrix of V_3 in R_N) and hence it might satisfy (4.2). As a simple example, we look at the case where $\alpha_l<0$, then for

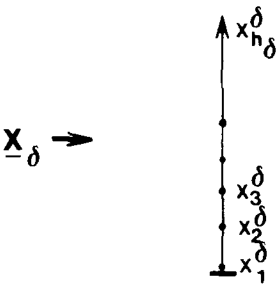


FIG. G5-1. Arrow of points represents a vector.

any n_l greater than one the doublet condition is violated, and V^* includes n_l central submatrices with negative determinants. Subspacing in this case means fixing n_{l-1} component of $\mathbf{x}^l$ to be zero and the last one to be ρ^l,

$$x_j^l = \delta_{1,j}\rho^l. \tag{4.3}$$

None of the $\mathbf{x}^l$ components is a variable any longer, and hence its contribution to V_3 is only through the terms $\mathbf{a}\cdot\mathbf{x}, c$. This is a general rule; any "subspacing" should be followed by throwing out the corresponding α_l, γ_l^{l+1}, γ_{l-1}^l terms from V^*, getting a new matrix V^*, and fixing n_6, N, N^* to their new actual values.

One should notice that the doublet condition is not a sufficient one. We know that identical central submatrices with negative determinants in V^* are forbidden, while an absence of any such matrix pairs means that a minimum exists in R_N* space and hence in R_N. Now one should contemplate the implications of a single negative determinant. In order to answer this question, we go to a subspace of the $\mathbf{y}$ vector, which is constructed from only two kinds of equivalent variables, the components of $\mathbf{x}^l$ and of $\mathbf{x}^{l+1}$. Suppose that $\gamma_l^{l+1} > 0$, the determinant of the matrix D is negative and appears in V^* only once,

$$D = \begin{pmatrix} \alpha_l & \gamma_l^{l+1} \\ \gamma_l^{l+1} & \alpha_{l+1} \end{pmatrix}.$$

This means that the interaction in R_N* space takes the form

$$V_3(\mathbf{x}_l, \mathbf{x}_{l+1}) = \sum_{j=1}^{h_l} \alpha_l x_j^{l^2} + \sum_{j=1}^{h(l+1)} \alpha_{l+1} x_j^{(l+1)^2} + 2\gamma_l^{l+1}\mathbf{x}^l\mathbf{x}'^{(l+1)}. \tag{4.4}$$

In the case that one of the α_l, α_{l+1} is negative, we should "subspace," and the corresponding variables become constants. In such a case the matrix D does not appear in V^* at all. In the case that α_l, α_{l+1} are positive, one of the eigenvalues of D, d_1, is positive, and the second, d_2, is negative. This is because $\mathrm{Tr}(D)$ is positive and $\det(D)$ is negative. The eigenvector corresponding to d_2 is (a^l, a^{l+1}). After subspacing according to the doublet condition, we are left with n_l^* matrices D_l, and n_{l+1}^* matrices D_{l+1} and a single D matrix,

$$D_l = \begin{pmatrix} \alpha_l & 0 \\ 0 & 0 \end{pmatrix}, \quad D_{l+1} = \begin{pmatrix} 0 & 0 \\ 0 & \alpha_{l+1} \end{pmatrix},$$

where, in order to eliminate the D matrix, one can either go to the border $x_j^l = 0$, or to the border $x_j^{l+1} = 0$. Now it is obvious that the minimum value of $V_3(\mathbf{x}_l, \mathbf{x}_{l+1})$ is in the direction

$$y_j^l = \begin{cases} -a_l, & j = j_0, \\ a_{l/h_l^x}, & j \neq j_0, \end{cases} \qquad y_j^{l+1} = \begin{cases} -a_{l+1}, & j = j_0', \\ a_{l+1/h_{l+1}^x}, & j \neq j_0', \end{cases}$$

where the interaction in the D matrix happened in j_0, j_0' indexes, respectively. The value of the second derivative in this direction is

$$V^*_{\min} = \alpha_l \left(1 + \frac{1}{h_l^x}\right)a_l^2 + \alpha_{l+1}\left(1 + \frac{1}{h_{l+1}^*}\right)a_{l+1}^2 + 2\gamma_l^{l+1}a_l a_{l+1} \tag{4.5}$$

and to have a minimum in R_N it should be positive. This is equivalent to the requirement that the determinant of the matrix

$$\begin{pmatrix} \alpha_l(1 + 1/h_l^x) & \gamma_l^{l+1} \\ \gamma_l^{l+1} & \alpha_{l+1}(1 + 1/h_{l+1}^x) \end{pmatrix} \tag{4.6}$$

should be positive. It is clear that $\det D^*$ can be positive, while $\det D$ is negative; so the answer to our question is dependent on the value of the parameters, and thus a single determinant might be negative.

$\underline{X}_\delta$ →

FIG. G5-2. Schematic representation of an arrow.

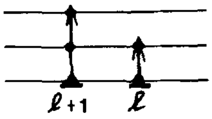


FIG. G5-3. Interaction of $\mathbf{x}^l$, with $\mathbf{x}^{l+1}$ for $\gamma_l^{l+1} < 0$.

Similar considerations are valid in spaces which contain more than two types of variables. The answer to the question as to whether a single negative determinant is permissible is that this always depends on the specific parameters under consideration.

V. GRAPHS

There are many ways to write the matrix V_3 and we find it constructive to use a graphical approach. In this way one obtains a deeper insight into V_3 and can treat it with greater facility. In the following we shall describe this graphical method. The objects of the graph are arrows which are constructed from the components of the $\mathbf{x}^6$ vectors by putting them in columns. Each column describes one vector $\mathbf{x}^6$ where its components, x_j^6, are spread in order as points along the arrow, x_1 at the tail and x_{n-6}^6 at the head of n_6 points arrow. For example, see Fig. G5-1.

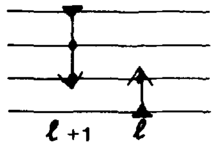


FIG. G5-4. Interaction of $\mathbf{x}^l$ with $\mathbf{x}^{l+1}$ for γ_l^{l+1} for $\gamma_l^{l+1} > 0$.

In the following we shall sketch such an arrow as in Fig. G5-2. The $\mathbf{x}^l$, $\mathbf{x}^{l+1}$ interaction is in N_l dimensional space but, as we observed, n_l, n_{l+1}, which are the actual number of non-vanishing components in $\mathbf{x}^{l+1}$, respectively, are not necessarily equal to N_l or to each other. In any case each of them is equal to, or less than, N_l. To illustrate their interaction, we take for example $N_l = 4$, $N_{(l-1)} = 2$, $N_{(l+1)} = 3$ and $n_{(l+1)} = 3$, $n_l = 2$. We shall sketch them in $N_l = 4$ unit interval in two different ways (Fig. G5-3) for $\gamma_l^{l+1} < 0$, and (Fig. G5-4) for $\gamma_l^{l+1} > 0$. One should read from these graphs the potential by reading row after row and summing them. A point in the i column and j step from the tail should be read: $\alpha_i (x_j^i)^2$ (self-interaction). Two points next to each other in the same row, the first j unit from its base and the second m unit from its base, should be read as:

$$\alpha_l (x_j^l)^2 + \alpha_{l+1} (x_m^{l+1})^2 + 2\gamma_l^{l+1} x_j^l x_m^{l+1}.$$

Therefore, Fig. G5-3 is equivalent to the potential

$$V_3(\mathbf{x}^l, \mathbf{x}^{l+1}) = [\alpha_{l+1} (x_3^{l+1})^2]$$
$$+ [\alpha_l (x_l^l)^2 + \alpha_{l+1} (x_2^{l+1})^2 + 2\gamma_l^{l+1} x_2^l x_2^{l+1}]$$
$$+ [\alpha_l (x_1^l)^2 + \alpha_{l+1} (x_1^{l+1})^2 + 2\gamma_{l+1}^l x_1^l x_1^{l+1}].$$

It is clear that such a situation occurs when $\gamma_l^{l+1} < 0$, since at the minimum the two vectors are ordered in the same sense. In Fig. G5-4 by summing all terms one obtains

$$V(\mathbf{x}^l, \mathbf{x}^{l+1}) = \sum_{j=1}^{2} \alpha_l (x_j^l)^2 + \sum_{j=1}^{3} \alpha_{l+1}^l (x_j^{l+1})^2 + 2\gamma_l^{l+1} x_2^l x_3^{l+1}.$$

The graph (Fig. G5-4) is relevant when $\gamma_l^{l+1} < 0$, and hence at the minimum the vectors are ordered in the opposite sense. We give another example for $\kappa = 5$, and $(N_5, N_4, N_3, N_2, N_1) = (3, 3, 3, 5, 5)$ and $(n_5, n_4, n_3, n_2, n_1) = (3, 3, 3, 5, 3)$, taking for all $\delta, \gamma_\delta^{\delta+1} > 0$, except $\delta = 2, \gamma_2^3 < 0$, in the graph (Fig. G5-5).

The potential in the sum of all the rows, where the first row is read as follows:

$$V_3(1) = \alpha_5 (x_1^5)^2 + \alpha_4 (x_3^4)^2 + \alpha_3 (x_1^3)^2 + \alpha_2 (x_1^2)^2$$
$$+ 2\gamma_4^5 x_1^5 x_3^4 + 2\gamma_3^4 x_3^4 x_1^3 + 2\gamma_2^3 x_1^3 x_1^2$$

and the fifth row contributes

$$V_3(5) = \alpha_2 (x_5^2)^2 + \alpha_1 (x_1^1)^2$$
$$+ 2\gamma_1^2 x_5^2 x_1^1 + 2\gamma_5^1 x_1^1 x_3^5.$$

It is important to sketch the last vector twice, the first time as the κ vector, treating it like the others, and the second time as the zero vector, reading it only for the "$\gamma_\kappa^1 \mathbf{x}^1 \cdot \mathbf{x}^\kappa$" terms, and not the "self-interaction terms". In order to remember it, we put the zero vector outside and separate it from all the other vectors by *a pair of lines*.

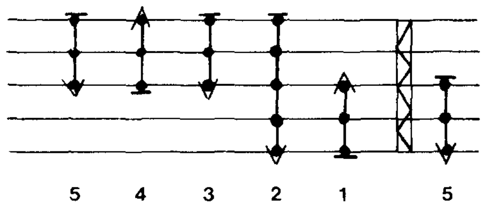


FIG. G5-5. Example to graphical representation of V_3.

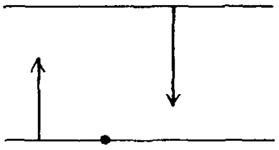


FIG. G5-6. Fixed point in a graph.

A vector $\mathbf{x}^c$ which has only one point, meaning that this point represents a constant ($x_j^c = \rho^c \delta_{j,1}$) and not a variable, will be represented by a point (and will be referred to as a "fixed point"), e.g., see Fig. G5-6. One can define such a graph by giving the length of each vector, n_δ, its direction, I_δ, (-1) for down and $(+1)$ for up, and the point where its tail lies, s_δ. These three quantities are given by

$$h_\delta^0 = \min(N_\delta, N_{\delta+1}), \tag{5.1}$$

$$I_\delta = -(\gamma_{\delta-1}^\delta / |\gamma_{\delta-1}^\delta|) I_{\delta-1}, \tag{5.2}$$

$$S_\delta = S_{\delta-1} + (I_{\delta-1} \circ N_{\delta-1}) \delta (I_\delta, -I_{\delta-1}). \tag{5.3}$$

I_1 is fixed to be (-1) and s_1 to be $(+1)$. From the above one finds that the number of rows, denoted by t, satisfies

$$\min\{N_\delta\}^\delta \leqslant t \leqslant [(\kappa/3 + 2)\max\{N_\delta\}^\delta] - \kappa/3. \tag{5.4}$$

One of the graph's features is that in the space of all the $\{x_j^\delta\}$ variables, R_{N^*}, it presents a most convenient basis by which we can construct the potential V_3 in a simple way.

In the following we construct the second derivation of V_3, V^* in this basis. We take every row of the graph as a vector, so we have t vectors of κ dimension $\mathbf{u}^l$ $(l = 1, \ldots, t)$. This means that component u_p^l will be equal to x_m^p, if in the graph the mth component of $\mathbf{x}^p (x_m^p)$ appears in the l row and in the p column. But if this is an empty place or filled by a fixed point, u_p^l is equal to zero, e.g., Fig. G5-7.

In this basis we obtain

$$\mathbf{x} V^* \mathbf{x} = \sum_l \mathbf{u}^l u^l \mathbf{u}^l + \gamma_\kappa^1 \mathbf{x}^\kappa \mathbf{x}^1, \tag{5.5}$$

u^l is a constant matrix which has nonvanishing elements in its three central diagonals in cases where the corresponding components of $\mathbf{u}^l$ are nonzero,

$$u_{\delta_1, \delta_2}^l = [\delta(\delta_1, \delta_2) \alpha_{\delta_1} + \delta(\delta_1 + 1, \delta_2) \gamma_{\delta_1}^{\delta_1 + 1} + \delta(\delta_1, \delta_2 + 1) \gamma_{\delta_1 - 1}^{\delta_1}]$$
$$\times \left(\frac{\partial \mathbf{u}_{\delta_1}^l}{\partial x^{\delta_1}} \right) \times \frac{\partial \mathbf{u}_{\delta_2}^l}{\partial x^{\delta_2}}. \tag{5.6}$$

The last two multipliers are nonzero if $\mathbf{u}^l$ depends on $\mathbf{x}^{\delta_1}$ and on $\mathbf{x}^{\delta_2}$. This is the only dependence of u_{δ_1, δ_2}^l on l. (In the above we use as an exception $\delta_{\kappa+1, 1} = 0$.) u^l

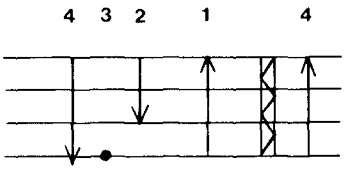


FIG. G5-7. Graphical basis for V^*.

has the form

$$u^l = \begin{pmatrix} \alpha_1 & \gamma_1^2 & & & & \\ \gamma_1^2 & \alpha_2 & \gamma_2^3 & & & \\ & \gamma_2^3 & \circ & \circ & & \\ & & \circ & \circ & \circ & \\ & & & \circ & \circ & \gamma_{\kappa-1}^{\kappa} \\ & & & & \gamma_{\kappa-1}^{\kappa} & \alpha_\kappa \end{pmatrix}.$$

One can write (5. 5) in the general form

$$\mathbf{x}V^\kappa\mathbf{x} = \mathbf{u}U\mathbf{u}. \qquad (5.7)$$

The vector $\mathbf{u}$ is a direct sum of all the $\mathbf{u}^l$ vectors, hence has $t\cdot\kappa$ components. U is a $(t\times t)\times(\kappa\times\kappa)$ dimension matrix where the U^l's appear as a $(\kappa\times\kappa)$ matrix in its diagonal. The nondiagonal $(\kappa\times\kappa)$ matrices are equal to zero, except for possible γ_κ^1 elements which might appear at most twice in each of them, and represents the terms "$2\gamma_\kappa^1\mathbf{x}^\kappa\mathbf{x}'^1$". Their sum is denoted by $U^{\kappa,1}$, e. g.,

$$U = \begin{pmatrix} (U^1)\cdots(0)\cdots(\gamma_\kappa^1 0) & & \\ (0)\vdots\quad(U^2)\quad\vdots & & 0 \\ (0^{\gamma_\kappa^1})\cdots\cdots(U^3) & & \\ & \circ & \\ 0 & & \circ \\ & & & \circ \end{pmatrix}.$$

VI. PROCEEDING

Now that we have the graphs, and V^* is represented in corresponding basis by the matrix U, we can proceed quite easily to find the structure of the minimum. One proceeds step by step, where in the jth step one looks for negative determinants in V^* constructed from j-type variables. $\{\mathbf{x}^{\delta_1},\ldots,\mathbf{x}^{\delta_j}\}$, where the δ_i type of variable means the components of the vector $\mathbf{x}^{\delta_i}$. If we find such a determinant, we should find out how many times it appears. We use for this purpose the graph which illustrates it explicitly. In the case in which this number is greater than one, NM occurs and we should go over the subspacing procedure, and adjust the graph, the potential V_3, the vectors $\mathbf{x}^\delta$ and their actual lengths n_δ, to the new situation. In case we arrive at a singlet negative determinant, we should find out if this situation is allowed, according to the description in the previous section. Nevertheless, if one wishes to avoid the tedious task of working it out he can disregard the above procedure of the negative singlet, but he should proceed from there on in the following two ways. The first assumes that this single negative determinant is allowed, and the second assumes that it is not. In the end he should obtain two extremum points and by comparing the two values of the two potentials he will know which is the minimum point and which is the right assumption. The first step was described as an example for the doublet condition, and it is shown graphically in Fig. G6-1. The second step was partially treated as an example for the "singlet negative determinant" condition, which we utilized. From the graph (Fig. G6-2) we see that there are two situations: (A) for $\gamma_l^{l+1}<0$, and (B) for $\gamma_l^{l+1}>0$. In the first case there are only two honest subspacings when the NM situation occurs ($\det D=\bar{D}<0$), irrespective of the sign of any possible $\bar{D}^*$ ($\det D^*=\bar{D}^*$)(4. 6) "and" means "consider all cases," "or" means "if single negative determinant is forbidden consider only the graph to the left and vice versa." We give another specific example of case (B) in Fig. G6-3. If one finds the sign of $\bar{D}_1^*$, $\bar{D}_2^*$, he has three honest subspaces, if he is the lazy type he should consider all five graphs as honest. One should not think that as a result only part of the $\mathbf{x}^l$ or $\mathbf{x}^{l+1}$ can be thrown out, as in that case D cannot be constructed (Fig. G6-4), because in any case we can again show that the potential, described by what is left in Fig. G6-4, is higher than the situation described by Fig. G6-5(B). Hence the previous conclusion that subspacing must take place only in the

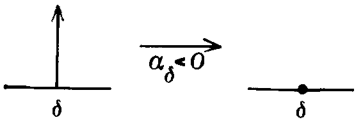


FIG. G6-1. Subspacing at the first step.

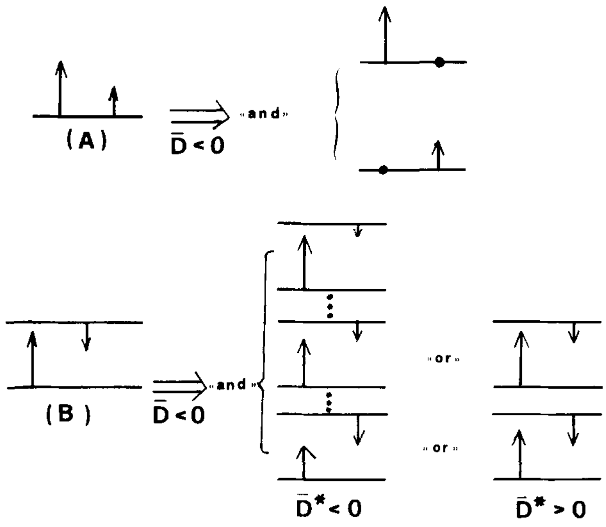


FIG. G6-2. Three kinds of subspacing in the second step.

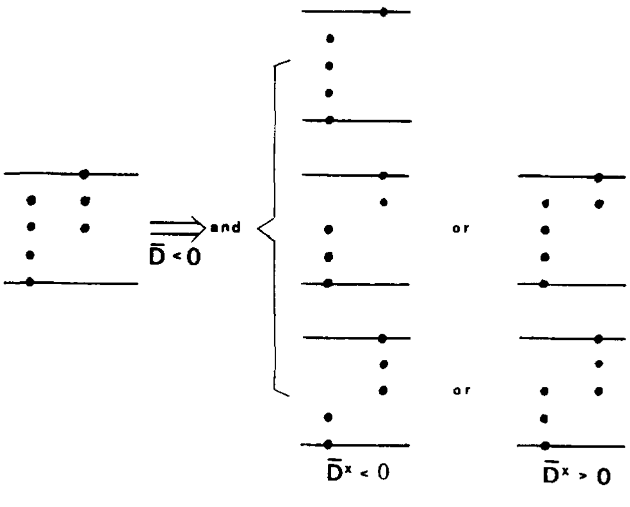


FIG. G6-3. Example of Fig. G6-2 (B).

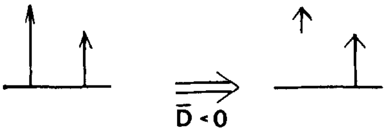


FIG. G6-4. Wrong subspacing.

upper indices of any vector $\mathbf{x}^\delta$ is unavoidable [Fig. G6-5(C)]. This conclusion arises from (3.5) where the vanishing components of $\mathbf{x}^\delta$ should always be at the head of the arrow; in any other way of making subspacing we miss the minimum point and waste our time in fruitless investigation.

The difference between subspacing at the first step ($\alpha_\delta < 0$) and that of the second ($\bar{D} < 0$) is that one must now consider a few different solutions.

In the case of Fig. G6-2(A) we have two independent solutions and in the case of Fig. G6-2(B) three. This is also the case in the following steps, where one has, in general, more than one allowed subspace with positive definite submatrices in V^*. We should list them and treat them all on the same footing.

In such a way we proceed step by step from the one dimension problem to the $(\kappa - 1)$ type of vector problem. (As we want to give a general picture, we avoid treating the κ step and leave it to the following discussions.) We consider each submatrix which arises from our graph; whenever it gives a negative determinant we do not let it appear more than once. That is done by subspacing the proper variables, which means fixing them to be zero, or, when we are left with only one variable in the $\mathbf{x}^I$ vector, it is fixed to be ρ^I. It should be clear that this procedure is not uniquely defined and is described by a tree-like diagram. This occurs because, in general, in every step we can choose several different borders in which the minimum might be found. Any such choice is followed in the next step by other choices, and hence, in general, every branch branches off into smaller twigs, etc. This game is finished when each line has a minimum at its end. The global minimum is found by comparing all these border minima and choosing the lowest one.

One should remember that even when condition (4.2) ($\mathbf{y}V^*\mathbf{y} \geqslant 0$) is fulfilled and a minimum point is found in R_N, this minimum can occur outside the domain of interest; $R_{\bar{N}}$. That means violation of (3.7) ($x_j^\delta > 0$) which is the PM situation. One should proceed from this PM as from the NM situation, but this time one does not need to check the positivity of V^*, where it remains positive definite in any subspace. Another hint one can get from the structure of V^* is that the new "subspacing" should be done only to the variables which get negative values at the minimum point of R_N. Moreover, in all the private cases which we considered it was found that if some extremum points are found to be inside the pyramid $R_{\bar{N}}$, the global minimum should be one of them; hence there is no need to treat any other branches which lead to the PM situation.

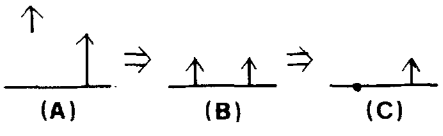


FIG. G6-5. Wrong subspacing predicts right one.

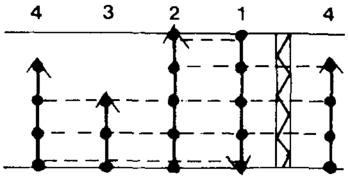


FIG. G7-1. Chains in the graph.

VII. TOPOLOGY

In this section we shall study the topological features of the graphs so we can use them to get easy answers to a few problems.

First, we connect each of the two points which are found adjacent in the same row. In this way one gets a set of "chains." For example, we look at the graph (Fig. G7-1). The chains are

$$x_5^2 - x_1^1,$$
$$x_4^2 - x_2^1 - x_4^4,$$
$$x_3^2 - x_3^1 - x_3^4 - x_3^3,$$
$$x_2^2 - x_4^1 - x_2^4 - x_2^3,$$
$$x_1^2 - x_5^1 - x_1^4 - x_1^3.$$

We should emphasize that in the topological treatment a single fixed point vector is treated equivalently to all other points, contrary to the minimum point problem. Any two points of the same vector which are found in similar chains and in the same place will be called "equivalent points." In Fig. G7-1 the four nontrivial equivalent sets are

$$(x_3^2, x_2^2, x_1^2), (x_5^1, x_4^1, x_3^1), (x_1^4, x_2^4, x_3^4), (x_1^3, x_2^3, x_3^3).$$

In this way we put a point in two different equivalent classes, "chain" and "equivalent set," and now connect both by defining as a "zone" the collection of all similar chains. Two chains are said to be similar if they have an equal number of components of each vector $\mathbf{x}^\delta$. So, the only zone in Fig. G7-1 which includes more than one chain is the collection of all four equivalent sets written above. This zone includes twelve points.

The length of a zone, which is the length of the chain in it, will be denoted by l_z, and its width, which is the number of components in any one of its equivalent sets, will be denoted by W_z. A few examples for the case $\kappa = 3$, $N_\delta = N$ (the zones are indicated) are given in Figs. G7-2–G7-4.

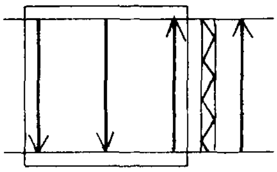


FIG. G7-2. A graph of one zone.

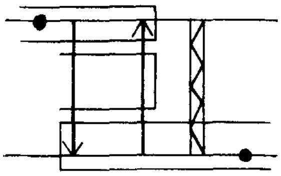


FIG. G7-3. A graph of two zones.

Part of the importance of this topological treatment can be seen by observing that all equivalent points have the same value at the minimum point. To prove it, we look at the space of all points which are in the same zone; they are interacting in V_3 between themselves only. We denote the sum of all the corresponding equivalent points in one vector $\mathbf{x}^6$ by ρ^6_z, and the potential in this zone by V^3_z. In a space where a minimum point exists, V^*_z is positive definite. Moreover, the problem is completely symmetric for the equivalent points; hence the minimum should be at the midpoint, where $x^6_{j_z}=\rho^6_z/w_z$. This rule is quite important in the PM situation, where one now knows that only subspacing of all the equivalent points together makes sense. Subspacing of only part of them will lead to another PM and hence one can treat all equivalent points as one variable only.

For example, in Fig. G7-5, where V^* is positive definite but at the minimum some of the x^6_j's are negative—only two subspacing make sense. We give another graphic illustration of equivalent chains and zones where $(N_1,N_2,N_3,N_4,N_5)=(6,10,10,6,6)$, in Fig. G7-6. The chains are

$$\begin{cases} x^3_9-x^4_2-x^5_2-x^1_2-x^2_5-x^3_5-x^4_6-x^5_6-x^1_6-x^2_1-x^3_1 \\ x^3_{10}-x^4_1-x^5_1-x^1_1-x^2_6-x^3_6-x^4_5-x^5_5-x^1_5-x^2_2-x^3_2 \end{cases}$$

$$\begin{cases} x^3_8-x^4_3-x^5_3-x^1_3-x^2_4-x^3_4 \\ x^3_7-x^4_4-x^5_4-x^1_4-x^2_3-x^3_3. \end{cases}$$

The first two chains are equivalent to each other and hence form a zone of two-point width, the last two are equivalent and form a zone of the same width as well. One should notice that two points should be in the same place in the chain in order to be equivalent; so x^3_8 is equivalent to x^3_7, but not to x^3_4, where they appear in the same chain but not in the same place.

We shall now try to learn more about the topological properties of the graphs. A chain becomes a "ring" if by picking some point in it and going from that point in one direction along the chain, we come back to the same original point. If we pass only κ points on our way back, it is called a "simple ring," if we pass other points from the same vector before closing the chain we call it a "spiral ring." The spiral ring contains $n\circ\kappa$ points, where n is greater than one. In the case where we do not come back to the same point, we call it either a "long line chain" or a "small line chain," according to whether its length is greater than κ or not. In both cases we call it an open chain. The zones are defined according to the type of chain they include. It is clear that if our original V^* is not positive definite and therefore subspacing had been done, the graph contains at most one "ring" zone, which is a simple ring of one-point width.

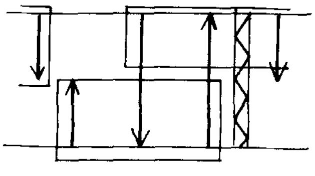


FIG. G7-4. A graph of two zones.

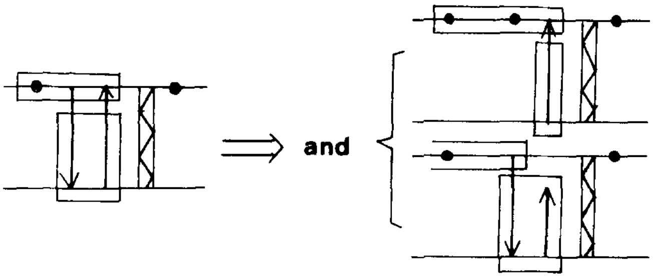


FIG. G7-5. Subspacing in PM situation.

In Fig. G7-2 there is no subspacing and the spiral ring zone is wide, while in Fig. G7-3 there is no ring. In Fig. G7-4 one makes subspacing of the second and the third vector and there is no ring, but in the case where "negative singlet" can exist, Fig. G7-4 would contain three zones, one of them a ring of one-point width. In order to observe what kind of a ring could be obtained in a specific graph we should draw only the last vector in its two places. If these are in the same direction and begin at the same level, as in Fig. G7-7, the graph can contain a single simple ring zone. In the case where their bases are in a different level, as in Fig. G7-8, one could obtain a few long line zones. In both cases we shall call the graph a "diagonal graph." In the second case the last vector is sketched in an opposite directions and we call the graph "antidiagonal." Such a graph (Fig. G7-9) can include a spiral ring zone with a length of $2\circ\kappa$ and sometimes a simple ring zone of one-point (Fig. G7-10) added to it. It is quite easy to show that in the minimum the two zones' points obtain the same value, and hence we treat it as one zone. We notice that only one ring-zone can appear in each graph.

Before concluding this discussion of the topology of the graph, we would like to emphasize again that the global minimum can never stay in a subspace which is obtained by subspacing only some of the components of an equivalent set in the PM situation. This also includes the case where such a situation arises by subspacing on another border of $R_{\bar{N}}$. After defining the topology of the

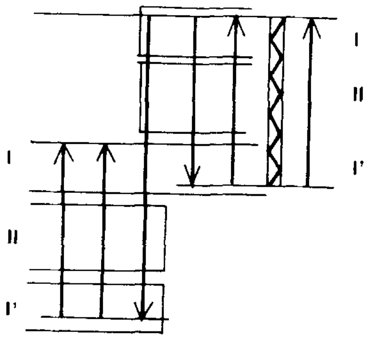


FIG. G7-6. Zones in a complicated graph.

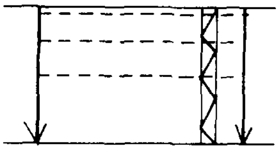


FIG. G7-7. Simple rings.

graph, the procedure becomes much easier. A chain in the graph corresponds to a chain of interactions in the great U matrix, e.g., see Fig. G7-11. A zone of width w_z in the graphs means that the chain appearing in it will perform w_z times in U. From this, one learns that the "doublet condition" should be carried out only for zones wider than one point, while the "negative singlet" procedure is relevant for a one-point width zone. Along the previous steps which treated $(\kappa - 1)$ dimensional matrices, we treated less than κ types of variables and hence worked only inside the $(\kappa \times \kappa)$ u^l matrices or their equivalent. But in the κ step one should take care of the positivity of the huge $(t\times t)\times(\kappa\times\kappa)$ matrix U. In the case where the graph includes short line zones and "simple ring" zones only, $U^{\kappa,1}$ can be made to vanish identically, and hence all the nonzero elements are concentrated in the u^l matrices and the κ step is identical to all other steps, e.g.,

$$u^l\big|_{\text{S.R.Z.}} = \begin{pmatrix} \ddots & & \gamma_\kappa^1 \\ & \ddots \; 0 & \\ 0 & \ddots & \\ \gamma_\kappa^1 & & \ddots \end{pmatrix}.$$

In the "spiral ring" zone (S.R.Z.) case, the U contains combinations of four matrices along its diagonal, e.g.,

$$u^{l_1,l_2}_{\text{S.R.Z.}} = \left(\begin{array}{c|c} \begin{matrix} \ddots & 0 \\ 0 & \ddots \end{matrix} & \begin{matrix} & \gamma_\kappa^1 \\ 0 & \\ \gamma_\kappa^1 & \end{matrix} \\ \hline \begin{matrix} & \gamma_\kappa^1 \\ 0 & \\ \gamma_\kappa^1 & \end{matrix} & \begin{matrix} \ddots & 0 \\ 0 & \ddots \end{matrix} \end{array}\right).$$

But as we have proved, at the minimum the two equivalent vectors $\langle \mathbf{u}_{l_1}\rangle$, $\langle \mathbf{u}_{l_2}\rangle$ are equal, hence one can go to the subspace described by $\mathbf{u}_{l_1} = \mathbf{u}_{l_2}$ and get the same problem as in the "simple ring" zone (SP.R.Z.) case, e.g.,

$$u^{l_1,l_2}_{\text{SP.R.Z.}} \to \left(\begin{array}{c|c} \begin{matrix} \ddots & 0 \; \gamma_\kappa^1 \\ \gamma_\kappa^1 \; 0 & \ddots \end{matrix} & 0 \\ \hline 0 & \begin{matrix} \ddots & 0 \; \gamma_\kappa^1 \\ \gamma_\kappa^1 \; 0 & \ddots \end{matrix} \end{array}\right).$$

This means that only where the graph contains long line zones the problem of the κth step spreads out from the u^l matrices and becomes complicated. We want to avoid such a situation, hence we try to bring an effective "γ_κ^1" into the u^l, and leave $U^{\kappa,1}$ equal to zero, obtaining an effective simple ring zone problem. This is done only for the purpose of checking the doublet and singlet conditions in the κ step. But before this it should be noticed that in a "diagonal graph" the width of the long line zone is just one point and therefore quite simple to avoid, in the way that the singlet condition was avoided. In the "antidiagonal" graph one should take, in the general case, three different effective "γ_κ^1", $0, -\gamma_\kappa^1, +\gamma_\kappa^1$. Only when the doublet condition is satisfied for all of them will a minimum point exist in R_N.

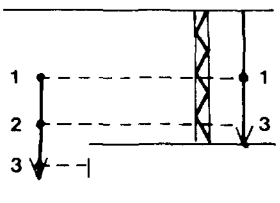


FIG. G7-8. A long line chain.

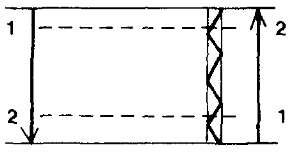


FIG. G7-9. A spiral ring.

VIII. CONCLUSION

A. Possible symmetry breaking

After going over the NM situation and finding spaces with a minimum point in R_N, going over the PM situation and finding spaces with a relevant minimum point in $R_{\overline{N}}$, comparing all the results and getting the global minimum point, one can construct the matrices $\langle A^{+\delta}\rangle$, $\langle A^{-\delta}\rangle$ which were described in Sec. III. As it is convenient to work with diagonal matrices $\langle A^\delta\rangle$, we should like to know whether we can diagonalize all these matrices simultaneously. We shall consider for a moment the case where all N_δ are equal to the same N, which means all $\langle A^\delta\rangle$ are square matrices.

Let us suppose that $\langle A^{-l}\rangle$ is antidiagonal $(\gamma_{l-1}^l > 0)$ and then make a gauge transformation which brings its first row to the last one, and the second to the last but one, etc. In such a transformation $\langle A^l\rangle$ becomes diagonal but as an additional result the columns of $\langle A^{l+1}\rangle$ are changed correspondingly. If γ_l^{l+1} is greater than zero, this change is plausible for our aim of diagonalization because $\langle A^{-l+1}\rangle$ also becomes diagonal. But if γ_l^{l+1} is negative we would have to make another gauge transformation, this time on the $\langle A^{-l+1}\rangle$ rows. In this manner we can diagonalize $(\kappa - 1)$ matrices, while the κth matrix $\langle A^{l-1}\rangle$ would possibly present a problem, where transformation on its rows will change the columns of $\langle A^l\rangle$ and the story begins again. It is quite easy to prove that such a procedure will diagonalize all the κ matrices if

$$\prod_{\delta=1}^{\kappa} (-\gamma_\delta^{\delta+1}) \geq 0, \tag{8.1}$$

which means that an even number of repulsion interactions in V_3 permits the diagonalization of all these matrices. It can be seen that for a diagonal graph the set of κ matrices is diagonalizable, and vice versa.

There is one exception to this rule and this occurs when all the nonvanishing components of each matrix

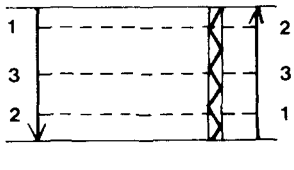
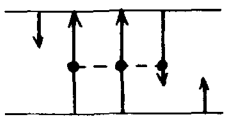


FIG. G7-10. A simple ring in a spiral zone.



$$\Longrightarrow \begin{pmatrix} 0 & 0 & & & \\ 0 & \alpha_2 & \gamma_2^3 & & 0 \\ & \gamma_2^3 & \alpha_3 & \gamma_3^4 & \\ & & \gamma_3^4 & \alpha_4 & 0 \\ 0 & & & 0 & 0 \end{pmatrix}.$$

FIG. G7-11. Chains of interactions.

$\langle A^\delta\rangle$ are equal, which means that the original V^* is positive definite, and all the $\langle \mathbf{x}^\delta\rangle$ components are in the same zone. In this specific case one can always diagonalize the whole set of matrices by a gauge transformation. In the general case, when not all the matrices are squares, the same procedure takes place where the transformed $\langle A^{-\delta}\rangle$ are diagonalized in the $(n_\delta \times n_\delta)$ square in which they lie, e.g.,

$$\langle A^{-\delta}\rangle \Rightarrow \left(\begin{array}{c|cccc} & \sqrt{x_1^\delta} & & & \\ & & \cdot & & \\ 0 & 0 & & \cdot & 0 \\ & & & \cdot & \sqrt{x^\delta_{n_\delta-1}} & \\ & & & 0 & & \sqrt{x_{n_\delta}} \end{array}\right).$$

In the undiagonalizable cases we diagonalize the first $(\kappa-1)$ matrices and leave $\langle A^\kappa\rangle$ in the $\langle A^{-\kappa}\rangle$ form. Now, when we have the diagonalized properties of the $\langle A^\delta\rangle$ matrices, it is quite easy to find the residual symmetry. We treat the case where S^{N_δ} is $U(N_\delta)$, as all other cases are quite similar.

From the above consideration we know that whenever the elements of $\sqrt{\langle \mathbf{x}^\delta\rangle}$ are equivalent they are equal, hence one can read explicitly the remaining symmetry from the graph and from the diagonal matrices.

Let us suppose that the graph contains p zones, we give each zone an index z, where $z=1,2,\ldots,p$ and denote the width of each zone by w_z. The residual symmetry of G is

$$G^R = \left[\prod_{z=1}^{p} \otimes\, U^*(W_z)\right] \otimes \left[\prod_{\delta=1}^{\kappa} U_\delta(N_\delta^*)\right]. \tag{8.2}$$

The group $U^*(W_l)$ is $U(w_l)$, except for the case in which the l zone is a spiral ring zone. In this case the $U^*(W_l)$ group is defined to be the subgroup of $U(w_l)$, where its corresponding unitary w_l dimension matrices f satisfy the equality

$$[d^{w_l}_{i,j} = \delta((i+j),(w_l+1))] \quad d^{w_l} f d^{w_l} = f. \tag{8.3}$$

The matrix d^{w_l} is the w_l dimensional antidiagonal matrix where all the elements of the antidiagonal are equal to one.

For example

$$U^*(2) = U(1)\otimes U(1), \quad U^*(4) = U(2)\otimes U(2).$$

The second bracket of (8.2) contains all the remaining parts of the $U(N_\delta)$ groups which leave all $\langle A^\delta\rangle$ invariant. Denote the actual length of the vectors by n_δ^*, the N_δ^*'s are found to be

$$N_\delta^* = \begin{cases} N_\delta - \max(n^*_{\delta+1}, n^*_\delta) & \gamma_\delta^{\delta+1} < 0 \\ N_\delta - (n^*_\delta + n^*_{\delta+1}) & \gamma_\delta^{\delta+1} > 0. \end{cases} \tag{8.4}$$

(In the case where N_δ^* is negative according to the above equation, one should take it to be zero.) The proof of (8.2) is based on the structure of the graph, or, equivalently, on that of the matrices $\langle A^\delta\rangle$. The group $U(N_l)$ acts from the left on $\langle A^l\rangle$ while $U(N_{l-1})$ acts from the right, so in order to obtain an invariant situation one should leave in G^R only those combinations under which $\langle A^l\rangle$ remains invariant. This combination is a direct product of groups of the form $U(N_j^l)\oplus U(N_j^{l-1})$ where N_j^l, N_j^{l-1} are equal to the number of the components of $\mathbf{x}^l$ which belong to the jth zone, which is w_j. The group $U(N^l)$ now acts from the right on $\langle A^{l+1}\rangle$ and is thereby connected with $U(N^{l+1})$ and along the same zones, hence goes on to bind together $U(N_j^l)\oplus U(N_j^{l+1}) \oplus U(N_j^{l+1})$, where N_j^{l+1} is again equal to N_j^{l+1}. In this way one binds together the subgroups of the $U(N_\delta)$'s in a chain similar to that of the graph, where each N_j^l is equal to the width of one zone. Hence by classifying the zones in the graph one gets the first part of G^R directly. In the above procedure there is one critical point which occurs upon returning to the first group $U(N_{l-1})$ through the $\langle A^{l-1}\rangle$ side. If there is no spiral zone in our graph, there is no change in the above program. Where there is such a zone, one is led to the conclusion that the remaining symmetry which arises from this zone is $U^*(W_z)$ instead of $U(w_z)$.

B. How to find or build a symmetry breaking

To conclude this general consideration we want to answer two questions: the first is how to find the symmetry breaking from a given potential, the second is how to find a potential which will give specific breaking.

We have difficulties answering both questions because our analysis is based on getting part of the breaking $\{\rho^\delta\}$ and part of the potential $\{\alpha_\delta, \gamma_\delta^{\delta+1}\}$ and then finding the full breaking. This is why our answers to both problems are not direct. First, given a potential V, all the $\{\alpha_\delta, \gamma_\delta^{\delta+1}\}$ are known and it is possible to construct all the allowed breakings by eliminating any NM situation. Let us denote these breakings by m $(m=1,\ldots,M)$. The global minimum is either inside these spaces or in one which is derived from them by PM situation. Then using the equivalent set method, we find the minimum of V_3 in terms of the $\{\alpha_\delta, \gamma_\delta^{\delta+1}\}$ parameters and the unknown $\{\rho^\delta\}$. This means finding the minimum of the M potentials, which are derived by imposing the plane condition

$$\chi_2^\delta = \rho^\delta - \sum_{\substack{j=1\\(j\neq 2)}}^{n_\delta} \chi_j^\delta$$

on V_3;

$$V_m^3 = \mathbf{x} v_m \mathbf{x} + \mathbf{a}_m(\boldsymbol{\rho})\mathbf{x} + c_m(\boldsymbol{\rho}). \tag{8.5}$$

[v_m is a constant matrix, $\mathbf{a}(\boldsymbol{\rho})$ is a constant vector which depends linearly on $\boldsymbol{\rho}$, and c is a constant which depends bilinearly on $\boldsymbol{\rho}$.] From this expression one finds $\langle \mathbf{x}_m^\delta\rangle$ and the minimum value V_m in terms of the $\boldsymbol{\rho}$,

$$\langle V_m\rangle = \boldsymbol{\rho} v_m^* \boldsymbol{\rho}. \tag{8.6}$$

This relation should hold in an open region surrounding $\boldsymbol{\rho}$ near the global minimum and thus gives the potential V_3 as a function of $\boldsymbol{\rho}$ only. Hence the whole original potential V can be expressed as a function of ρ only,

$$V_m = \mu\rho + \rho V_1 \rho + \rho v_m^* \rho. \tag{8.7}$$

Solving these m minimum problems, one gets $\langle \rho_m \rangle$ and the minimum values of the V_m; $\langle V_m \rangle$ as a function of the coupling constants only. Then going back to find the $\langle \mathbf{x}_m^\delta \rangle$ in terms of these parameters one selects the relevant solutions which are inside $R_{\bar{N}}$. By comparing their corresponding minima $\langle V_m \rangle$, the global minimum, which is the deeper one, can be selected.

In the case where such relevant solutions are found, the real minimum is one of them, and not any of the PM situations. The proof of this is based on the structure of the potential and by noticing that PM situations lead to a subspace of the considered spaces. But in the case where there is no legitimate solution in $R_{\bar{N}}$ one must pursue the PM procedure which means finding new matrices v_m^* for each legitimate solution and go over again with them in the same way, until an extremum is found in $R_{\bar{N}}$.

Going over to the second question of finding the coupling constants of the whole potential when the symmetry breaking is given, one must first construct the graph of the given $\langle \mathbf{x}^\delta \rangle$. In the case where it is a legitimate graph one can find from it the proper signs of the coupling constants and of the corresponding determinants in V^*. Then it is easy to find the structure of the corresponding v_m^* from (8.5) and, substituting into (8.7), one finds the following relation:

$$\rho = -\tfrac{1}{2}(V_1 + v_m^*)^{-1}\mu. \tag{8.8}$$

The proper v_m^* should also satisfy

$$\langle \mathbf{x} \rangle = -\tfrac{1}{2}(v_m^*)^{-1}\mathbf{a}(\rho). \tag{8.9}$$

It is not a hard problem to find V_1, μ, and v_m^* from these two equations; on the contrary, they leave much freedom for choosing the coupling constants. However, this is not a direct procedure because the potential so obtained may turn out to have other and deeper minima with different associated symmetries.

In order to avoid this, one must choose the v_m^* and V_1 elements carefully and consider all the legitimate graphs which might arise. The correct choice of these matrices will put the absolute minimum in the proper point.

The aim of this program is to give an algorithm and thus should be treated as a guide in handling complicated symmetry breaking. Nevertheless, many cases which were most difficult to handle properly can be treated quite easily according to the above rules. Moreover for any specific problem one can work with the help of this algorithm to get the exact solutions in a precise way.

In subsequent articles we are going to give a few examples both of the mathematical procedure and of the most attractive physical situation which might arise in such models. Anyhow, it is possible to conclude from the above discussion that the hierarchy of symmetry breaking arises from the structure of the zones and the location of the minimum point, which depends on all the parameters. Thus the strength of the breaking $\langle \rho^\delta \rangle$ is spread on the equivalent sets in different ways. The usual mechanism to get a hierarchy by broadening the space of the global minimum, which contains in the tree approximation several kinds of breaking, is present here also. One can multiply each A^δ by any unitary matrix constructed from sub-unitary matrices each one acts on an equivalent set only

$$u^\delta \cdot A^\delta = \begin{pmatrix} U_1^\delta & & & \\ & U_2^\delta & & \\ & & \cdot & \\ & & & \cdot \\ & & & & \cdot \end{pmatrix} \cdot \begin{pmatrix} a_1^\delta \cdot I & & & \\ & a_2^\delta \cdot I & & \\ & & \cdot & \\ & & & \cdot \\ & & & & \cdot \end{pmatrix}.$$

The potential depends only on $A^{\delta -}A^\delta$ and $A^{\delta +}A^\delta$, therefore such a multiplication will leave it invariant. By gauge transformation one can eliminate all these phases in any open zone, but in ring zones the phases of one of the new matrices are left. We conclude from it that the remaining symmetry G^R which is defined in (8.2) is the maximal one, and $\mathrm{U}^*(w_g)$ can reduce step by step to $\mathrm{U}(1) \otimes \mathrm{U}(1) \cdots \otimes \mathrm{U}(1)$. The minimum space of the potential V is degenerate in the following way:

$$\langle V(\mathrm{U}^*(w_g))\rangle = \langle V(\mathrm{U}^*(n_1) \otimes \mathrm{U}^*(n_2) \cdots \otimes \mathrm{U}^*(n_g))\rangle,$$
$$\sum_i n_i = w_g. \tag{8.10}$$

For example, a simple ring zone of width 4 predicts a degenerate minimum space which contains the groups:

$$\mathrm{U}(4), \mathrm{U}(3) \otimes \mathrm{U}(1), \mathrm{U}(2) \otimes \mathrm{U}(2), \mathrm{U}(1) \otimes \mathrm{U}(1) \otimes \mathrm{U}(1) \otimes \mathrm{U}(1).$$

The decision as to which one of these groups is the remaining symmetry, is left to higher order corrections. In this fashion a hierarchy in the symmetry breaking can be produced in a way that only the photon emerges without mass. Moreover, it is not hard to see that by changing the surrounding[12–14] situation, several kinds of phases can be derived as a consequence of NM and PM situations. In this way, a massless "gluon" which likes to emerge from one phase has to acquire mass in order to be free in another phase.

We hope to consider all these subjects more broadly in the following articles and to discuss the different probabilities which can arise in different domains of interactions and environments.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor Abdus Salam and Professor J. Strathdee for many useful discussions, suggestions, and encouragement. Special thanks are due to Professor J. Strathdee for a critical reading of the manuscript. The author is grateful to the International Atomic Energy Agency and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste.

[1] P. W. Higgs, Phys. Rev. **145**, 1156 (1966); T. W. Kibble, *ibid.* **155**, 1554 (1967); see also E. Abers and B. W. Lee, Phys. Rep. C 9, 1 (1973) and references therein.

[2] A. Salam, in *Elementary Particle Theory*, edited by N. Svartholm (Almquist and Wiksells, Stockholm, 1968).

[3] S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967).

[4] D. Gross and F. Wilczek, Phys. Rev. Lett. **30**, 1343 (1973); H.D. Politzer, *ibid.* 30, 1396 (1973).

[5] H. Georgi and S. Glashow, Phys. Rev. **32**, 348 (1974).

[6] J.C. Pati and A. Salam, Phys. Rev. D 10, 275 (1974).

[7] E. Ma, Phys. Rev. D **11**, 322 (1975); E. Gildener, *ibid.* **13**, 1025 (1976).

[8] J. Ellis, M.K. Gaillard, and D.V. Nanopoulos, Ref. TH. 2093 CERN preprint.

[9] I.V. Krive and A.D. Linde, Lebedev Physical Inst. Preprint No. 11 (February 1976); S. Weinberg, Phys. Rev. Lett. **36**, 294 (1976).

[10] L.-F. Li, Phys. Rev. D 9, 1723 (1974).

[11] S. Coleman and E. Weinberg, Phys. Rev. D 7, 1888 (1973).

[12] L. Dolan and R. Jackiw, Phys. Rev. D 9, 3320 (1974).

[13] T.D. Lee and M. Margolis, Phys.Rev. D **11**, 1591 (1975).

[14] A. Salam and J. Strathdee, Nucl. Phys. B 90, 203 (1975); D.A. Kirzhnits and A.D. Linde, ICTP, Trieste, preprint IC/75/28 (in preparation).

[15] Many of the results in this paper are obtained as a consequence of the positivity of the considered variables.

[16] We ignore the possibility of vanishing or equalities between some of the coupling constants which are not imposed by the underlying symmetry.

[17] Here, as throughout this article, the positivity of a matrix means nonexistence of its negative eigenvalues.

Spontaneous symmetry breaking of $U_1(M)\otimes U_2(M)\otimes U_3(M)$ with colored Higgs particles

Amir Schorr

International Centre for Theoretical Physics, Trieste, Italy

(Received 20 September 1976)

In a previous article a general method for finding the symmetry-breaking direction in a theory which is symmetric under a group $U(N_1)\otimes\cdots\otimes U(N_j)\otimes SU(M_1)\otimes\cdots\otimes\cdots\otimes SU(M_k)$ was presented. In order to explain and exhibit its utility, we derive in this paper the possible directions of spontaneous symmetry breaking in a $U_1(M)\otimes U_2(M)\otimes U_2(M)$ theory with "colored" Higgs particles. As an example we discuss the symmetry breaking in the Pati–Salam model which is found to be a legitimate one. A simple explanation of the stability of the quarks inside a "particle-phase" is exhibited.

I. INTRODUCTION

In a unified theory based on a product of simple gauge groups the Higgs particles can interact with more than one kind of vector meson.[1,2] For such a theory, we presented in a previous article[3] a technique by which one can find the spontaneous symmetry breaking by drawing a few graphs and analyzing the positivity of a simple matrix.

In order to demonstrate the method, we analyze the possible directions in which the group $U_1(M)\otimes U_2(M)\otimes U_3(M)$ can be broken. This symmetry is an important one in itself and has been utilized by Pati and Salam[1] in an $SU(4)_{\text{right}}\otimes SU(4)_{\text{left}}\otimes SU(4)_{\text{color}}$ theory of unified strong, weak, and electromagnetic interactions.

As we would like this paper to be self-contained, we shall present our technique in a simple manner, by giving in Sec. II a brief review without proofs. In Sec. III we analyze one of the four basic modes in which the symmetry $U_1(M)\otimes U_2(M)\otimes U_3(M)$ can be broken and demonstrate the graphical method.

In Sec. IV the other three basic modes are analyzed, and all the possible breakings which can arise within them are graphically drawn. Section V in dedicated to an analysis of the symmetry breaking that characterizes the Pati—Salam model.

II. RELEVANT VARIABLES FOR SEEKING A MINIMUM

The Lagrangian of the theory in which we are interested is a renormalize one and invariant under the gauge group: $G=U_1(M)\otimes U_2(M)\otimes U_3(M)$. The theory contains fermions, Higgs—Kibble particles $\{A^{\delta}_{j,k}, A^{\delta+}_{j,k}\}$ and gauge fields $\{W^{\delta}_{j,k}\}$ $(\delta=1,2,3)$ $(j,k=1,\ldots,M)$. The scalars transform under G according to the representation

$$A^1\sim(M,1,\overline{M}),\quad A^2\sim(\overline{M},M,1),\quad A^3\sim(1,\overline{M},M). \tag{2.1}$$

The Lagrangian contains all the renormalizable interactions which are invariant under G and also under three discrete symmetries which exclude cubic scalar interactions. The invariant potential V of the Higgs fields is

$$\begin{aligned}V&=\mu^{\delta}\,\mathrm{Tr}(A^{\delta}A^{\delta+})+a_{\delta,\beta}\,\mathrm{Tr}(A^{\delta}A^{\delta+})\,\mathrm{Tr}(A^{\beta}A^{\beta+})\\&\quad+\alpha_{\delta}\,\mathrm{Tr}(A^{\delta}A^{\delta+}A^{\delta}A^{\delta+})+2\gamma_{\delta}^{\delta+1}\,\mathrm{Tr}(A^{\delta}A^{\delta+}A^{(\delta+1)^+}A^{\delta+1})\\&=\mu\cdot\rho^*+\rho^*C^0\rho^*+V_3,\end{aligned} \tag{2.2}$$

where

$$\rho^{*\delta}=\mathrm{Tr}(A^{\delta}A^{\delta+}),\quad C^0_{\delta,\beta}=a_{\delta,\beta},\quad A^0\equiv A^3.$$

Our purpose is to find the minima of this potential as a function of the $6M^2$ variables $\{A^{\delta}_{j,k}, A^{\delta+}_{j,k}\}$. We restrict our discussion to potentials which are bounded from below and hence must have global minimum. The general problem is too difficult, hence we search for the minimum in subspaces of the $6M^2$ dimensional one. These subspaces are defined by fixing some of the variables, and thereby eliminating them from the main discussion. This is a basic technique in the whole procedure and will be called "subspacing."

The first subspacing is done by the "plane condition"; each $\rho^{*\delta}$ is fixed to be equal to $\mathrm{Tr}\langle A^{\delta}A^{\delta+}\rangle$ which is denoted by ρ^{δ},

$$\langle\mathrm{Tr}(A^{\delta}A^{\delta+})\rangle=\rho^{\delta}\quad\text{(plane condition)}. \tag{2.3}$$

In the subspace where all $\rho^{*\delta}$ are fixed, the first row of the potential (2.2) is a constant, hence the location of the global minimum in this subspace is determined by the second row, V_3 only. (The value of ρ^{δ} is not yet known; however, we treat it as a constant.)

Most of the subspacings will be based on the following technique: find the values of some of the variables at the global minimum; then fix them to have these values; next seek out the minimum in this subspace, which necessarily coincides with the global minimum.

Taking the $\{(A^{\delta}A^{\delta+}),\ (A^{\delta+}A^{\delta})\}$ matrices to be the independent variables, V_3 is a function of them only. It was shown that one can diagonalize simultaneously all of these variables in a subspacing procedure. Moreover, for each δ the diagonal elements of $\langle A^{\delta+}A^{\delta}\rangle$ are equal to those of $\langle A^{\delta}A^{\delta+}\rangle$, except possibly for their order which is determined by the sign of $\gamma^{\delta}_{\delta-1}$. Hence one can make a subspacing, and put all the nondiagonal elements of the matrices $\{(A^{\delta}A^{\delta+}),\ (A^{\delta+}A^{\delta})\}$ equal to zero and the diagonal elements of $(A^{\delta+}A^{\delta})$ equal to a rearrangement of those of $(A^{\delta}A^{\delta+})$. The subspace R_N which we obtain has $3(M-1)$ dimensions $\{\mathbf{x}^{\delta}\}$;

$$\mathbf{x}^{\delta}=\mathrm{diag}(A^{\delta}A^{\delta+})=(x_1^{\delta},x_2^{\delta},\ldots,x_M^{\delta}), \tag{2.4}$$

subject to the plane conditions

$$\sum_{j=1}^{M}x_j^{\delta}=\rho^{\delta}. \tag{2.5}$$

The diagonal components of $(A^{\delta+}A^{\delta})$ are expressed by the $\mathbf{x}^\delta$ variables in the following way:

$$x_j'^\delta = \mathrm{diag}\,(A^{\delta *}A^\delta)_j = \begin{cases} x_j^\delta, & \gamma_{\delta-1}^\delta < 0, \\ x_{(M+1-j)}^\delta, & \gamma_{\delta-1}^\delta > 0. \end{cases} \tag{2.6}$$

The space of $3M$ free variables $\{x_j^\delta\}$ is denoted by R_{N^*} and contains the R_N space.

The potential V_3 in R_N, as well as in R_{N^*}, can be expressed in terms of the vectors $\mathbf{x}^\delta$,

$$V_3 = \alpha_\delta \mathbf{x}^\delta \mathbf{x}^\delta + \gamma_\delta^{\delta+1}\mathbf{x}^\delta \mathbf{x}'^{\delta+1}. \tag{2.7}$$

Our interest is to find the minimum point of V_3 as a function of the $\mathbf{x}^\delta$ components. But as all the $\langle \mathbf{x}^\delta\rangle$ elements are positive, the relevant space is $R_{\bar{N}}$, a bounded pyramid in R_N, a bounded pyramid in R_N, defined by

$$\rho^\delta \geq x_j^\delta \geq 0 \quad \text{(positivity condition)}. \tag{2.8}$$

The topology is $R_{N^*} \supset R_N \supset R_{\bar{N}}$.

We know that there is a minimum point in $R_{\bar{N}}$ denoted by $\langle \mathbf{x}\rangle$, since it is a bounded pyramid. But if there is no minimum point in R_N (nonminimum situation—NM) $\langle \mathbf{x}\rangle$ should be on the borders of the pyramid—$R_{\bar{N}}$. Hence some of the $\langle \mathbf{x}\rangle$ components are zero, and another subspacing should be done. If R_N contains a minimum point but this point is outside of $R_{\bar{N}}$, then the minimum value of V_3 in $R_{\bar{N}}$ is attained on its boundary. The same conclusions as in NM obtain, and we go over the subspacing procedure as well; we call this phenomenon PM (pseudominimum). In order to obtain a minimum in R_N, the matrix of the second derivatives of V_3 denoted by $\bar{V}_3$ should be positive definite. To obtain $\bar{V}_3$, one must impose the plane conditions (2.5) and hence it is most difficult to verify the positivity of $\bar{V}$. In R_{N^*} the matrix of the second derivatives of V_3, deboted by V^*, is much simpler than $\bar{V}$, since the plan condition is not imposed here. We therefore prefer to compute the positivity of $\bar{V}$ by working through V^*. (The Lagrange multiplier method is not effective in this problem.) This means that, instead of the plane condition, we present the "doublet condition" which defines the relation between the positivity of $\bar{V}$ and that of V^* in the following way. There are three situations: (a) V^* is positive definite and therefore $\bar{V}$ is positive definite; (b) V^* is not positive definite and has at least two identical central submatrices with negative determinants, hence also $\bar{V}$ is not positive definite; (c) V^* is not positive definite and contains no more than one copy of a specific negative determinant. In this case $\bar{V}$ may or may not be positive definite, and the "singlet condition" decides this point.

The singlet condition is described in a previous paper[3] and will not be dealt with here. In any case one can do without it and treat the problem in two different ways, firstly as in case (a), and secondly as in case (b). By comparing the two values of the predicted minima one finds which is the right one.

The doublet condition should be handled step by step. In the jth step one takes into consideration the submatrix of V^* which is appropriate to the components of j vectors $\{\mathbf{x}^{\delta_1}\cdots\mathbf{x}^{\delta_j}\}$ and then goes on to $(j+1)$ vectors. Whenever situation (b) (NM) arises one should subspace variables, take their border values, (2.8), as zero or ρ^δ, and their interactions disappear from V^*, which is the second derivation according to the true variables. In this way a new positive definite $\bar{V}$ is created and one passes to case (a). Topologically it means that whenever there is no minimum in the whole R_N, the minimum in $R_{\bar{N}}$ is on its border, therefore by subspacing one transfers the problem to those borders which include a minimum point. In general there are several such legitimate borders, and one should treat them all on the same footing, finally choosing the right one as that which predicts the lowest minimum value.

On each subspace of R_N the location of the minimum point should be found. If it is inside $R_{\bar{N}}$ the procedure of subspacing terminates, but if it is outside of $R_{\bar{N}}$ (PM) one must search for the relevant minimum on $R_{\bar{N}}$'s borders, and the subspacing procedure continues. Anyhow, only a few special borders must be considered and the positivity of V^* remains unchanged. In general, one finds that a few graphs which are denoted by g^m $(m=1,\ldots,L)$ can describe the symmetry breaking and one must choose the right graph according to the value of ρ. But at this stage we can write V_3^m for each graph g^m as a function of ρ only

$$V_3^m = \rho C^m \rho. \tag{2.9}$$

Inserting it into (2.2) we find the value of the potential V as a function of ρ in L different subspaces (note that ρ is treated as a variable and not as a constant only in the context of Eq. (2.10)),

$$V^m = \mu\rho^m + \rho^m(C^0 + C^m)\rho^m. \tag{2.10}$$

The L minima values $V_{\min}^m$ of these functions are obtained at the minimum points $\rho_{\min}^m$ which correspond to the vectors $\mathbf{x}_{\min}^{\delta(m)}$. The lowest $V_{\min}^m$ for which all the $\rho_{\min}$ and $\mathbf{x}_{\min}^{\delta(m)}$ components are positive is the minimum value of the potential $\langle V\rangle$. Its associated graph describes the symmetry breaking.

III. THE (–,–,–) MODE

The main purpose of this paper is to find out the possible symmetry breakings for the group $G = \mathrm{U}_1(M) \otimes \mathrm{U}_2(M) \otimes \mathrm{U}_3(M)$ in the framework which was described above. In order to begin the investigation, the graph of the problem must be drawn. Each of the $\mathbf{x}^\delta$ vectors has M components and is described by a graphical arrow which contains M points. A graph of these three vectors can be described in an M point width domain.

We shall consider at first the $(-,-,-)$ case, where all the $\gamma_\delta^{\delta+1}$ are negative.

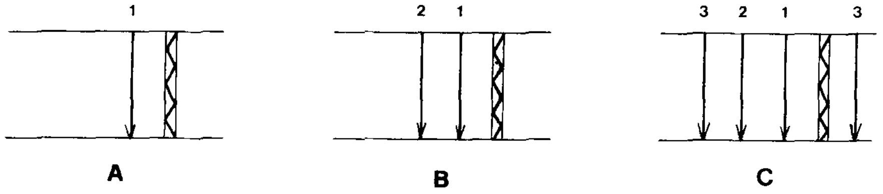


FIG. G3-1. The potential for one, two, and three vectors.

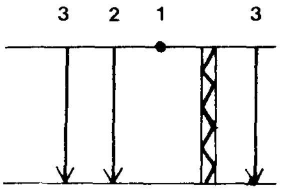


FIG. G3-2. Single negative α^6.

In Fig. G3-1(A) we draw the first vector $\mathbf{x}^1$. One reads it row by row, where, in the tail, sits x_1^1 and, in the head, sits x_M^1,

$$V_3^A=\alpha_1(x_1^1)^2+\alpha_1(x_2^1)^2+\alpha_1(x_3^1)^2+\cdots+\alpha_1(x_M^1)^2. \qquad (3.1)$$

There are only "self-interaction" terms of the single vector $\mathbf{x}^1$ present in Fig. G3-1(A). In the second step [Fig. G3-1 (B)] $\mathbf{x}^2$ is drawn. Its tail is in the same row as that of $\mathbf{x}^1$ because γ_1^2 is negative. The potential is now read again row by row, from top to bottom. It describes the self-interacting terms $(\alpha_1\mathbf{x}^1\cdot\mathbf{x}^1+\alpha_2\mathbf{x}^2\circ\mathbf{x}^2)$ and the $\mathbf{x}^1,\mathbf{x}^2$ interactions $\gamma_1^2\mathbf{x}^1\cdot\mathbf{x}'^2$,

$$\begin{aligned}V_3^B=&[\alpha_1(x_1^1)^2+\alpha_2(x_1^2)^2+2\gamma_1^2x_1^1x_1^2]+[\alpha_1(x_2^1)^2\\&+\alpha_2(x_2^2)^2+2\gamma_1^2x_2^1x_2^2]+\cdots+[\alpha_1(x_M^1)^2+\alpha_2(x_M^2)^2\\&+2\gamma_1^2x_M^1x_M^2].\end{aligned} \qquad (3.2)$$

In the third stage [Fig. G3-1(C)] $\mathbf{x}^3$ is drawn next to the vector $\mathbf{x}^2$, where their tails are in the same row, as γ_2^3 is negative. This vector $\mathbf{x}^3$ represents the self-interaction "$\alpha_3\mathbf{x}^3\mathbf{x}^3$" and the interaction with its neighbor $\mathbf{x}^2$; "$\gamma_2^3\mathbf{x}^2\mathbf{x}'^3$". The vector $\mathbf{x}^3$ is the last vector and should be drawn twice, the second time separated by two lines, and it represents the interaction with $\mathbf{x}^1$ only. Again γ_3^1 is negative and the tail of $\mathbf{x}^3$ is in the same row as that of $\mathbf{x}^1$. The full potential is now drawn and it can be read row by row,

$$\begin{aligned}V=&[\alpha_1(x_1^1)^2+\alpha_2(x_1^2)^2+\alpha_3(x_1^3)^2+2\gamma_1^2x_1^1x_1^2+2\gamma_2^3x_1^2x_1^3\\&+2\gamma_3^1x_1^3x_1^1]+\cdots+[\alpha_1(x_M^1)^2+\alpha_2(x_M^2)^2\\&+\alpha_3(x_M^3)^2+2\gamma_1^2x_M^1x_M^2+2\gamma_2^3x_M^2x_M^3+2\gamma_3^1x_M^3x_M^1].\end{aligned} \qquad (3.3)$$

After the graph is drawn, V^*, the second derivative matrix of V_3 in R_N*, should be written. As we claim, V^* in the graphical basis is quite simple:

$$\begin{aligned}V_3&=\mathbf{x}V^*\mathbf{x},\\ \mathbf{x}&=[(x_1^1,x_1^2,x_1^3),(x_2^1,x_2^2,x_2^3),\ldots,(x_M^1,x_M^2,x_M^3)]\\&=(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_M),\\ V^*&=\begin{pmatrix}U^1&&&\\&U^2&&\\&&\ddots&\\&&&U^M\end{pmatrix}.\end{aligned}$$

All the matrices U^l which incorporate the interactions represented by the lth row are equal to one matrix D,

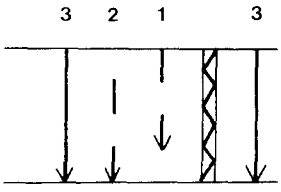


FIG. G3-3. Wrong graph for $\epsilon_1^2=(-)$.

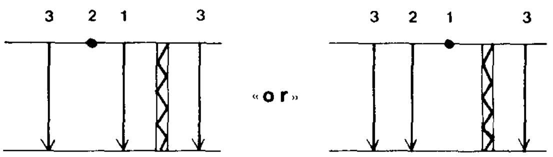


FIG. G3-4. Right solutions for negative D_2^1.

$$U^l=D=\begin{pmatrix}\alpha_1&\gamma_1^2&\gamma_3^1\\ \gamma_1^2&\alpha_2&\gamma_2^3\\ \gamma_3^1&\gamma_2^3&\alpha_3\end{pmatrix}. \qquad (3.4)$$

There are three basic kinds of submatrices in U. The first one is α_δ itself, whose determinant sign is denoted by ϵ^δ. The second kind of submatrices is

$$D_\delta^{\delta+1}=\begin{pmatrix}\alpha_\delta&\gamma_\delta^{\delta+1}\\ \gamma_\delta^{\delta+1}&\alpha_{\delta+1}\end{pmatrix},$$

the sign of whose determinants is denoted by $\epsilon_\delta^{\delta+1}$. The third one is D itself and the sign of its determinant is denoted by ϵ. Before making any investigation one should observe that in the (–,–,–) mode all the vectors $\mathbf{x}^\delta$ are topologically equivalent. Each one has two neighbors, $\mathbf{x}^{\delta+1}$ and $\mathbf{x}^{\delta-1}$, which point in its direction, so all three vectors can be treated on the same footing.

Suppose now that ϵ^1 is minus, hence there are M negative central determinants in V^*, all equal to α_1. One should subspace and go to the border $x_M^1=0$. This is the relevant border, since at the global minimum point the components of $\mathbf{x}^1$ are ordered; hence the last components should be the smallest. This is a general rule; subspacing should always be done at the head of the arrow. In this way one finds that the single border allowed by the doublet condition is

$$\begin{aligned}&x_1^1=\rho^1\\&x_j^1=0,\quad j=2,\ldots,M,\end{aligned} \qquad (3.5)$$

where at this border the x_1^1 should fulfill the plane condition alone.

All the $\mathbf{x}^1$ components are no longer variables and the graph becomes as in Fig. G3-2. The resulting V^* matrix is

$$V^*=\begin{pmatrix}L_1&&&\\&L_1&&\\&&\ddots&\\&&&L_1\end{pmatrix},\quad L_1=\begin{pmatrix}0&0&0\\0&\alpha_2&\gamma_2^3\\0&\alpha_2^3&\alpha_3\end{pmatrix}. \qquad (3.6)$$

There are no more NM situations which are caused by ϵ^1.

In another case, where ϵ_1^2 is minus, one should eliminate from V^* all the D_1^2 matrices which represent the terms $\gamma_1^2\mathbf{x}_1\mathbf{x}_2$. The only way to do this is by putting x_M^1

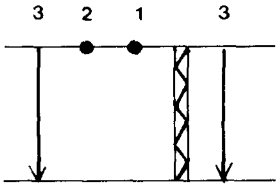


FIG. G3-5. Singlet positive α^5.

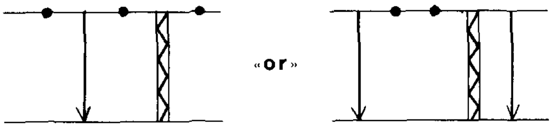


FIG. G3-6. Two nonconnected negative determinants.

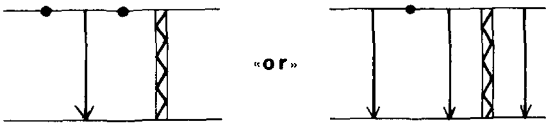


FIG. G3-7. Two connected negative determinants.

$=0$ or $x_M^2=0$ and then $x_{M-1}^1=0$ or $x_{M-1}^2=0$, etc. One can assume that in this way some of the components of $\mathbf{x}^1$ and some of the components of $\mathbf{x}^2$ are made to vanish and the graph takes the form of Fig. G3-3. However, this is not the case, and the only relevant subspacing is accomplished by fixing either all the components of $\mathbf{x}^1$, or all those of $\mathbf{x}^2$, since subspacing is done at the head of the arrows only, as in Fig. G3-4. In any specific problem one should consider both possibilities and then compare the value of V_3 at the extremum points which result. Since for our purpose these vectors are equivalent, we consider only one of these possibilities. The third situation which forces a subspacing is that in which ϵ is minus. In this case, fixing of any one of the $\mathbf{x}^\delta$ vectors will do, and one gets the same final graph as in the first two cases (Fig. G3-4).

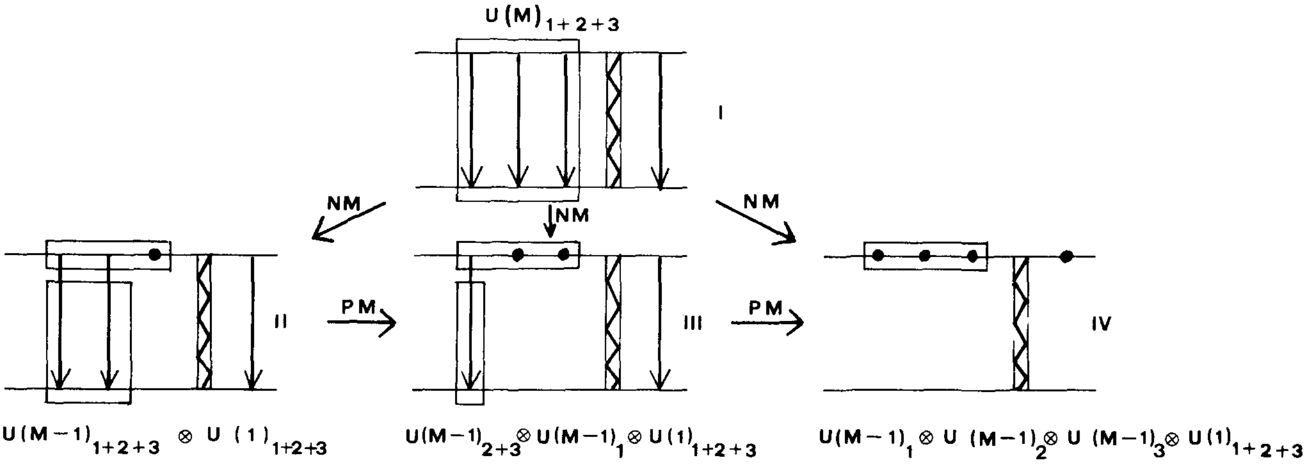
In a situation in which two different "ϵ" are negative, one might be forced to make a second subspacing; for example, in the case where $\epsilon^1=(-)$, $\epsilon^2=(-)$, the graph takes the form of Fig. G3-5. For $\epsilon^1=(-)$, $\epsilon_2^3=(-)$ the graph is represented by Fig. G3-6. For $\epsilon_1^2=(-)$, $\epsilon_2^3=(-)$ the graph is represented by Fig. G3-7. As we descend from the spaces to their subspaces by the subspacing procedure, any subspacing which is not forced by the doublet condition is not relevant. We can define this principle by two working rules. The first sates that subspacing should be done first according to ϵ^δ, then according to $\epsilon_\delta^{\delta+1}$ and finally according to ϵ. In this way one does the minimal amount of subspacing. The second "working rule" states that whenever a set of vectors point in the same direction and they have to be subspaced, only the components of one of them are fixed to be constant, e.g., $x_j^{\delta_0}=\delta_{j,1}\rho^{\delta_0}$. In this way one can conclude that the final graph which describes the symmetry breaking in the $(-,-,-)$ case must be one of the kinds shown in Fig. G3-8 (the maximal remaining symmetry is written on the left). The residual symmetry is found from the graphs in the following way. We define three equivalence relations: "chain", "equivalent set" and "zone". A "chain" is the collection of all points which are connected to one another by the "interaction link" $\gamma_\delta^{\delta+1}\mathbf{x}^\delta\mathbf{x}^{\delta+1}$, e.g., see Fig. G3-9. In graph G3-9(A) there are two different kinds of chain (open chains and simple ring), whereas graph G3-9(B) represents a third kind (spiral rings). A "zone" is the collection of all the chains of the same kind in one graph. (Two chains are from the same kind if they contain the same number of components of each vector.) In G3-9(a) there are two zones, in G3-9(B) there is one (see Fig. G3-8). An "equivalent set" is the collection of all points which belong to the same vector $\mathbf{x}^\delta$ and to the same zone. We shall give more examples in the following.

The length of a zone z is the length of its chains l_z, and the width of a zone z is the width of its equivalent sets w_z. For example in Fig. G3-8(1) there is one zone with $l_z=3$, $w_z=M$; in Fig. G3-8(2) there are two zones, $l_1=3$, $w_1=1$ and $l_2=2$, $w_2=M-1$; in Fig. G3-8(3) there are two zones: $l_1=3$, $w_1=1$ and $l_2=1$, $w_2=M-1$; in Fig. G3-8(4) there is one zone: $l=3$, $w=1$.

At the minimum point $\langle\mathbf{x}\rangle$, all the w_z components of $\langle\mathbf{x}^\delta\rangle$ which belong to the same zone z are equal. Hence in any "diagonalized mode" where the arrow, $\mathbf{x}^3$, vector points in the same direction on both sides of the two separating lines, the maximal residual symmetry G^R is

$$G^R=\left[\prod_{z=1}^{p}\times U(w_z)\right]\otimes\left[\prod_{i=1}^{3}U_i(M_i^*)\right] \tag{3.7}$$

(p is the number of zones in the graph).


U(M)$_{1+2+3}$
I
NM
NM
NM
II
PM
III
PM
IV
U(M−1)$_{1+2+3}$ ⊗ U (1)$_{1+2+3}$
U(M−1)$_{2+3}$⊗U(M−1)$_1$⊗U(1)$_{1+2+3}$
U(M−1)$_1$⊗U (M−1)$_2$⊗ U (M−1)$_3$⊗ U(1)$_{1+2+3}$

Zones are indicated by closed rectangles

FIG. G3-8. Minimal symmetry breaking in the $(-,-,-)$ case.

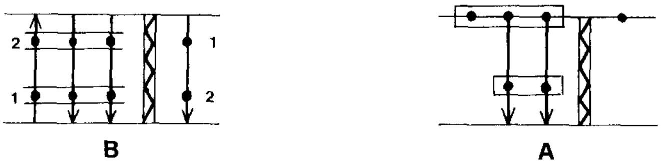


FIG. G3-9. Examples of chains.

The group $U_\delta(M_\delta^*)$ is the subgroup of $U_\delta(M)$, which leaves the graph invariant

$$M_\delta^* = \begin{cases} M - \max(m_\delta, m_{\delta+1}) & \gamma_\delta^{\delta+1} < 0 \\ M - (m_\delta + m_{\delta+1}) & \gamma_\delta^{\delta+1} > 0 \end{cases}$$

(m_δ is the actual length of $\langle \mathbf{x}^\delta \rangle$ in the graph).

Referring back to Fig. G3-8(1) we find: $p=1$, $w_1 = M$, $M_\delta^* = 0$ $(\delta = 1,2,3)$, and obtain $G_1^R = \mathrm{U}(M)_{1+2+3}$. In Fig. G3-8(2) $p=2$, $w_2 = 1$, $w_1 = M-1$, $M_\delta^* = 0$ $(\delta = 1,2,3)$ and obtain $G_2^R = \mathrm{U}(M-1)_{1+2+3} \otimes \mathrm{U}(1)_{1+2+3}$. In Fig. G3-8(3) $p = 2$, $w_2 = 1$, $w_1 = (M-1)$, $M_\delta^* = 0$ $(\delta = 2,3)$, $M_1^* = (M-1)$ and obtain $G^R = \mathrm{U}(M-1)_{2+3} \otimes \mathrm{U}(M-1)_1 \otimes \mathrm{U}(1)_{1+2+3}$. In Fig. G3-8(4) $p=1$, $w_1 = 1$, $M_\delta^* = (M-1)$ $(\delta = 1,2,3)$ and obtain $G^R = \mathrm{U}(M-1)_1 \otimes \mathrm{U}(M-1)_2 \otimes \mathrm{U}(M-1)_3 \otimes \mathrm{U}(1)_{1+2+3}$.

IV. THE (+,+,–), (+,–,–), (+,+,+) MODES

The second mode which is considered is the $(+,+,-)$ case; $\gamma_1^2 > 0$, $\gamma_2^3 > 0$, $\gamma_3^1 > 0$. The basic graph is shown in Fig. G4-1. It is a diagonal mode, as are all modes which contain an even number of positive $\gamma_\delta^{\delta+1}$ terms. The vectors $\mathbf{x}^1$, $\mathbf{x}^3$ are equivalent in the sense that both of them have one neighboring vector pointing in its direction, and one pointing in opposite direction. From that point of view there are two kinds of NM situations. The first kind is a consequence of negative ϵ^δ or ϵ_3^1, which means subspacing of parallel vectors as in the $(-,-,-)$ situation; the second kind can arise from negative ϵ_1^2 which means subspacing of opposite vectors. These subspacings caused by negative D_2^1 are of two types and are represented in Fig. G4-2.
In Fig. G4-2(A) the singlet condition does not permit a negative singlet ϵ_1^2 to be present in V^*, whereas in Fig. G4-2(B) such a singlet is allowed. The structure of V^* after subspacing in the two cases is [D is given by (3.4)]:

$$V_{A,B}^* = \begin{pmatrix} L_1 & & & & 0 \\ & \ddots & & & \\ & & L_1 & & \\ 0 & & & L_2 & \\ & & & & \ddots \\ & & & & & L_2 \end{pmatrix}, \begin{pmatrix} L_1 & & & & & 0 \\ & \ddots & & & & \\ & & L_1 & & & \\ 0 & & & D & & \\ & & & & L_2 & \\ & & & & & \ddots \\ & & & & & & L_2 \end{pmatrix},$$

$$L_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \alpha_2 & \alpha_2^3 \\ 0 & \alpha_2^3 & \alpha_3 \end{pmatrix}, \qquad L_2 = \begin{pmatrix} \alpha_1 & 0 & \gamma_3^1 \\ 0 & 0 & 0 \\ \gamma_3^1 & 0 & \alpha_3 \end{pmatrix}. \qquad (4.1)$$

In Fig. G4-1(A) there are two zones of width: $w_1 = l$, $w_2 = M - l$, and in Fig. G4-1(B) there are three zones of width: $w_1 = l$, $w_2 = 1$, $w_3 = M - l - 1$ $(1 \leqslant l \leqslant M-2)$. One must consider all the relevant borders on the same footing, which means it is necessary to consider each value of l for which the w's are nonnegative. An NM situation which arises because of negative ϵ has the subspacing characterized by Figs. G3-2 and G4-2 as its solution.

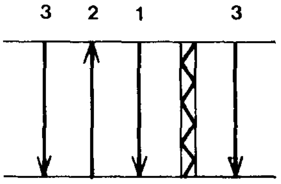


FIG. G4-1. Basic graph for the (+,+,–) mode.

In the following scheme (Fig. G4-3) we present all the possible graphs which can arise as solutions to NM and PM situations in the $(+,+,-)$ mode. The parameter l can assume any value between zero to $(M-1)$, which makes sense.

The correspondence residual symmetry breakings are

I. $\mathrm{U}(M)_{1+2+3}$ (this is the maximal one)

II. $\mathrm{U}(l)_{1+2+3} \otimes \mathrm{U}(M-l-1)_{1+2+3} \otimes \mathrm{U}(1)_{1+2+3}$

III. $\mathrm{U}(l)_{1+2+3} \otimes \mathrm{U}(M-\ l)_{1+2+3}$

IV. $\mathrm{U}(l)_{1+2} \otimes \mathrm{U}(l)_3 \otimes \mathrm{U}(M-l-1)_{1+2+3} \otimes \mathrm{U}(1)_{1+2+3}$

V. $\mathrm{U}(l)_{1+2} \otimes \mathrm{U}(l)_3 \otimes \mathrm{U}(M-l-1)_{1+2+3} \otimes \mathrm{U}(1)_{1+2+3}$

VI. $\mathrm{U}(l)_{1+2} \otimes \mathrm{U}(l)_3 \otimes \mathrm{U}(M-l)_{1+2+3}$

VII. $\mathrm{U}(l)_{1+2} \otimes \mathrm{U}(l)_3 \otimes \mathrm{U}(M-l-2)_{1+3} \otimes \mathrm{U}(M-l-2)_3$
$\otimes \mathrm{U}(1)_{1+2+3} \otimes \mathrm{U}(1)_{1+2+3}$

VIII. $\mathrm{U}(l)_{1+2} \otimes \mathrm{U}(l)_3 \otimes \mathrm{U}(M-l-1)_{1+3} \otimes \mathrm{U}(M-l-1)_2$
$\otimes \mathrm{U}(1)_{1+2+3}$

IX. $\mathrm{U}(M-2)_1 \otimes \mathrm{U}(M-2)_2 \otimes \mathrm{U}(M-2)_3 \otimes \mathrm{U}(1)_{1+2+3} \otimes \mathrm{U}(1)_{1+2}$

Now we shall consider the $(-,-,+)$ mode, taking γ_1^2, γ_2^3 negative and γ_3^1 positive. The basic graph is Fig. G4-4. It shows that this is an antidiagonal case, since the two vectors representing $\mathbf{x}^3$ are in opposite directions. This is due to the single positive γ_3^1, where in general, for an odd number of positive γ parameters the graph is nondiagonal. As a consequence, the maximal remaining symmetry does not satisfy Eq. (3.7) as it stands, and we have instead:

$$G^R = \left[\prod_{z=1}^{p} \otimes \mathrm{U}^*(w_z)\right] \otimes \left[\prod_{\delta=1}^{3} \otimes \mathrm{U}_\delta(M_\delta^*)\right]. \qquad (4.2)$$

The group $\mathrm{U}^*(w_z)$ is equal to $\mathrm{U}(w_z)$ if the chains which are included in the z zone are open chains. But if they are closed spiral chains, $\mathrm{U}^*(w_z)$ is $\mathrm{U}(w_z/2) \otimes \mathrm{U}(w_z/2)$

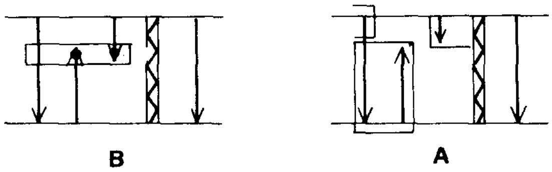


FIG. G4-2. Subspacing of negative D_2^1.

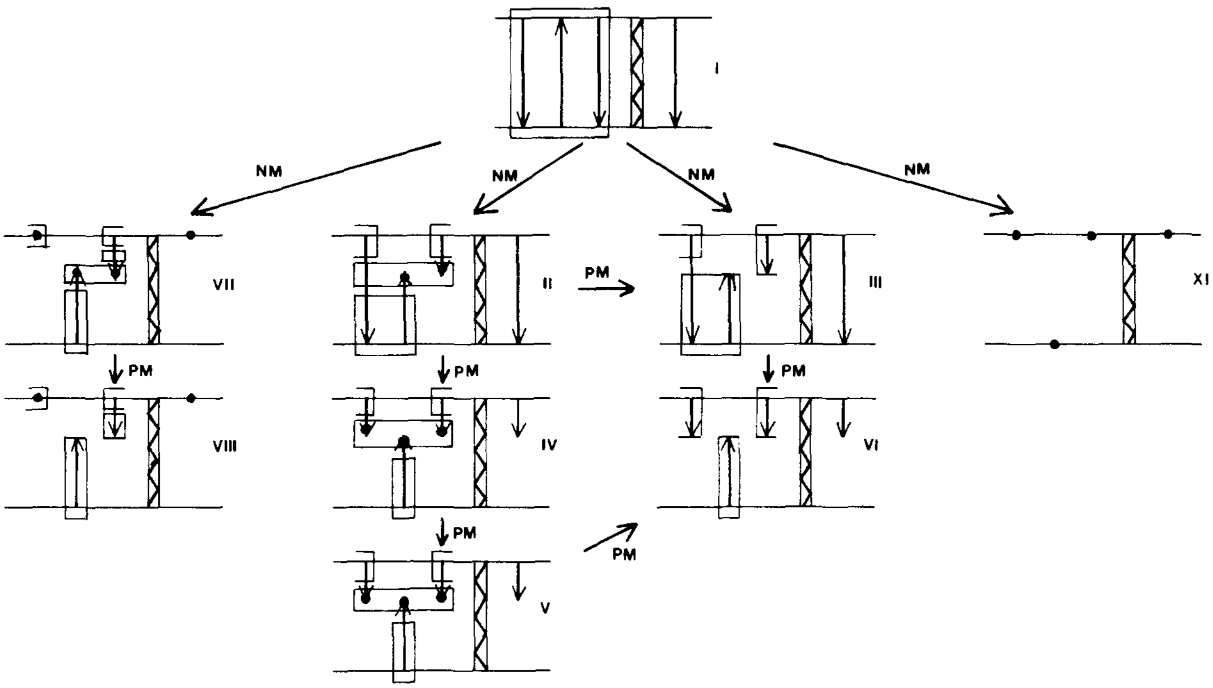


FIG. G4-3. Symmetry breaking in the (+,+,−) case.

when w_g is even, and $\mathrm{U}[(w_g-1)/2]\otimes\mathrm{U}[(w_g-1)/2]\otimes\mathrm{U}(1)$ for odd w_g. The only two cases in the $\mathrm{U}_1(M)\otimes\mathrm{U}_2(M)\otimes\mathrm{U}_3(M)$ problem where such a situation arises in a nontrivial way ($w_g\neq 1$), are those in which all the signs of $\{D_1^\delta, D_2^\delta, D_3\}$ ($\delta=1,2,3$) are positive and the signs of the $\gamma_\delta^{\delta+1}$ are either $(-,-,+)$ or $(+,+,+)$. Hence a minimum point is found to be in R_N and the entire graph is a single zone. Therefore all the points of each vector are in one equivalent set and take the same value

$$x_j^\delta=\rho^\delta/M. \qquad (4.3)$$

This point is in $R_{\bar N}$, and therefore it is the global minimum. A point in the A^δ space which realizes that minimum is presented in (4.4) ($k=1,2$) $a^\delta=\sqrt{\rho^\delta/M}$,

$$\langle A_\kappa\rangle=\begin{pmatrix} a_\kappa & & & 0\\ & a_\kappa & & \\ & & \ddots & \\ 0 & & & a_\kappa\end{pmatrix},\quad \langle A_3\rangle=\begin{pmatrix} 0 & & & a_3\\ & & a_3 & \\ & \cdot^{\cdot^{\cdot}} & & \\ a_3 & & & 0\end{pmatrix}. \qquad (4.4)$$

In this degenerate case we can make another gauge transformation and bring $\langle A_3\rangle$ to $\langle A_3^E\rangle$, if M is even, and to $\langle A_3^0\rangle$, if M is odd,

$$\langle A_3^E\rangle=\begin{pmatrix} a_3 & & & & 0\\ & \ddots & & & \\ & & a_3 & & \\ & & & -a_3 & \\ & & & & \ddots \\ 0 & & & & & -a_3\end{pmatrix},\ \langle A_3^0\rangle=\begin{pmatrix} a_3 & & & & & 0\\ & \ddots & & & & \\ & & a_3 & & & \\ & & & 0 & & \\ & & & & -a_3 & \\ & & & & & \ddots \\ 0 & & & & & & -a_3\end{pmatrix}. \qquad (4.5)$$

The residual symmetry in the even case is $G^R=\mathrm{U}(M/2)\otimes\mathrm{U}(M/2)$ and in the odd case it is $G^R=\mathrm{U}[(M-1)/2]\otimes\mathrm{U}[(M-1)/2]\otimes\mathrm{U}(1)$. (In the first two modes which are diagonalized, all three matrices $\langle A^\delta\rangle$ result from multiplication of the identity matrix by a scalar $\sqrt{\rho^\delta/M}$, if V^* is originally positive definite.)

The possible directions of symmetry breaking in the $(-,-,+)$ mode are shown in Fig. G4-5.

The corresponding residual symmetries are: (M^* equal to $M/2$ for even M, and to $[(M-1)/2]$ for odd M)

I. $\mathrm{U}(M^*)\otimes\mathrm{U}(M^*)\otimes\mathrm{U}(M-2M^*)$,

II. $\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(M-2l-2)\otimes\mathrm{U}(1)$,

II*. (for odd M) $\mathrm{U}(M^*)\otimes\mathrm{U}(M^*)\otimes\mathrm{U}(M^*)$,

III. $\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(M-2l-2)\otimes\mathrm{U}(1)$,

IV. $\mathrm{U}(l+1)\otimes\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(M-2l-2)\otimes\mathrm{U}(1)$,

V. $\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(M-l-1)\otimes\mathrm{U}(M-2l-2)\otimes\mathrm{U}(1)$,

V*. (for odd M) $\mathrm{U}(M^*)\otimes\mathrm{U}(M^*)\otimes\mathrm{U}(M^*)\otimes\mathrm{U}(1)$,

VI. $\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(M-l-1)\otimes\mathrm{U}(M-2l-1)\otimes\mathrm{U}(1)$,

VI*. (for odd M) $\mathrm{U}(M^*)\otimes\mathrm{U}(M^*)\otimes\mathrm{U}(M^*)\otimes\mathrm{U}(1)$,

VII. $\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(M-2l)$,

VIII. $\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(M-2l)$,

IX. $\mathrm{U}(M-l)\otimes\mathrm{U}(M-l)\otimes\mathrm{U}(M-2l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(l)$,

X. $\mathrm{U}(M-2)\otimes\mathrm{U}(1)$,

XI. $\mathrm{U}(M-2)\otimes\mathrm{U}(1)\otimes\mathrm{U}(1)$,

XII. $\mathrm{U}(M-2)\otimes\mathrm{U}(1)\otimes\mathrm{U}(1)\otimes\mathrm{U}(1)$,

XIII. $\mathrm{U}(M-l-2)\otimes\mathrm{U}(M-l-2)\otimes\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(1)\otimes\mathrm{U}(1)$,

XIV. $\mathrm{U}(M-l-1)\otimes\mathrm{U}(M-l-1)\otimes\mathrm{U}(l)\otimes\mathrm{U}(l)\otimes\mathrm{U}(1)$,

XV. $\mathrm{U}(M-1)\otimes\mathrm{U}(M-2)\otimes\mathrm{U}(1)$,

XVI. $\mathrm{U}(M-1)\otimes\mathrm{U}(M-1)\otimes\mathrm{U}(M-2)\otimes\mathrm{U}(1)$.

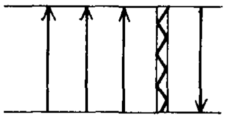


FIG. G4-4. Basic graph of the (−,−,+) mode.

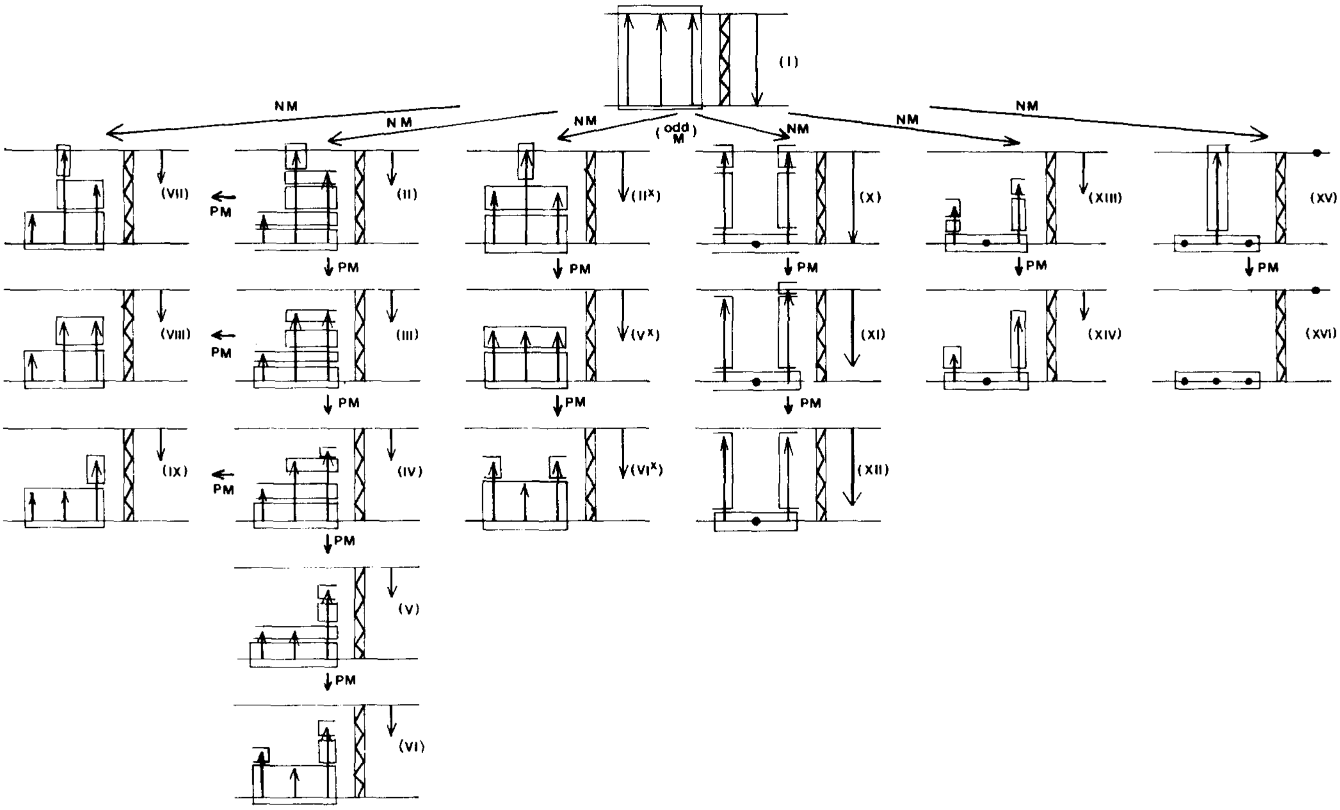


FIG. G4-5. Minimal symmetry breaking in the (−, −, +) mode.

The fourth mode is the (+, +, +) case in which all the $\gamma_\delta^{\delta+1}$ coupling constants are positive, the basic graph is described by Fig. G4-6. All three vectors are topologically equivalent and the graph is antidiagonal. When all the "ϵ" are positive and no subspacing is needed, the maximal remaining symmetry is the same as in the (−, −, +) mode.

We give below the graphs of the breakings which arise in this mode, and the corresponding residual symmetries (Fig. G4-7):

I. $U(M^*)\otimes U(M^*)\otimes U(M-2M^*)$,

II. $U(M-2l-2)\otimes U(l)\otimes U(l)\otimes U(1)$,

II*. (for odd M) $U(M^*)\otimes U(M^*)\otimes U(1)$,

III. $U(M-2l-2)\otimes U(l)\otimes U(l)\otimes U(l)\otimes U(1)$,

IV. $U(M-2l-2)\otimes U(l)\otimes U(l)\otimes U(l+1)\otimes U(1)$,

V. $U(M-l-1)\otimes U(M-2l-2)\otimes U(l)\otimes U(l)\otimes U(1)$,

V*. (for odd M) $U(M^*)\otimes U(M^*)\otimes U(M^*)\otimes U(1)$,

VI. $U(M-l-1)\otimes U(M-2l-1)\otimes U(l)\otimes U(l)\otimes U(1)$,

VII. $U(M-l)\otimes U(M-2l)\otimes U(l)\otimes U(l)$,

VIII. $U(M-2l)\otimes U(l)\otimes U(l)$,

IX. $U(M-2l)\otimes U(l)\otimes U(l)\otimes U(l)$,

X. $U(M-2)\otimes U(M-2)\otimes U(M-2)$.

In the above investigation it was emphasized that the residual symmetry (which we write in correspondence with a specific graph) is the maximal one. This breaking is the only one which is described by the graph, except in the four situations where V^* is originally positive definite. In these four situations (Figs. G3-8-I, G4-3-I, G4-5-I, G4-7-I) any component $\sqrt{\langle x_1^3\rangle}$ ($\sqrt{\langle x_j^6\rangle}$ appears in $\langle A^6\rangle$) can acquire a different phase, hence the M point-width zone in the graph can split into any combination, without changing the value of the potential at the minimum. The minimum is degenerate,

$$\langle V(U^*(M))\rangle = \langle V(U^*(M_1)\otimes U^*(M_2)\otimes\cdots U^*(M_n))\rangle \qquad \left(\sum_i M_i = M\right).$$

The decision as to which of these groups the symmetry G will be reduced is left to higher order (quantum) contributions.

V. THE PATI-SALAM MODEL

As an example of this procedure of analyzing a symmetry breakdown we treat the Pati–Salam model.[1] From our point of view the problem is the following: in a gauge theory with $U_1(4)\otimes U_2(4)\otimes U_3(4)$ $[U(4)_{\mathrm{right}}\otimes$

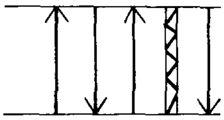


FIG. G4-6. Basic graph of the (+, +, +) mode.

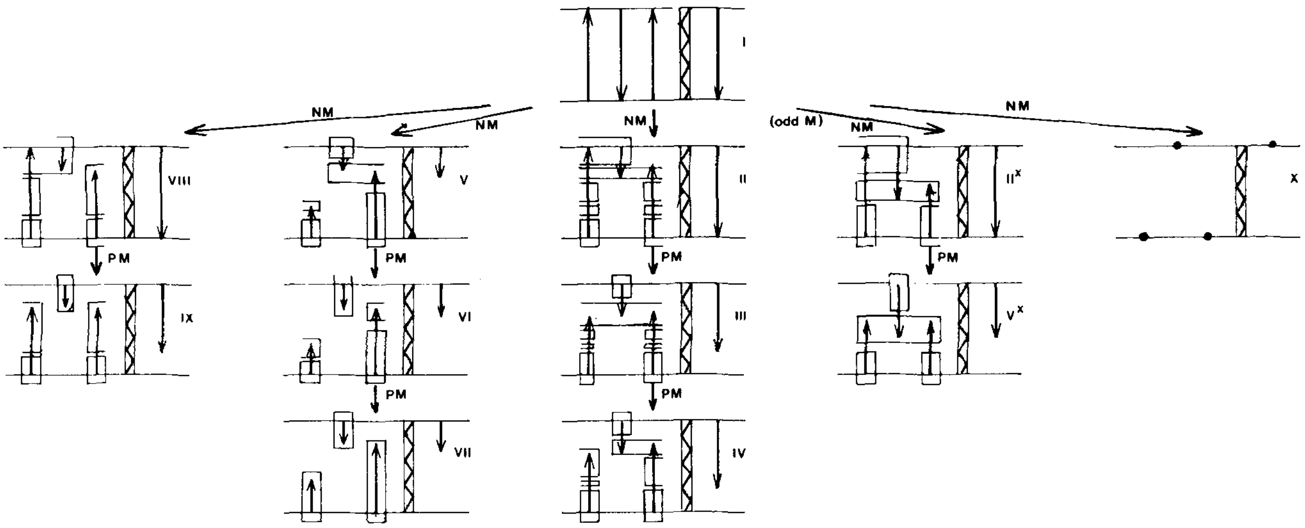


FIG. G4-7. Minimal symmetry breaking in the (+, +, +) mode.

$U(4)_{\text{color}} \otimes U(4)_{\text{left}}]$ as the underlying symmetry, is the symmetry breaking described by (5.1) allowed? If it is a legitimate solution, what should be the character of the coupling constants?

$$\langle A_1\rangle = \begin{pmatrix} a_1 & & & 0 \\ & a_1 & & \\ & & a_1 & \\ 0 & & & a_4 \end{pmatrix}, \quad \langle A_2\rangle = \begin{pmatrix} 0 & & & 0 \\ & 0 & & \\ & & 0 & \\ 0 & & & b \end{pmatrix},$$

$$\langle A_3\rangle = \begin{pmatrix} c_1 & & & 0 \\ & c_1 & & \\ & & c_1 & \\ 0 & & & c_4 \end{pmatrix}. \tag{5.1}$$

At first sight, the $\langle A^\delta\rangle$ matrices show that the breaking is of the diagonal type, hence it belongs either to the first mode $(-,-,-)$ or the second one $(-,+,+)$. We consider first the $(-,-,-)$ case in which the $\gamma_\delta^{\delta+1}$ are all negative. The basic diagram (Fig. G5-1) should transform to Fig. G5-2. A set of coupling constants which predicts such a situation is

$$\gamma_\delta^{\delta+1} < 0, \quad \alpha_1, \alpha_3 > 0, \quad \alpha_2 < 0, \tag{5.2}$$

$$\operatorname{sign}(\det D_3^1) = \epsilon_3^1 = (+), \quad D_3^1 = \begin{pmatrix} \alpha_3 & \gamma_3^1 \\ \gamma_3^1 & \alpha_1 \end{pmatrix}. \tag{5.3}$$

Instead of the $\alpha_2 < 0$ condition, a negative determinant of D_1^2 or D_2^3 or D could also predict this NM subspacing. However, we shall consider as an example only condition (5.2). The conditions (5.2) and (5.3) are sufficient to guarantee that the extremum point of R_N is a minimum. The requirements that this minimum point is inside $R_{\bar N}$, and hence that no PM situation should arise, must be found.

For this purpose we write the potential V_3 in terms of the coupling constants, the magnitudes ρ^δ and the four variables in R_{N*}, as they appear in Fig. G5-3.

$$V_3 = \alpha_1(x_1^2 + 3x_2^2) + \alpha_3(z_1^2 + 3z_2^2) + 2\gamma_3^1(x_1 z_1 + 3x_2 z_2) + 2\gamma_1^2\rho_2 x_1 + 2\gamma_2^3 z_1\rho_2 + \alpha_2\rho_2^2. \tag{5.4}$$

The potential in terms of the R_N variable should be written, which means expressing x_2, z_2 as a function of x_1, z_1. We find it profitable to express all four variables in terms of two "normalized" ones, σ, ν, and "normalized" coupling constants;

$$\begin{aligned} x_1 &= \rho_1(\tfrac14 - 3\sigma), & z_1 &= \rho_3(\tfrac14 - 3\nu), \\ x_2 &= \rho_1(\tfrac14 + \sigma), & z_2 &= \rho_3(\tfrac14 + \nu), \\ \bar\alpha_\delta &= \rho_\delta^2\alpha_\delta, & \bar\gamma_\delta^{\delta+1} &= \rho_\delta\rho_{\delta+1}\gamma_\delta^{\delta+1}. \end{aligned} \tag{5.5}$$

In terms of these variables the potential V_3 in R_N is

$$\bar V = 12\bar\alpha_1\sigma^2 + 12\bar\alpha_3\nu^2 + 24\gamma_3^1\sigma\nu - 6\bar\gamma_1^2\sigma - 6\bar\gamma_2^3\nu + \tfrac12[\bar\alpha_1/2 + \bar\alpha_3/2 + 2\bar\alpha_2 + \bar\gamma_1^2 + \bar\gamma^3{}_2 + \bar\gamma_3^1] \tag{5.6}$$

and the minimum point is found to be

$$\begin{aligned} \langle\sigma\rangle &= \tfrac12(\bar\alpha_3\bar\gamma_1^2 + |\bar\gamma_3^1|\,\bar\gamma_2^3)\ (\det\bar D_3^1)^{-1}, \\ \langle\nu\rangle &= \tfrac12(\bar\alpha_1\bar\gamma_2^3 + |\bar\gamma_3^1|\,\bar\gamma_1^2)\ (\det\bar D_3^1)^{-1}. \end{aligned} \tag{5.7}$$

The requirement that no PM situation exists and that the minimum point is *inside* the "pyramid" $R_{\bar N}$ is

$$-\tfrac14 < \langle\sigma\rangle, \langle\nu\rangle < \tfrac1{12} \tag{5.8}$$

We already know that $\langle x_1\rangle$ is greater than $\langle x_2\rangle$, so $\langle\sigma\rangle$, $\langle\nu\rangle$ should be negative, that is guaranteed in Eq. (5.7) by the inequality (5.2). The left-hand sides of the inequalities (5.8), bound the solutions of (5.7) by

$$\begin{aligned} \rho_1/2\rho_2 &> (\alpha_3|\gamma_1^2| + \gamma_3^1\gamma_2^3)(\det D_3^1)^{-1}, \\ \rho_3/2\rho_2 &> (\alpha_1|\gamma_2^3| + \gamma_3^1\gamma_1^2)(\det D_3^1)^{-1}. \end{aligned} \tag{5.9}$$

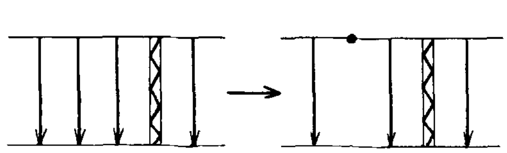


FIG. G5-1. G5-2. NM situation of the $(-,-,-)$ type for the Pati–Salam model.

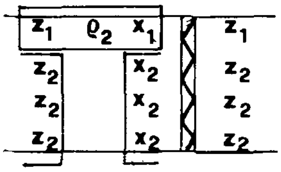


FIG. G5-3. Equivalent-sets variables.

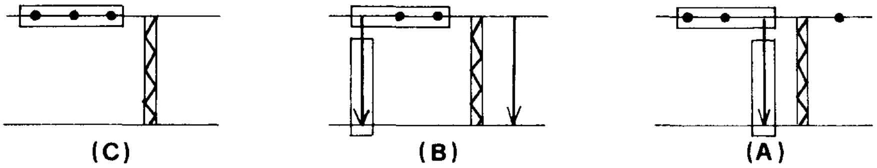


FIG. G5-4. PM situation of the (–, –, –) type.

To summarize: in the (–, –, –) mode, it is seen that the signs of all six coupling constants are fixed by (5.2), where their order of magnitude is partially fixed by the inequalitites (5.3) and (5.9). This means that α_2 and two of the other five can be fixed within broad limits, while the other three are constrained by upper (or lower) bounds (5.9). It seems that with the actual values[1] of ρ^6 the γ_2^3 and γ_1^2 should be much smaller than all the other coupling constants in the theory. In the case where the inequalities (5.9) are not satisfied, the minimum point in R_N is outside of $R_{\bar{N}}$ and one of the following symmetry breakings, as a solution to the PM situation, is predicted in Fig. G5-4. In order to avoid PM solutions we must be careful when fixing the other eight coupling constants of the whole potential (2.2), so that one gets the proper magnitudes ρ^6 of the breakings (5.1).

It is seen that much freedom remains in any step and each coupling constant is free to be changed in some domain without spoiling the structure of the symmetry breaking. The correct way of fixing these coupling constants is to try to verify that this domain is compatible with the corrections due to higher order contributions.

We go over now to the second diagonalized mode (–, +, +) which also leads to the (5.1) breaking. The basic graph is shown in Fig. G5-5(A) and it should transform to Fig. G5-5(B). The (+, –, +), (+, +, –) modes are not considered here.) Such a transition can take place in a few basic ways. The first one occurs when α_2 is negative, the other happens when some of the matrices D_1^2, D_2^3, D has negative determinants. The treatment of the first situation is equivalent to that of the (–, –, –) mode. Therefore we concentrate here on another situation described by (5.10) and (5.11)

$$\gamma_2^3, \gamma_1^2 > 0, \quad \gamma_3^1 < 0,$$

$$\alpha_1, \alpha_2, \alpha_3 > 0, \tag{5.10}$$

$$\epsilon_3^1 = (+), \quad \epsilon_2^3 = \epsilon_1^2 = (-). \tag{5.11}$$

The situation in which singlet negative determinants are permitted is carried out in the same way as that in which such determinants are excluded; hence we choose to treat the second case, which is simpler. Restriction (5.11) should be changed to a stronger one, (5.12), according to the discussion in Ref. 3, in order to prohibit these singlets

$$\epsilon_3^1 = (+), \ \det(D_1^2) < -2\alpha_1\alpha_2, \quad \det(D_2^3) < -2\alpha_2\alpha_3. \tag{5.12}$$

The structure of the breaking with the above condition is one of those shown in Fig. G5-6. The potential $\bar{V}$ in the Fig. G5-6(B) case is a function of n [the length (ρ_1, ρ_3) on the $\mathbf{x}^1, \mathbf{x}^3$ vectors taken to be constant]

$$\bar{V}^B(n) = \frac{\alpha_1}{n}\rho_1^2 + \frac{\alpha_3}{n}\rho_3^2 + \frac{2\gamma_3^1}{n}\rho_3\rho_1 + \frac{\alpha_2}{4-n}\rho_2^2$$

$$= \frac{d_1}{n} + \frac{d_2}{4-n}, \quad d_1, d_2 > 0. \tag{5.13}$$

$\bar{V}$ has a minimum point n_0 in the region $(0 < n < 4)$, this point should be fixed to be in the region $(3 < n < 4)$, in order to get the desired breaking. It is found from (5.13) to be

$$n_0 = 4\left(\frac{1-\sqrt{\alpha}}{1-\alpha}\right), \quad \alpha = \frac{d_2}{d_1} < 1. \tag{5.14}$$

Hence the ratio of d_2 to d_1 is bounded by

$$0 < \alpha < \tfrac{1}{9}. \tag{5.15}$$

With these conditions the symmetry breaking should occur either in the form G5-6(A) or in the form G5-6(B) with $n_0 = 3$. In order to exclude the second possibility, we shall calculate the minimum values of the potential in both cases and compare them. The G5-6(B) minimum is

$$V_{\min}^B = \frac{\alpha_1}{3}\rho_1^2 + \frac{\alpha_3}{3}\rho_3^2 + \frac{2\gamma_3^1}{3}\rho_1\rho_3 + \alpha_2\rho_2^2. \tag{5.16}$$

The minimum point of V^A is found in the same way as that of the potential of (5.4) and it is given by (5.7). In the present mode one gets positive $\langle\sigma\rangle$ and $\langle\nu\rangle$ and the upper bound should be imposed

$$\rho_1/6\rho_2 > (\alpha_3\gamma_1^2 + |\gamma_3^1|\,\gamma_2^3)(\det D_3^1)^{-1},$$

$$\rho_3/6\rho_2 > (\alpha_1\gamma_2^3 + |\gamma_3^1|\,\gamma_1^2)(\det D_3^1)^{-1}. \tag{5.17}$$

A straightforward calculation shows that in this situation $V_{\min}^A$ is lower than $V_{\min}^B$, hence conditions (5.10), (5.12), (5.15), (5.17) predict the desired breaking. Similar conclusions to those in the previous examples are obtained where the main actual difficulty should arise from the low bounds $(\rho_1/12\rho_2)$ and $(\rho_3/12\rho_2)$, where $\rho_2 \gg \rho_1$, $\rho_2 \gg \rho_3$. The main physical consequence of the violation of(5.17), which predicts $\langle x_1\rangle = \langle z_1\rangle = 0$ [Fig. G5-6(B)] is the conservation of baryon and lepton numbers separately.[1] Hence the proton can become stable as a consequence of the pseudominimum situation.

To conclude we should like to know what could be the consequences of higher order calculations, on the one hand, and changing of the environments, on the other, e.g., external fields,[4] high temperature,[5] or density.[6] In the present model the "restoration" phenomena can arise as a consequence of imposing NM or PM situations (related phenomena are described by Weinberg[7]).

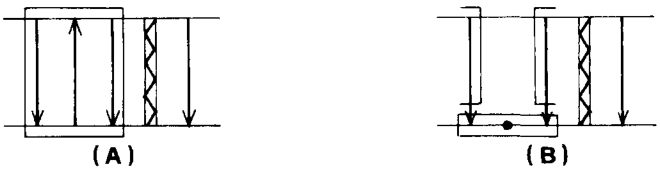


FIG. G5-5. NM situation of the (–, +, +) type in the Pati-Salam model.

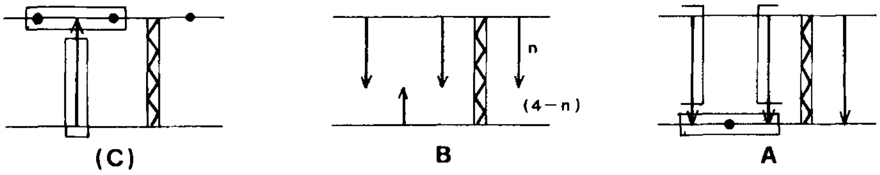


FIG. G5-6 Possible NM situation imposed by (5.12).

This can occur in different ways and predicts transitions between different breakings through first-or second-order phase transitions.[8] In order to illustrate it, we go back to the Pati–Salam breaking in the (−, −, −) mode. Considering the lowest values of $\langle \rho_1/\rho_2\rangle$ and $\langle \rho_3/\rho_2\rangle$ it seems reasonable to assume that $\langle\sigma\rangle$ and $\langle\nu\rangle$ are in the neighborhood of the point $\langle\sigma\rangle=\langle\nu\rangle=-\frac{1}{4}$. As a consequence it might follow that different "environments" will predict different hierarchies in the breaking, or second-order phase transition by passing to other domains in the $\sigma_{\min}$, $\nu_{\min}$ plane. These transitions are presented in Fig. G5-7. The domains in the graph G5-7 represent the following residual symmetries:

I. $U(3)_{R+L+C}\otimes U(1)_{R+L+C}$,

II. $U(3)_{L+C}\otimes U(3)_R\otimes U(1)_{R+L+C}$,

III. $U(3)_C\otimes U(3)_{R+L}\otimes U(1)_{R+L+C}$,

IV. $U(3)_C\otimes U(3)_R\otimes U(3)_L\otimes U(1)_{C+L+R}$.

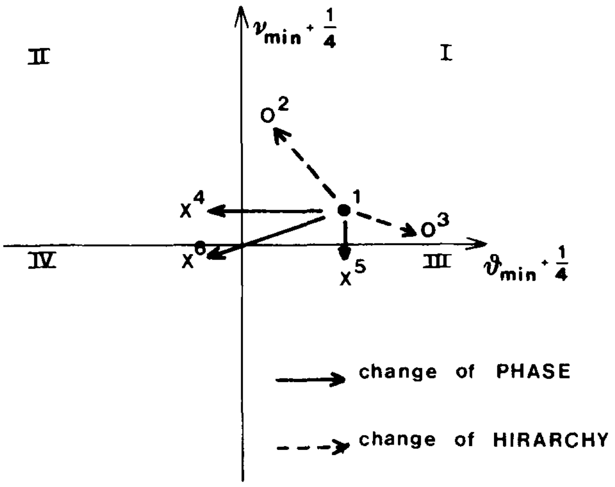


FIG. G5-7. $\sigma_{\min}$ and $\nu_{\min}$ in R_{N*} space.

The point (1) in the graph represents the actual Pati–Salam breaking, where the arrows to (2) and (3) represent transitions to other hierarchies of symmetry breaking according to

$$(2)\ U(3)_R \underset{\text{IV}\to\text{II}\to\text{I}}{\otimes} U(3)_L\otimes U(3)_C\otimes U(1)_{R+L+C}\to U(3)_{C+L}$$
$$\otimes U(3)_R\otimes U(1)_{R+L+C}\to U(3)_{R+L+C}\otimes U(1)_{R+C+L},$$

$$(3)\ U(3)_C \underset{\text{IV}\to\text{III}\to\text{I}}{\otimes} U(3)_L\otimes U(3)_R\otimes U(1)_{(C+L+R)}\to U(3)_C$$
$$\otimes U(3)_{R+L}\otimes U(1)_{R+L+C}\to U(3)_{R+L+C}$$
$$\otimes U(1)_{R+L+C}.$$

The arrows from (1) to (4), (5), (6) represent possible phase transition to II, III, IV, respectively.

We hope to investigate the dynamics of these phenomena in detail in a future article and to learn if and how they occur.

ACKNOWLEDGMENTS

The author would like to thank Professor Abdus Salam and Professor J. Strathdee for useful discussions and comments. Special thanks are due to Professor J. Strathdee for a critical reading of the manuscript. He would also like to thank the International Atomic Energy Agency and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste.

[1] J.C. Pati and A. Salam, Phys. Rev. D 10, 275 (1974).
[2] L.-F. Li, Phys. Rev. D 9, 1723 (1974).
[3] A. Schorr, J. Math. Phys. 18, 1627 (1977).
[4] A. Salam and J. Strathdee, Nucl. Phys. B 90, 230 (1975).
[5] D.A. Kirzhnits and A.D. Linde, Phys. Lett. B 42, 471 (1972); L. Doland and R. Jackiw, Phys. Rev. D 9, 3320 (1974).
[6] T.D. Lee and M. Margolis, Phys. Rev. D 11, 1591 (1975).
[7] S. Weinberg, Phys. Rev. D 9, 3357 (1974).
[8] We assume here that the higher order corrections will not lead to other types of phases.

Correlation functions in the spherical and mean spherical models

Mark Kac

The Rockefeller University, New York, New York 10021

Colin J. Thompson

Department of Mathematics, University of Melbourne, Parkville, Victoria 3052, Australia

(Received 18 February 1977)

A transformation is obtained relating spherical and mean and spherical averages. The kernel of the transformation is the probability density of $N^{-1}\Sigma_{i=1}^{N} x_i{}^2$ in the mean spherical model. The transformation is inverted to obtain a simple method for computing spherical averages from mean spherical averages. Averages in the two ensembles are identical except in zero field below the critical temperature.

1. INTRODUCTION

Since its inception in 1952[1] the spherical model has enjoyed considerable popularity[2] as an exactly soluble model displaying critical phenomena.

The model consists of a set of N "spins" $-\infty < x_i < \infty$, $i = 1, 2, \ldots, N$ located on the vertices of a lattice with "Hamiltonian"

$$H = -\sum_{i<j} \rho_{ij} x_i x_j - H \sum_i x_i \tag{1.1}$$

and the spins subject to the spherical constraint

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{N} x_i^2 = N. \tag{1.2}$$

The canonical partition function for the model is given by

$$Q_s(N, \beta, H) = \int \exp\left(\tfrac{1}{2}\beta \sum_{i,j=1}^{N} \rho_{ij} x_i x_j + \beta H \sum_{i=1}^{N} x_i\right) d\sigma_{\sqrt{N}}, \tag{1.3}$$

where the integral is taken over the surface of the N-dimensional sphere (1.2) and for simplicity we have assumed that $\rho_{ij} = \rho_{ji}$ and set $\rho_{ii} = 0$.

2. REVIEW OF SOME OLD RESULTS

The partition function (1.3) is usually evaluated by writing

$$\begin{aligned} Q_s(N, \beta, H) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta(\|\mathbf{x}\| - N^{1/2}) \\ &\quad \times \exp\left(\tfrac{1}{2}\beta \sum_{i,j=1}^{N} \rho_{ij} x_i x_j + \beta H \sum_{i=1}^{N} x_i\right) dx_1 \cdots dx_N \\ &= 2N^{1/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta\left[\sum_{i=1}^{N} x_i^2 - N\right] \\ &\quad \times \exp\left(\tfrac{1}{2}\beta \sum_{i,j=1}^{N} \rho_{ij} x_i x_j + \beta H \sum_{i=1}^{N} x_i\right) dx_1 \cdots dx_N \end{aligned} \tag{2.1}$$

and using the integral representation

$$\delta(x) = (1/2\pi i) \int_{-i\infty}^{i\infty} \exp(-sx)\, ds \tag{2.2}$$

to obtain

$$Q_s(N, \beta, H) = (2N^{1/2}/2\pi i) \int_{s_0 - i\infty}^{s_0 + i\infty} \exp(Ns) Q_{ms}(N, \beta, H, s)\, ds, \tag{2.3}$$

where

$$\begin{aligned} &Q_{ms}(N, \beta, H, s) \\ &\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-s \sum_{i=1}^{N} x_i^2 + \tfrac{1}{2}\beta \sum_{i,j=1}^{N} \rho_{ij} x_i x_j \right. \\ &\qquad \left. + \beta H \sum_{i=1}^{N} x_i\right) dx_1 \cdots dx_N. \end{aligned} \tag{2.4}$$

The Gaussian integrals in (2.4) are readily performed, and finally the integral in (2.3) is evaluated by the method of steepest descents.[1]

For the sake of definiteness we will assume that ρ_{ij} depends only on $\mathbf{r}_i - \mathbf{r}_j$, where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the position vectors of lattice points i and j on a regular hypercubical lattice. In this case the quadratic form in the exponent of (2.4) is easily diagonalized, and one obtains

$$\begin{aligned} Q_{ms}(N, \beta, H, s) &= (2\pi/\beta)^{N/2} \prod_{\mathbf{q}} [z - \lambda(\mathbf{q})]^{-1/2} \\ &\quad \times \exp\{(N\beta H^2/2)[z - \lambda(0)]^{-1}\}, \end{aligned} \tag{2.5}$$

where

$$z = 2s/\beta, \tag{2.6}$$

$\lambda(\mathbf{q})$ are the eigenvalues of the matrix ρ, i.e., the Fourier coefficients of $\rho(\mathbf{r})$, and the product over $\mathbf{q}$ is taken over allowed wave vectors in the first Brillouin zone of the reciprocal lattice.[2]

In the thermodynamic limit one then obtains for the canonical free energy[1,2]

$$\begin{aligned} -\beta\psi &= \lim_{N\to\infty} N^{-1} \log Q_s(N, \beta, H) \\ &= \tfrac{1}{2}\log(2\pi/\beta) + \tfrac{1}{2}\beta H^2 [z_s - \lambda(0)]^{-1} + \tfrac{1}{2}[\beta z_s - f(z_s)], \end{aligned} \tag{2.7}$$

where, in d dimensions,

$$f(z) = (2\pi)^{-d} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \log[z - \lambda(\boldsymbol{\theta})]\, d\theta_1 \cdots d\theta_d. \tag{2.8}$$

The saddle point z_s is determined from

$$\beta = \beta H^2[z_s - \lambda(0)]^{-1} + (2\pi)^{-d} \int_{-\pi}^{\pi} \cdots$$
$$\times \int_{-\pi}^{\pi} [z_s - \lambda(\theta)]^{-1}\, d\theta_1 \cdots d\theta_d, \tag{2.9}$$

and, in order for the above steps to be valid, one needs

$$z > \lambda(0) = \max \lambda(\mathbf{q}). \tag{2.10}$$

3. THE MEAN SPHERICAL MODEL

Shortly after the original paper[1] appeared, it was pointed out by Lewis and Wannier[3] that the steepest descent method could be avoided by considering a grand canonical rather than canonical ensemble in which the variable s in (2.4) is now conjugate to $\sum x_i^2$. That is, the grand canonical partition function is given by (2.4) with s, or equivalently z defined by (2.6), fixed by the condition

$$\left\langle \sum_{i=1}^{N} x_i^2 \right\rangle = -\frac{\partial}{\partial s} \log Q_{ms}(N, \beta, H, s) = N \tag{3.1}$$

so that the spherical constraint (1.2) is satisfied now in the mean; hence the term mean spherical model. In reality, of course, there is only one model. The terms spherical and mean spherical refer to the two ensembles, canonical and grand canonical, respectively. We will, however, adopt the common practice of referring to two models.

In the thermodynamic limit the condition (3.1) reduces to (2.9), and hence the spherical and mean spherical models are thermodynamically equivalent. It was quickly realized,[4] however, that thermodynamic averages, such as $\langle x_1^4 \rangle$ are different in the two ensembles, and a little later a more serious discrepancy, involving the zero field probability distribution for a single spin, was noted.[5] More recently it was shown[6] that consistent results for the two ensembles are obtained in nonzero field and even in zero field provided one takes the limit $H \to 0$ after the thermodynamic limit. Our aim here is to further investigate the relationship between the two models in zero field. In the following sections we obtain an expression for the probability distribution of $N^{-1}\sum_{i=1}^{N} x_i^2$ in the mean spherical model (which, not surprisingly, is not simply a delta function) and a transformation relating spherical and mean spherical averages. In nonzero field and in zero field above the critical temperature the averages are identical, so that the transformation is only nontrivial in zero field below the critical temperature.

4. THE DISTRIBUTION OF $N^{-1}\Sigma_{i=1}^{N} x_i^2$ IN THE MEAN-SPHERICAL MODEL

In zero field the saddle point condition (2.4) becomes

$$\beta = (2\pi)^{-d} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} [z_s - \lambda(\theta)]^{-1}\, d\theta_1 \cdots d\theta_d. \tag{4.1}$$

When [cf. (2.10)]

$$\beta_c = (2\pi)^{-d} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} [\lambda(0) - \lambda(\theta)^{-1}\, d\theta_1 \cdots d\theta_d < \infty, \tag{4.2}$$

the expression (2.7) for the free energy and the condition (4.1) only holds for $\beta < \beta_c$ (i.e., high temperatures). When $\beta > \beta_c$, the saddle point "sticks" at $z_s = \lambda(0)$. A similar situation occurs for the mean spherical model, but now some care needs to be exercised in proceeding to the thermodynamic limit. For finite N the mean spherical constraint (3.1) in zero field becomes, from (2.5),

$$\beta = N^{-1} \sum_{\mathbf{q}} (z - \lambda(\mathbf{q}))^{-1} \tag{4.3}$$

and as long as $\beta < \beta_c$, (4.3) becomes (4.1) in the limit $N \to \infty$. When $\beta > \beta_c$, one must separate off the $\mathbf{q} = 0$ term in (4.3) and set

$$z = \lambda(0) + [N(\beta - \beta_c)]^{-1}. \tag{4.4}$$

Then in the thermodynamic limit z sticks once more at $\lambda(0)$ and identical thermodynamic properties for the two models are obtained. The distribution of spins, however, is not at all well behaved in zero field.

In order to investigate the zero field probability distribution for $N^{-1}\sum_{i=1}^{N} x_i^2$ we consider the mean spherical average characteristic function $\langle \exp(i\xi N^{-1}\sum_{i=1}^{N} x_i^2)\rangle_{ms}$. From (2.4) and (2.6) we readily obtain

$$\left\langle \exp\left(i\xi N^{-1} \sum_{i=1}^{N} x_i^2 \right) \right\rangle_{ms} = \frac{Q_{ms}(N, z - 2i\xi(\beta N)^{-1})}{Q_{ms}(N, z)}, \tag{4.5}$$

where

$$Q_{ms}(N, z) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}\beta \sum_{i,j=1}^{N} (z\delta_{ij} - \rho_{ij}) x_i x_j \right)$$
$$\times dx_1 \cdots dx_N. \tag{4.6}$$

Using (2.5), it is easily verified that for $\beta < \beta_c$ and $N \to \infty$

$$\left\langle \exp\left(i\xi N^{-1} \sum_{i=1}^{N} x_i^2 \right) \right\rangle_{ms} \sim \exp\left(i\xi(\beta N)^{-1} \sum_{\mathbf{q}} [z - \lambda(\mathbf{q})]^{-1} \right)$$
$$= \exp(i\xi) \quad (\beta < \beta_c), \tag{4.7}$$

where use has been made of (4.3). When $\beta > \beta_c$, the $\mathbf{q} = 0$ terms must be separated off. Then on substituting (4.4) one obtains

$$\left\langle \exp\left(i\xi N^{-1} \sum_{i=1}^{N} x_i^2 \right) \right\rangle_{ms} \sim \left[1 - 2i\xi \left(1 - \frac{\beta c}{\beta} \right) \right]^{-1/2}$$
$$\times \exp\left(i\xi(\beta N)^{-1} \sum_{\mathbf{q}\neq 0} [\lambda(0) - \lambda(\mathbf{q})]^{-1} \right.$$
$$\sim \left[1 - 2i\xi \left(1 - \frac{\beta c}{\beta} \right) \right]^{-1/2} \exp(i\xi\beta_c/\beta), \tag{4.8}$$

where use has been made of (4.2). The above results may be summarized by writing

$$\lim_{N\to\infty} \left\langle \exp\left(i\xi N^{-1} \sum_{i=1}^{N} x_i^2 \right) \right\rangle_{ms} = \int_0^{\infty} \exp(i\xi x) K(x, \beta)\, dx \tag{4.9}$$

where for $\beta < \beta_c$

$$K(x, \beta) = \delta(x - 1) \tag{4.10}$$

and for $\beta > \beta_c$

$$K(x, \beta) = \begin{cases} [2\pi(\beta x - \beta_c)(\beta - \beta_c)]^{-1/2}\beta \exp[-(\beta x - \beta_c)/2(\beta - \beta_c)], \\ \qquad\qquad x > \beta_c/\beta, \\ 0, \quad x < \beta_c/\beta. \end{cases} \tag{4.11}$$

The probabilistic interpretation of the kernel K is clear:

$$K(x,\beta)\,dx=\lim_{N\to\infty}\operatorname{Prob}\left\{x<N^{-1}\sum_{i=1}^{N}x_i^2<x+dx\right\}. \tag{4.12}$$

As expected, the distribution is concentrated at the spherical value for $\beta<\beta_c$. This is no longer the case, however, below the critical temperature $\beta>\beta_c$.

Before deriving a transformation relating averages between the two models it is perhaps worth noting at this point that a similar situation to the above holds for the ideal Bose gas when one compares the canonical and grand canonical descriptions.[7,8] Thus in the grand canonical ensemble the probability distribution for the density N/V

$$\sigma(x)\,dx=\lim_{\substack{N,V\to\infty\\ \rho=N/V}}\operatorname{Prob}\{x<\rho<x+dx\} \tag{4.13}$$

is given for $\beta<\beta_c$ by

$$\sigma(x)=\delta(x-\rho) \tag{4.14}$$

and for $\beta>\beta_c$ by

$$\begin{cases}\sigma(x)= & (\rho-\rho_c)\exp[-(x-\rho_c)/(\rho-\rho_c)], \quad x>\rho_c,\\ & 0, \qquad\qquad\qquad\qquad\qquad x<\rho_c.\end{cases} \tag{4.15}$$

The similarity of the ideal Bose gas and the spherical model has of course been known[9] for some time.

5. RELATION BETWEEN AVERAGES

For a function $f(x_1,\ldots,x_N)$ we define the zero field mean spherical average by

$$f_{ms}^{(N)}=[Q_{ms}(N,z)]^{-1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,\ldots,x_N)$$
$$\times\exp\left(-\tfrac{1}{2}\beta\sum_{i,j=1}^{N}(z\delta_{ij}-\rho_{ij})x_ix_j\right)dx_1\cdots dx_N \tag{5.1}$$

with z determined from (4.3), and the spherical average by

$$f_s^{(N)}=[Q_s(N,\beta)]^{-1}\int f(x_1,\ldots,x_N)$$
$$\times\exp\left(-\tfrac{1}{2}\beta\sum_{i,j=1}^{N}\rho_{ij}x_ix_j\right)d\sigma_{\sqrt{N}}. \tag{5.2}$$

Beginning with (5.1), we write

$$f_{ms}^{(N)}=\sum_{k=0}^{\infty}\int\underset{k\Delta\xi<\|\mathbf{x}\|^2/N<(k+1)\Delta\xi}{\cdots}\int f(x_1,\ldots,x_N)$$
$$\times\exp\left(-\tfrac{1}{2}\beta\sum_{i,j=1}^{N}(z\delta_{ij}-\rho_{ij})x_ix_j\right)$$
$$\times dx_1\cdots dx_N[Q_{ms}(N,z)]^{-1}. \tag{5.3}$$

Multiplying and dividing the summand in (5.3) by

$$\int\underset{k\Delta\xi<\|\mathbf{x}\|^2/N<(k+1)\Delta\xi}{\cdots}\int\exp\left(-\tfrac{1}{2}\beta\sum_{i,j=1}^{N}(z\delta_{ij}-\rho_{ij})x_ix_j\right)$$
$$\times dx_1\cdots dx_N \tag{5.4}$$

and taking the limit $\Delta\xi\to0$ then yields

$$f_{ms}^{(N)}=\int_0^{\infty}f_s^{(N)}(\xi)\,d\tau_N(\xi), \tag{5.5}$$

where

$$\tau_N(\xi)=[Q_{ms}(N,z)]^{-1}$$
$$\times\int\underset{\|\mathbf{x}\|^2/N<\xi}{\cdots}\int\exp\left(-\tfrac{1}{2}\beta\sum_{i,j=1}^{N}(z\delta_{ij}-\rho_{ij})x_ix_j\right)$$
$$\times dx_1\cdots dx_N \tag{5.6}$$

and $f_s^{(N)}(\xi)$ is the spherical average of $f(x_1\cdots x_N)$ taken over the sphere $\|\mathbf{x}\|^2=N\xi$, i.e.,

$$f_s^{(N)}(\xi)=[Q_s(N,\xi,\beta)]^{-1}\int f(x_1,\ldots,x_N)$$
$$\times\exp\left(\tfrac{1}{2}\beta\sum_{i,j=1}^{N}\rho_{ij}x_ix_j\right)d\sigma_{\sqrt{N\xi}} \tag{5.7}$$

with

$$Q_s(N,\xi,\beta)=\int\exp\left(\tfrac{1}{2}\beta\sum_{i,j=1}^{N}\rho_{ij}x_ix_j\right)d\sigma_{\sqrt{N\xi}}. \tag{5.8}$$

Finally, taking the limit $N\to\infty$, assuming that

$$f_{ms}=\lim_{N\to\infty}f_{ms}^{(N)}\quad\text{and}\quad f_s(\xi)=\lim_{N\to\infty}f_s^{(N)}(\xi) \tag{5.9}$$

exist and that limits can be interchanged, we obtain, from (5.5) and (5.6),

$$f_{ms}=\int_0^{\infty}f_s(\xi)K(\xi,\beta)\,d\xi, \tag{5.10}$$

where we have used that fact that with z determined from (4.3)

$$\lim_{N\to\infty}\tau_N(\xi)=\int_0^{\xi}K(\eta,\beta)\,d\eta, \tag{5.11}$$

where $K(\eta,\beta)$ is the distribution function defined by (4.12) and given by (4.10) and (4.11).

The transformation (5.10) is exactly what one might have expected. In particular, it is to be noted that above the critical point $\beta<\beta_c$, $K(\xi,\beta)=\delta(\xi-1)$ and hence from (5.7) and (5.2)

$$f_{ms}=f_s\quad(\beta<\beta_c). \tag{5.12}$$

Below the critical point $(\beta>\beta_c)$ the transformation (5.10) is, of course, nontrivial.

6. AN APPLICATION

Since it is always easier to calculate mean spherical rather than spherical averages, the transformation (5.10) should be inverted as a practical means of evaluating spherical averages. We will consider a particular class of functions here which covers most cases of interest, namely f a homogeneous function of some fixed degree, $2p$, say, so that

$$f(\lambda x_1,\lambda x_2,\ldots,\lambda x_N)=\lambda^{2p}f(x_1,x_2,\ldots,x_N) \tag{6.1}$$

for all λ. As an example, $2p$-spin averages satisfy (6.1).

In order to highlight the temperature dependence, we write

$$f_s(\xi)=f_s(\xi,\tau),\quad\tau=\beta_c/\beta \tag{6.2}$$

for spherical averages defined by (5.7). The ordinary spherical average defined by (5.2) is obtained by setting

$\xi = 1$ in (6.2), and by changing variables $x_i \to \xi^{1/2} x_i$ in (5.7) it readily follows from (6.1) that

$$f_s(\xi) = \xi^p f_s(1, \tau/\xi). \tag{6.3}$$

From (4.10) it follows immediately that

$$f_s(1, \tau) = f_{ms}(\tau) \quad \text{when} \quad \tau > 1 \tag{6.4}$$

so that, above the critical temperature, averages are identical for the two models. When $\tau < 1$, however, (4.11) and (5.10) give

$$f_{ms}(\tau) = \int_\tau^\infty \xi^p f_s(1, \tau/\xi) \exp[-(\xi - \tau)/2(1-\tau)] \times [2\pi(\xi - \tau)(1-\tau)]^{-1/2} d\xi. \tag{6.5}$$

To invert (6.5), we make the change of variables

$$\xi = \tau(x+1) \tag{6.6}$$

to obtain

$$\tilde{f}_{ms}(S) = [2s/(1+2s)]^p (s/\pi)^{1/2} \times \int_0^\infty f_s(1, (1+x)^{-1})\, x^{-1/2}(1+x)^p \exp(-sx)\, dx, \tag{6.7}$$

where

$$s = \tau/2(1-\tau) \tag{6.8}$$

and

$$\tilde{f}_{ms}(s) = f_{ms}(2s/(1+2s)). \tag{6.9}$$

Inverting the Laplace transform (6.7) then gives, on replacing $(1+x)^{-1}$ by τ,

$$f_s(1, \tau) = \tau^p (1/\tau - 1)^{1/2} (2\pi i)^{-1} \times \int_{c-i\infty}^{c+i\infty} (1 + 1/2s)^p (\pi/s)^{1/2} \tilde{f}_{ms}(s) \exp[s(1/\tau - 1)] ds. \tag{6.10}$$

As an example consider the case

$$f(x_1, \ldots, x_N) = N^{-1} \sum_{i=1}^N x_i^4 \tag{6.11}$$

on a three-dimensional lattice with nearest neighbour interactions only. It is known[4] that

$$f_{ms}(\tau) = 3 \quad \text{for all } \tau, \tag{6.12}$$

whereas for the spherical model[1]

$$f_s(1, \tau) = \begin{matrix} 3, & \tau > 1, \\ 3 - 2(1-\tau)^2, & \tau < 1. \end{matrix} \tag{6.13}$$

Equation (6.13) follows immediately from (6.10) by noting that $p = 2$ in this case, and

$$(2\pi i)^{-1} \int_{c-i\infty}^{c+i\infty} (1 + 1/2s)^2 (\pi/s)^{1/2}\, 3 \exp(sx)\, ds = x^{-1/2}(3 + 6x + x^2), \tag{6.14}$$

thus closing the cycle back to the original "discrepancy" between the two models, noted by Lewis and Wannier.[4]

ACKNOWLEDGMENTS

Colin J. Thompson would like to thank the Rockefeller University for their hospitality during a brief visit where part of this work was carried out.

[1] T.H. Berlin and M. Kac, Phys. Rev. 86, 821 (1952).
[2] G.S. Joyce, in *Phase Transitions and Critical Phenomena Vol.* 2, edited by C. Domb and M.S. Green (Academic, London, 1973).
[3] H.W. Lewis and G.H. Wannier, Phys. Rev. 88, 682 (1952).
[4] H.W. Lewis and G.H. Wannier, Phys. Rev. 90, 1131E (1953).
[5] M. Lax, Phys. Rev. 97, 1419 (1955).
[6] C.C. Yan and G.H. Wannier, J. Math. Phys. 6, 1833 (1965).
[7] M. Kac, in *The Physicist's Conception of Nature*, edited by J. Mehra (Reidel, Dordrecht, 1973), p. 521.
[8] J.T. Cannon, Commun. Math. Phys. 29, 89 (1973).
[9] J. Ford and T.H. Berlin, J. Chem. Phys. 27, 931 (1957).

Solitons on moving space curves[a)]

G. L. Lamb, Jr.

Department of Mathematics and Optical Sciences Center, The University of Arizona, Tucson, Arizona 85721

(Received 4 April 1977)

It is shown that the motion of certain types of helical space curves may be related to the sine–Gordon equation and to the Hirota equation (and consequently to the nonlinear Schrödinger equation and to the modified Korteweg–de Vries equation). The intrinsic equations that govern the motion of space curves are shown to provide the various linear equations that have been introduced to solve these evolution equations by inverse scattering methods.

I. INTRODUCTION

Although the partial differential equations that exhibit soliton behavior have been solved in a variety of ways during the past decade,[1–14] the underlying reasons why the known soliton equations should be the elite equations that are endowed with such remarkable properties have remained unclear. Some success in this matter has recently been achieved by utilizing the relativistic invariance of the sine—Gordon equation[15] as well as concepts taken from group theory[16] and modern differential geometry.[17,18] The purpose of this paper is to offer a somewhat more elementary geometric approach by showing that certain of the more common nonlinear evolution equations are precisely the ones that arise in a consideration of the motion of certain types of helical space curves. When the evolution equations are viewed in this context, the linear equations that have previously been associated with them in an ad hoc manner for solution by inverse scattering methods are found to arise in a natural way from the standard intrinsic equations that govern the motion of twisted space curves.

The development presented here is a natural extension of a result obtained by Hasimoto,[19] who showed that the intrinsic equations governing the curvature and torsion of an isolated thin vortex filament moving without stretching in an incompressible inviscid fluid can be reduced to a nonlinear Schrödinger equation. The many similarities in the properties of the various soliton equations immediately tempt one to conjecture that other equations exhibiting soliton behavior may also be related to helical space curves. In the present paper this is shown to be the case. The sine—Gordon equation is shown to be related to curves of either constant curvature or constant torsion with the curvature (or torsion) playing the role of eigenvalue parameter in the inverse scattering formulism. A certain class of curves in which both curvature and torsion may vary are found to be related to a soliton equation that was first obtained by Hirota.[20] The nonlinear Schrödinger equation and the modified Korteweg—de Vries equation are special cases of this equation. The modified Korteweg—de Vries equation is obtained for the case in which the torsion is required to have a constant value which then plays the role of the eigenvalue parameter in the inverse scattering formulism. For the Hirota equation and for the nonlinear Schrödinger equation, it is the asymptotic value assumed by the torsion at large distances from the disturbance that is the eigenvalue parameter.

The present paper provides an approach to these evolution equations that, in some respects, is similar to one used previously for treating soliton behavior in coherent optical pulse propagation.[21] In that context the Bloch equations, which have the form of a single component of the vector Serret—Frenet equations,[22] provide a natural motivation for the introduction of linear equations to which inverse scattering techniques may then be applied. In addition, the physical process of population inversion of the two-level atoms and subsequent return of this inverted population to the ground state provides a natural motivation for associating the notion of vanishing reflection coefficient with lossless (soliton) propagation when applying inverse scattering methods. In the present instance, the orientation of the trihedral of tangent, normal, and binormal vectors must return to its original orientation after passage of the loop of helical motion along the space curve. This may be used to motivate the association of vanishing reflection coefficient with soliton propagation for the equations under consideration here.

In Sec. II the fundamental equations governing the motion of twisted space curves are introduced and then combined to yield the general relations that are ultimately specialized to yield the various evolution equations. In Sec. III, specializations are carried out for the sine—Gordon equation and for the Hirota equation. In the course of effecting these specializations, certain auxiliary functions are also obtained. In Sec. IV, the intrinsic equations governing the motion of twisted curves, which are linear vector equations, are used to obtain various sets of linear scalar equations. These equations are precisely the equations that have previously been associated with the evolution equations for solution by inverse scattering methods. The coefficients in these equations are the auxiliary functions referred to above. Hence the specialization procedure used to obtain the evolution equations also immediately yields the linear equations. In Sec. V, various linear equations associated with the equations obtained in Sec. III are listed. As is to be expected, contact with the Korteweg—de Vries equation may be obtained through the Miura transformation.[23] Finally, the velocity with which the curves move is considered in Sec. VI.

[a)]This work was partly sponsored by the National Science Foundation under Contract No. GP43070 with the University of Arizona.

II. INTRINSIC EQUATIONS FOR MOVING SPACE CURVES

The motion of a twisted curve may be described by specifying the curvature and torsion of each point on the curve as a function of time. At each instant the spatial variations of the unit tangent, normal, and binormal vectors **t**, **n**, and **b**, respectively, are given by the Serret–Frenet equations[22]

$$\mathbf{t}_s = \kappa \mathbf{n}, \tag{2.1a}$$

$$\mathbf{b}_s = -\tau \mathbf{n}, \tag{2.1b}$$

$$\mathbf{n}_s = \tau \mathbf{b} - \kappa \mathbf{t}. \tag{2.1c}$$

The subscript denotes partial differentiation with respect to the arc length parameter s. The curvature κ and torsion τ are now functions of time as well as s.

The latter two of Eqs. (2.1) are conveniently combined into the complex form

$$(\mathbf{n} + i\mathbf{b})_s + i\tau(\mathbf{n} + i\mathbf{b}) = -\kappa \mathbf{t}. \tag{2.2}$$

If the torsion approaches a constant value τ_0 in regions of the curve that are remote from the disturbances of interest, then Eq. (2.2) may be rewritten

$$\mathrm{N}_s + i\tau_0 \mathrm{N} = -\psi \mathbf{t}, \tag{2.3}$$

where

$$\mathrm{N} = (\mathbf{n} + i\mathbf{b}) \exp\left[i \int_{-\infty}^{s} ds'(\tau - \tau_0)\right] \tag{2.4}$$

and

$$\psi = \kappa \exp\left[i \int_{-\infty}^{s} ds'(\tau - \tau_0)\right]. \tag{2.5}$$

Since the complex quantity $\psi(s, t)$ contains both curvature and torsion, it provides a complete description of the twisted curve. It is for this quantity, or functions closely related to it, that various evolution equations will be obtained. Extraction of the term τ_0 is convenient at a later stage in the analysis when the linear equations associated with inverse scattering methods are considered. If one's primary concern, however, is with the geometric configuration of the curves themselves, then the definition of $\psi(s, t)$ used by Hasimoto[19] may be preferable.

In terms of N and ψ, the first of the Serret–Frenet equations takes the form

$$\mathbf{t}_s = \tfrac{1}{2}(\psi^* N + \psi N^*), \tag{2.6}$$

where the asterisk indicates complex conjugate. In place of the customary vectors **n**, **b**, and **t**, it will be found convenient to describe the curve in terms of the three linearly independent vectors N, N*, and **t**. They are readily shown to satisfy the relations

$$\mathrm{N} \cdot \mathrm{N}^* = 2, \quad \mathrm{N} \cdot \mathbf{t} = \mathrm{N}^* \cdot \mathbf{t} = \mathrm{N} \cdot \mathrm{N} = 0. \tag{2.7}$$

Following the procedure employed by Hasimoto, the temporal variation of N, N*, and **t** may also be expressed as linear combinations of these vectors, i.e.,

$$\mathrm{N}_t = \alpha \mathrm{N} + \beta \mathrm{N}^* + \gamma \mathbf{t}, \tag{2.8a}$$

$$\mathbf{t}_t = \lambda \mathrm{N} + \mu \mathrm{N}^* + \nu \mathbf{t}. \tag{2.8b}$$

Restrictions on the coefficients in these equations are readily obtained. Multiplication of Eqs. (2.8) by N and **t** and use of Eqs. (2.7) yields $\alpha + \alpha^* = 0$, $\beta = \nu = 0$, and $\gamma = -2\mu$. Hence

$$\mathrm{N}_t = iR\mathrm{N} + \gamma \mathbf{t}, \tag{2.9a}$$

$$\mathbf{t}_t = -\tfrac{1}{2}(\gamma^* N + \gamma N^*), \tag{2.9b}$$

where $R(s, t)$ is real. Equating N_{st} and N_{ts} as obtained from Eqs. (2.3) and (2.9a) as well as $\mathbf{t}_{st}$ and $\mathbf{t}_{ts}$ from Eqs. (2.6) and (2.9b), one finds

$$\psi_t + \gamma_s + i(\tau_0 \gamma - R\psi) = 0, \tag{2.10a}$$

$$R_s = \tfrac{1}{2} i(\gamma \psi^* - \gamma^* \psi). \tag{2.10b}$$

Equations (2.10) provide three equations for the five functions included in R, ψ, and γ. (The identifications $\psi = 2q$, $\gamma = -2B$, $R = -2iA$, $\gamma^* = 2C$, and $\tau_0 = 2\xi$ yield the equations employed in Ref. 7.) This indeterminacy may be used to specialize the functions so as to yield various types of space curves.[24] In particular, if the auxiliary functions R and γ can be expressed in terms of ψ and its spatial derivatives, then Eq. (2.10a) will provide an evolution equation for the spatial and temporal variation of the curvature and torsion of the curve as expressed through ψ. Also, as shown below in Sec. IV, Eqs. (2.9) and the Serret–Frenet equations are easily related to the linear equations that have been associated with these evolution equations for solution by inverse scattering methods.

III. CONNECTION WITH CERTAIN STANDARD EVOLUTION EQUATIONS

A. The sine-Gordon equation

It will now be shown that the sine–Gordon equation may be associated with certain curves of either constant curvature or constant torsion. If the curvature is a specified time independent function of the arc length parameter, then

$$\psi = \kappa(s) \exp(i\sigma), \tag{3.1}$$

where $\kappa(s)$ is assumed known. In addition, if τ_0 is chosen to be zero, then

$$\sigma(s, t) = \int_{-\infty}^{s} ds' \tau(s', t). \tag{3.2}$$

Also, if γ is allowed to be an arbitrary real function of time, then Eq. (2.10a) yields

$$R = \sigma_t \tag{3.3}$$

while Eq. (2.10b) becomes

$$R_s = \gamma(t) \kappa(s) \sin\sigma. \tag{3.4}$$

Introducing the new independent variables $ds' = \kappa(s)\, ds$ and $dt' = \gamma(t)\, dt$, one sees that σ satisfies the sine–Gordon equation in the form

$$\sigma_{s't'} = \sin\sigma. \tag{3.5}$$

More simply, one may merely set $\kappa = \kappa_0$, a constant, and $\gamma = 1/\kappa_0$.

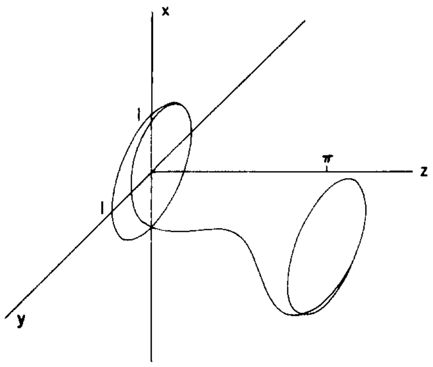


FIG. 1. Example of single soliton curve of constant curvature $\kappa=1$, $\tau=2\,\mathrm{sech}s$.

As an example of the type of space curves that may be associated with the sine–Gordon equation, consider the solution of Eq. (3.5) that yields the single soliton solution, namely

$$\sigma=4\tan^{-1}\exp(as+t/a), \qquad (3.6)$$

where a is an arbitrary constant. According to Eq. (3.2), the corresponding torsion is

$$\tau=2a\,\mathrm{sech}(as+t/a). \qquad (3.7)$$

The simplest space curve will be the one associated with a constant value of κ. Both curvature and torsion are thus specified. Determination of a coordinate representation for a curve having specified torsion and curvature is a standard problem in elementary differential geometry.[22] A calculation for the single soliton curves considered here is given in Appendix A. The result for $\kappa_0=1$ and $a=1$ is shown in Fig. 1.

A curve of constant torsion may also be associated with the sine–Gordon equation. Setting $\tau=\tau_0$, one sees from Eq. (2.5) that ψ is real and equal to the curvature. If one sets $\kappa=\sigma_s$ and chooses $\gamma=(i/\tau_0)\sin\sigma$, the expression for R_s is integrable and Eq. (2.10a) again yields the sine–Gordon equation. A coordinate representation for a curve with constant torsion and with curvature $\kappa(s,t)=2a\,\mathrm{sech}(as+t/a)$ is also derived in Appendix A. The result for $\tau_0=1$ and $a=1$ is shown in Fig. 2. Figures of this type for various values of a are given by Hasimoto in his consideration[19] of helical curves associated with the single soliton solution of the nonlinear Schrödinger equation.

B. The Hirota equation

Certain other well-known evolution equations are obtainable by specifying γ in such a way that the expression for R_s given in Eq. (2.10b) is a perfect derivative. An obvious choice of some generality is

$$\gamma=f\psi+ik\psi_s+a\psi_{ss}, \qquad (3.8)$$

where f is a real function that will be determined subsequently while k and a are real constants. Then,

$$R=-\tfrac{1}{2}k|\psi|^2+\tfrac{1}{2}ia(\psi^*\psi_s-\psi\psi_s^*)+\Gamma(t), \qquad (3.9)$$

where $\Gamma(t)$ arises upon integration of Eq. (2.10b) with respect to s. Substitution of Eqs. (3.8) and (3.9) into Eq. (2.10a) shows that the associated evolution equation must be of the form

$$\psi_t+(f\psi)_s+i(\tau_0f-\Gamma)\psi+i(k+\tau_0a)\psi_{ss}+a\psi_{sss}-\tau_0k\psi_s$$
$$+\tfrac{1}{2}ik|\psi|^2\psi+\tfrac{1}{2}a|\psi|^2\psi_s-\tfrac{1}{2}\psi^2\psi_s^*=0. \qquad (3.10)$$

To obtain an equation involving only ψ or $|\psi|$ and derivatives of ψ, the term $\psi^2\psi_s^*$ in this equation must be eliminated. This is readily accomplished by setting $f=f(\psi,\psi^*)$ and requiring

$$\psi\frac{\partial f}{\partial\psi^*}=\tfrac{1}{2}a\psi^2. \qquad (3.11)$$

Then, since f is real

$$f=\tfrac{1}{2}a|\psi|^2+c, \qquad (3.12)$$

where c is a real constant. The equation for ψ now becomes

$$\psi_t+\tfrac{3}{2}a|\psi|^2\psi_s+p\psi_s+iq|\psi|^2\psi+ir\psi+2iq\psi_{ss}+a\psi_{sss}=0, \qquad (3.13)$$

where

$$p=c-\tau_0k,\quad q=\tfrac{1}{2}(k+\tau_0a),\quad r=c\tau_0-\Gamma. \qquad (3.14)$$

The constants p and r may be set equal to zero without loss of generality since terms involving only ψ or ψ_s are readily removed from Eq. (3.13) by simple changes of the dependent and independent variables. (Such transformations are, of course, of prime concern in determining the geometric significance of the dependent variable ψ.) Solving for $c, k,$ and Γ in terms of q and a, one finds $c=2\tau_0q-\tau_0^2a$, $k=2q-\tau_0a$, $\Gamma=2\tau_0^2q-\tau_0^3a$. The final form of the evolution equation is then

$$\psi_t+\tfrac{3}{2}a|\psi|^2\psi_s+iq|\psi|^2\psi+2iq\psi_{ss}+a\psi_{sss}=0 \qquad (3.15)$$

with

$$R=-(q-\tfrac{1}{2}\tau_0a)|\psi|^2+\tfrac{1}{2}ia(\psi^*\psi_s-\psi\psi_s^*)+\tau_0^2(2q-\tau_0a) \qquad (3.16)$$

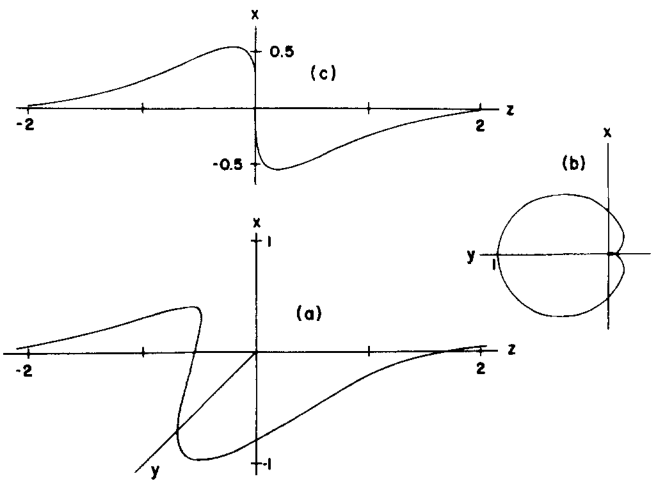


FIG. 2 (a) Example of single soliton curve of constant torsion, $\tau=1$, $\kappa=2\,\mathrm{sech}s$; (b) projection of curve on xy plane; (c) projection of curve on xz plane.

and

$$\gamma = f\psi + i(2q - \tau_0 a)\psi_s + a\psi_{ss}, \qquad (3.17)$$

in which

$$f = \tfrac{1}{2}a|\psi|^2 + \tau_0(2q - \tau_0 a). \qquad (3.18)$$

Equation (3.15) may be put into the three-parameter form given by Hirota[20] by rescaling the arc length variable. Setting $s = \lambda x$, one obtains

$$\psi_t + 3A|\psi|^2\psi_x + iB|\psi|^2\psi + iC\psi_{xx} + D\psi_{xxx} = 0 \qquad (3.19)$$

where $A = c/2\lambda$, $B = q$, $C = 2q/\lambda^2$, $D = a/\lambda^3$. The elimination of λ between these four coefficients yields the restriction $AC = BD$.

As is well known[8,20] Eqs. (3.15) or (3.19) contain both the nonlinear Schrödinger equation and the modified Korteweg—de Vries equation. The results for these specializations are summarized below.

(1) *Nonlinear Schrödinger equation*: Setting $a = 0$ and $q = -1$ in Eq. (3.15)—(3.17), one obtains

$$i\psi_t + |\psi|^2\psi + 2\psi_{ss} = 0 \qquad (3.20)$$

and

$$R = |\psi|^2 - 2\tau_0^2, \qquad (3.21a)$$

$$\gamma = -2\tau_0\psi - 2i\psi_s. \qquad (3.21b)$$

As noted above, the helical curves corresponding to the single soliton solution of the nonlinear Schrödinger equation has been described by Hasimoto.[19]

(2) *Modified Korteweg—de Vries equation*: Setting $q = 0$ and $a = 1$, one obtains

$$\psi_t + \tfrac{3}{2}|\psi|^2\psi_s + \psi_{sss} = 0 \qquad (3.22)$$

and

$$R = \tfrac{1}{2}\tau_0|\psi|^2 - \tau_0^3 + \tfrac{1}{2}i(\psi^*\psi_s - \psi\psi_s^*), \qquad (3.23a)$$

$$\gamma = \psi(\tfrac{1}{2}|\psi|^2 - \tau_0^2) - i\tau_0\psi_s + \psi_{ss}. \qquad (3.23b)$$

If ψ is assumed to be real, which according to the definition of ψ given in Eq. (2.5), corresponds to a curve having a constant torsion equal to τ_0, one obtains results which pertain to the modified Korteweg—de Vries equation.

Finally, in addition to the differential equations that follow from the integrable form of R that results from Eq. (3.8), certain integral equations associated with nonintegrable R have also been considered.[8,14] A simple example is obtained here with the choice $\gamma = i\psi$ and $\tau_0 = 0$. Then Eq. (2.10a) becomes

$$\psi_t + i\psi_s + i\psi\int_{-\infty}^{s} ds'|\psi|^2 = 0. \qquad (3.24)$$

The soliton nature of such integral equations has been noted previously.[8,14]

IV. CONNECTION WITH LINEAR EQUATIONS OF INVERSE SCATTERING

Since the Serret—Frenet equations, as well as Eqs. (2.9) that express the time dependence of N and t, are sets of linear equations, it is natural to inquire as to the manner in which they may be related to the systems of linear equations that have previously been identified with the soliton equations. A correspondence between these linear equations is readily established. One may begin by noting that any of the three scalar components of Eqs. (2.3) and (2.6) as well as Eqs. (2.9) possess the first integral $|N|^2 + t^2 = 1$, where N is now some one of the three components of N and similarly for t. Setting $N = u + iv$ and following a standard procedure used in curve theory,[22] one may factor this first integral and write

$$\frac{u+it}{1-v} = \frac{1+v}{u-it} \equiv \varphi, \qquad (4.1a)$$

$$(u-it)/(1-v) = (1+v)/(u+it) \equiv -1/\chi = \varphi^*. \qquad (4.1b)$$

Solving for u, v, and t in terms of φ and χ, one finds

$$u = (1-\varphi\chi)/(\varphi - \chi), \qquad (4.2a)$$

$$v = (\varphi + \chi)/(\varphi - \chi), \qquad (4.2b)$$

$$t = i(1+\varphi\chi)/(\varphi - \chi). \qquad (4.2c)$$

These expressions are now introduced into the scalar equations for u, v, and t that are obtainable from Eqs. (2.9), and from Eqs. (2.3) and (2.6). Writing $\gamma = \gamma_r + i\gamma_i$ and $\psi = \psi_r + i\psi_i$, one obtains

$$\dot{u} = -Rv + \gamma_r t, \qquad (4.3a)$$

$$\dot{v} = Ru + \gamma_i t, \qquad (4.3b)$$

$$\dot{t} = -(\gamma_r u + \gamma_i v) \qquad (4.3c)$$

and

$$u_s = \tau_0 v - \psi_r t, \qquad (4.4a)$$

$$v_s = -\tau_0 u - \psi_i t, \qquad (4.4b)$$

$$t_s = \psi_r u + \psi_i v. \qquad (4.4c)$$

(The dot indicates differentiation with respect to time.) From Eq. (4.3) one finds that ψ satisfies the Riccati equation

$$\varphi_t + i\gamma_r\varphi + \tfrac{1}{2}(i\gamma_i - R)\varphi^2 - \tfrac{1}{2}(i\gamma_i + R) = 0. \qquad (4.5)$$

The function χ is found to satisfy the same equation. Similarly, substitution into Eqs. (4.4) leads to the Riccati equation

$$\varphi_s - i\psi_r\varphi + \tfrac{1}{2}(\tau_0 - i\psi_i)\varphi^2 + \tfrac{1}{2}(i\psi_i + \tau_0) = 0. \qquad (4.6)$$

Each of these Riccati equations may be replaced by a pair of linear first-order equations by setting $\varphi = v_2/v_1$. One finds

$$v_{1t} = \tfrac{1}{2}i\gamma_r v_1 + \tfrac{1}{2}(i\gamma_i - R)v_2, \qquad (4.7a)$$

$$v_{2t} = \tfrac{1}{2}(i\gamma_i + R)v_1 - \tfrac{1}{2}i\gamma_r v_2 \qquad (4.7b)$$

and

$$v_{1s} = -\tfrac{1}{2}i\psi_r v_1 + \tfrac{1}{2}(-i\psi_i + \tau_0)v_2, \qquad (4.8a)$$

$$v_{2s} = -\tfrac{1}{2}(i\psi_i + \tau_0)v_1 + \tfrac{1}{2}i\psi_r v_2. \qquad (4.8b)$$

Besides the function ψ, the coefficients in these linear equations are the auxiliary functions R and γ that have already been obtained in the previous specializations to obtain the various evolution equations. Hence, once R and γ have been determined, the linear equations that are customarily associated with these evolution equations are immediately available. Not all of the familiar forms for these linear equations follow merely from Eqs. (4.7) and (4.8), however, since additional linear equations are also obtainable if one replaces Eq. (4.1) by alternative factorizations and repeats the calculation outlined above. In particular, the choice

$$(u+iv)/(1-t)=(1+t)/(u-iv)\equiv\varphi, \qquad (4.9\text{a})$$

$$(u-iv)/(1-t)=(1+t)/(u+iv)\equiv-1/\chi=\varphi^* \qquad (4.9\text{b})$$

leads to the Riccati equations

$$\varphi_t-iR\varphi+\tfrac{1}{2}\gamma^*\varphi^2+\tfrac{1}{2}\gamma=0, \qquad (4.10\text{a})$$

$$\varphi_s+i\tau_0\varphi-\tfrac{1}{2}\psi^*\varphi^2-\tfrac{1}{2}\psi=0 \qquad (4.10\text{b})$$

and the associated linear equations

$$v_{1t}=-\tfrac{1}{2}iRv_1+\tfrac{1}{2}\gamma^*v_2, \qquad (4.11\text{a})$$

$$v_{2t}=-\tfrac{1}{2}\gamma v_1+\tfrac{1}{2}iRv_2 \qquad (4.11\text{b})$$

and

$$v_{1s}-\tfrac{1}{2}i\tau_0v_1=-\tfrac{1}{2}\psi^*v_2, \qquad (4.12\text{a})$$

$$v_{2s}+\tfrac{1}{2}i\tau_0v_2=\tfrac{1}{2}\psi v_1 \qquad (4.12\text{b})$$

where again $\varphi=v_2/v_1$.

Finally a third factorization

$$(v+it)/(1-u)=(1+u)/(v-it)\equiv\varphi, \qquad (4.13\text{a})$$

$$(v-it)/(1-u)=(1+u)/(v+it)\equiv-1/\chi=\varphi^* \qquad (4.13\text{b})$$

leads to the Riccati equations

$$\varphi_t+i\gamma_i\varphi+\tfrac{1}{2}(R+i\gamma_r)\varphi^2+\tfrac{1}{2}(R-i\gamma_r)=0 \qquad (4.14\text{a})$$

and

$$\varphi_s-i\psi_i\varphi-\tfrac{1}{2}(\tau_0+i\psi_r)\varphi^2-\tfrac{1}{2}(\tau_0-i\psi_r)=0 \qquad (4.14\text{b})$$

with associated linear equations

$$v_{1t}=\tfrac{1}{2}i\gamma_iv_1+\tfrac{1}{2}(i\gamma_r+R)v_2, \qquad (4.15\text{a})$$

$$v_{2t}=\tfrac{1}{2}(i\gamma_r-R)v_1-\tfrac{1}{2}i\gamma_iv_2 \qquad (4.15\text{b})$$

and

$$v_{1s}=-\tfrac{1}{2}i\psi_iv_1-\tfrac{1}{2}(\tau_0+i\psi_r)v_2, \qquad (4.16\text{a})$$

$$v_{2s}=\tfrac{1}{2}(\tau_0-i\psi_r)v_1+\tfrac{1}{2}i\psi_iv_2. \qquad (4.16\text{b})$$

Comparison with Eqs. (4.7) and (4.8) shows that they merely serve to interchange the roles of the real and imaginary parts of ψ and γ.

Use of the first two of these sets of equations to recover the standard linear equations associated with the evolution equations is outlined in the next section.

V. SUMMARY OF RESULTS FOR VARIOUS EVOLUTION EQUATIONS

Results obtained in the previous sections may be used in various ways to associate linear equations with the nonlinear evolution equations under consideration here.

A. The sine-Gordon equation

As noted in Sec. III, the sine—Gordon equation may be associated with curves of constant curvature ($=\kappa_0$). If one also sets $\tau_0=0$ and $\gamma=1/\kappa_0$ in Eq. (3.4) then, since now $R=\sigma_t$, Eq. (3.4) becomes the sine—Gordon equation. The linear equations of the first factorization, namely Eqs. (4.7) and (4.8), yield

$$v_{1t}-(i/2\kappa_0)v_1=-\tfrac{1}{2}\sigma_tv_2, \qquad (5.1\text{a})$$

$$v_{2t}+(i/2\kappa_0)v_2=\tfrac{1}{2}\sigma_tv_1 \qquad (5.1\text{b})$$

along with

$$v_{1s}=-\tfrac{1}{2}i\kappa_0v_1\cos\sigma-\tfrac{1}{2}i\kappa_0v_2\sin\sigma, \qquad (5.2\text{a})$$

$$v_{2s}=-\tfrac{1}{2}i\kappa_0v_1\sin\sigma+\tfrac{1}{2}i\kappa_0v_2\cos\sigma. \qquad (5.2\text{b})$$

Equations of this same form but with the role of spatial and temporal derivatives interchanged may be obtained by considering a curve of constant torsion τ_0. As noted in Sec. IIIA, the quantity ψ is then real. With ψ equal to σ_s as before and also with $\gamma=(i/\tau_0)\sin\sigma$, $R=(1/\tau_0)\cos\sigma$ the linear equations corresponding to the second factorization, i.e., Eqs. (4.11) and (4.12), take the form

$$v_{1s}-\tfrac{1}{2}i\tau_0v_1=-\tfrac{1}{2}\sigma_sv_2, \qquad (5.3\text{a})$$

$$v_{2s}+\tfrac{1}{2}i\tau_0v_2=\tfrac{1}{2}\sigma_sv_1 \qquad (5.3\text{b})$$

and

$$v_{1t}=-(i/2\tau_0)v_1\cos\sigma-(i/2\tau_0)v_2\sin\sigma, \qquad (5.4\text{a})$$

$$v_{2t}=-(i/2\tau_0)v_1\sin\sigma+(i/2\tau_0)v_2\cos\sigma. \qquad (5.4\text{b})$$

These same choices for γ and R also yield linear equations for the one-component inverse method. Use of the first factorization, i.e., Eqs. (4.8a), (4.8b), yields

$$v_{1s}=-\tfrac{1}{2}i\sigma_sv_1+\tfrac{1}{2}\tau_0v_2, \qquad (5.5\text{a})$$

$$v_{2s}=-\tfrac{1}{2}\tau_0v_1+\tfrac{1}{2}i\sigma_sv_2. \qquad (5.5\text{b})$$

The second-order equation obtained for v_1 is

$$v_{1ss}+\tfrac{1}{4}(\tau_0^2+\sigma_s^2+2i\sigma_{ss})v_1=0, \qquad (5.6)$$

to which the customary inverse methods for the Schrödinger equation[2,21] may be applied.

B. The Hirota equation and specializations thereof

Here the most convenient choice for linear equations appears to be Eqs. (4.11) and (4.12) with R and γ given by Eqs. (3.16)—(3.18). As noted in Sec. III, the specialization $a=0$ in the Hirota equation leads to a nonlinear Schrödinger equation. The most convenient linear equations are again Eqs. (4.11) and (4.12) and the standard results[8] are obtained.

On setting $q=0$ and requiring ψ to be real, i.e., by setting $\tau=\tau_0$, the results for the Hirota equation reduce to those for the modified Korteweg—de Vries equation. Here the choice of Eqs. (4.11) and (4.12) for the linear equations leads to the customary first-order linear equation.[7] On the other hand, if Eqs. (4.7) and (4.8) are employed then Eqs. (4.8) may be combined to yield

$$v_{1ss}+(\tfrac{1}{4}\tau_0^2+z)v_1=0, \qquad (5.7\text{a})$$

$$v_{2ss}+(\tfrac{1}{4}\tau_0^2+z^*)v_2=0, \qquad (5.7\text{b})$$

where

$$z=\tfrac{1}{4}(\psi^2+2i\psi_s). \qquad (5.8)$$

This, of course, is the Miura transformation,[23] and one can readily show that z satisfies the Korteweg–de Vries equation in the form $z_t + 6zz_s + z_{sss} = 0$ if ψ satisfies the modified Korteweg–de Vries equation as given in Eq. (3.22). When γ, R, and ψ are expressed in terms of z, one finds that Eq. (4.7a) yields

$$v_{1t} = -\left[4i\frac{\partial^3}{\partial s^3} + 3i\left(z\frac{\partial}{\partial s} + \frac{\partial}{\partial s}z\right)\right]v_1, \qquad (5.9)$$

i.e., the expected result[4] for the time dependence of the eigenfunctions associated with the Korteweg–de Vries equation.

VI. KINEMATICS OF THE CURVES

In the hydrodynamic problem considered by Hasimoto, the motion of each point of the vortex filament was known to be parallel to the binormal to the curve at that point. The magnitude of the filament velocity, in appropriate units, was equal to the curvature of the filament at each point. If $\mathbf{X}(s,t)$ represents a position vector to the filament, then

$$\mathbf{X}_t(s,t) = \kappa(s,t)\mathbf{b}. \qquad (6.1)$$

This provided the starting point for the physical considerations which ultimately led to the nonlinear Schrödinger equation.

In the present analysis, on the other hand, analytical specializations have been introduced to yield the standard evolution equations. Consideration of these results would not be complete without some indication of how the velocity of the curve may be obtained.

Writing the velocity vector in the general linear form

$$\mathbf{X}_t = h^*\mathbf{N} + h\mathbf{N}^* + g\mathbf{t} \qquad (6.2)$$

and noting that $\mathbf{X}_s = \mathbf{t}$, one finds that the equality of mixed second derivatives of $\mathbf{X}$ imposes the relations

$$h_s + i\tau_0 h + \tfrac{1}{2}\psi g = -\tfrac{1}{2}\gamma, \qquad (6.3a)$$

$$h_s^* - i\tau_0 h^* + \tfrac{1}{2}\psi^* g = -\tfrac{1}{2}\gamma^*, \qquad (6.3b)$$

$$g_s - \psi^* h - \psi h^* = 0. \qquad (6.3c)$$

The solution of this system of differential equations is conveniently expressed in terms of the functions v_1 and v_2 introduced with the second factorization in Sec. IV. The calculation is summarized in Appendix B.

For the nonlinear Schrödinger equation in the form employed here, i.e., Eq. (3.20), one obtains

$$\mathbf{X}_t = 2\kappa\mathbf{b} + 4\tau_0\mathbf{t}. \qquad (6.4)$$

The constant τ_0 is seen to introduce a "slippage" along the curve.

For the modified Korteweg–de Vries equation the result is

$$\mathbf{X}_t = -\kappa_s\mathbf{n} + 2\tau_0\kappa\mathbf{b} + (3\tau_0^2 - \tfrac{1}{2}\kappa^2)\mathbf{t}. \qquad (6.5)$$

The results for the sine–Gordon equation cannot be expressed in such a transparent form since the integrals involved are not amenable to evaluation in any direct way. Specific results for the steady state solution obtained by setting $\sigma(s,t) = \sigma(as + t/a)$ are obtainable as indicated in Appendix B. For the curve of constant torsion one finds

$$\mathbf{X}_t = -a^{-2}[(\kappa/\tau_0)\mathbf{b} + \mathbf{t}]. \qquad (6.6)$$

ACKNOWLEDGMENT

The author wishes to thank P.G. Saffman for drawing attention to the research by Hasimoto concerning solitons on vortex filaments.

APPENDIX A

If the curvature κ and torsion τ of a twisted curve are given as functions of arc length s along the curve, then the curve may be referred to a rectangular coordinate system through relations of the form

$$x = \int^s U\,ds', \quad y = \int^s V\,ds', \quad z = \int^s W\,ds', \qquad (A1)$$

where the integrands are related to κ and τ through four intermediate functions. These latter four functions comprise the general solution of the Riccati equation

$$f_s + i\kappa f - \tfrac{1}{2}i\tau f^2 + \tfrac{1}{2}i\tau = 0 \qquad (A2)$$

i.e.,

$$f = (cP + Q)/(cR + S), \qquad (A3)$$

where c is an arbitrary constant. The expressions for U, V, and W are[22]

$$U = [P^2 - R^2 - (Q^2 - S^2)]/2T, \qquad (A4a)$$

$$V = i[P^2 - R^2 + (Q^2 - S^2)]/2T, \qquad (A4b)$$

$$W = (RS - PQ)/T, \qquad (A4c)$$

in which

$$T = PS - QR. \qquad (A5)$$

The single soliton curves considered in Sec. IIIa may be obtained by solving Eq. (A2) for $\kappa = \kappa_0$ and $\tau = 2a \times \operatorname{sech}(as + bt)$ and for $\kappa = 2a\operatorname{sech}(as + bt)$ and $\tau = \tau_0$. The general solutions of the two associated Riccati equations are closely related since a function g related to f in Eq. (A2) by the transformation

$$g = (f + 1)/(f - 1) \qquad (A6)$$

satisfies a Riccati equation with κ and τ interchanged, i.e.,

$$g_s + i\tau g - \tfrac{1}{2}i\kappa g^2 + \tfrac{1}{2}i\kappa = 0. \qquad (A7)$$

A particular solution of Eq. (A2) with $\kappa = 2a\operatorname{sech}(as + bt)$ and $\tau = \tau_0$ has been given previously for a problem in coherent optical pulse propagation.[25] In the notation being employed here, that particular solution is

$$f = -[a - i\exp(-i\sigma/2)]/[a - i\exp(i\sigma/2)] \qquad (A8)$$

where $\sigma = 4\tan^{-1}[\exp(as + bt)]$. One may readily show that if f_1 is a particular solution of Eq. (A2), then another particular solution is $f_2 = -1/f_1^*$. With two particular solutions known, the general solution is then given by a single quadrature.[22] The solution is of the form of Eq. (A3) with

$$P = \exp(-i\nu\xi)[\nu + i\exp(-i\sigma/2)] \qquad (A9a)$$

$$Q = \nu - i\exp(-i\sigma/2) \qquad (A9b)$$

$$R = \exp(-i\nu\xi)[\nu + i\exp(i\sigma/2)] \tag{A9c}$$

$$S = -[\nu - i\exp(i\sigma/2)] \tag{A9d}$$

where $\xi = as + bt$ and $\nu = \tau_0/a$. When the expressions for U, V and W are formed from these results, one finds that the integrals in Eqs. (A1) may be expressed in terms of elementary functions. The results, which have been discussed in some detail by Hasimoto,[19] are

$$x = -2\alpha\,\mathrm{sech}\,\xi\,\sin\nu\xi, \tag{A10a}$$

$$y = 2\alpha\,\mathrm{sech}\,\xi\,\cos\nu\xi, \tag{A10b}$$

$$z = s - 2\alpha\tanh\xi, \tag{A10c}$$

where $\alpha = 2/[a(\nu^2+1)]$.

An example of the type of helical curve described by this result is shown in Fig. 2. For additional figures, the paper by Hasimoto may be consulted. Equations (A10) may also be found in a paper by Hoppe.[26]

For $\kappa = \kappa_0$ and $\tau = 2a\,\mathrm{sech}(as + t/a)$, the transformation given in Eq. (A6) leads to a solution of Eq. (A3) in the form

$$f = (c\bar{P} + \bar{Q})/(c\bar{R} + \bar{S}) \tag{A11}$$

where

$$\bar{P} = 2\exp(-i\nu\xi)[\nu + i\cos(\sigma/2)], \tag{A12a}$$

$$\bar{Q} = -2\sin(\sigma/2), \tag{A12b}$$

$$\bar{R} = 2\exp(-i\nu\xi)\sin(\sigma/2), \tag{A12c}$$

$$\bar{S} = 2[\nu - i\cos(\sigma/2)]. \tag{A12d}$$

Equations (A4) now yield

$$\bar{U} = [(\nu^2-1)\cos\nu\xi + 2\cos(\sigma/2)\sin\nu\xi]/(\nu^2+1), \tag{A13a}$$

$$\bar{V} = [(\nu^2-1)\sin\nu\xi - 2\cos(\sigma/2)\cos\nu\xi]/(\nu^2+1), \tag{A13b}$$

$$\bar{W} = 2\nu\sin(\sigma/2)/(\nu^2+1). \tag{A13c}$$

Figure 1 is typical of the results that may be obtained in this case. Integrations for x and y were performed numerically.

APPENDIX B

Equation (6.3) may be written in the vector form

$$V_s + AV = F \tag{B1}$$

where $V = (h, h^*, g)^T$ and $F = -\frac{1}{2}(\gamma, \gamma^*, 0)^T$, in which T represents the transpose operation and

$$A = \begin{pmatrix} i\tau_0 & 0 & \frac{1}{2}\psi \\ 0 & -i\tau_0 & \frac{1}{2}\psi^* \\ -\psi^* & -\psi & 0 \end{pmatrix}. \tag{B2}$$

The solution of Eq. (B1) is conveniently expressed in terms of a fundamental matrix[27] that is composed of three column vectors which satisfy the homogeneous counterpart of Eq. (B1). The solution of Eq. (B1) is then

$$V = \Phi\left[C + \int_{-\infty}^{s} ds'\,\Phi^{-1}(s')\,F(s')\right], \tag{B3}$$

where Φ is the fundamental matrix satisfying $\Phi_s + A\Phi = 0$ and C is a constant vector. The solution vectors required for Φ are readily constructed from Eqs. (4.12) since these equations may be combined in three different ways to yield equations of the form

$$\phi_{n1s} + i\tau_0\phi_{n1} + \tfrac{1}{2}\psi\phi_{n3} = 0, \tag{B4a}$$

$$\phi_{n2s} - i\tau_0\phi_{n2} + \tfrac{1}{2}\psi^*\phi_{n3} = 0, \tag{B4b}$$

$$\phi_{n3s} - \psi^*\phi_{n1} - \psi\phi_{n2} = 0, \tag{B4c}$$

where $n = 1, 2, 3$. Three choices are

$$(\phi_{11}, \phi_{12}, \phi_{13}) = (v_1^{*2}, -v_2^{*2}, 2v_1^*v_2^*) \tag{B5a}$$

$$(\phi_{21}, \phi_{22}, \phi_{23}) = (-v_2^2, v_1^2, 2v_1v_2) \tag{B5b}$$

and

$$(\phi_{31}, \phi_{32}, \phi_{33}) = (-v_1^*v_2, -v_1v_2^*, |v_1|^2 - |v_2|^2). \tag{B5c}$$

Choosing the normalizations of v_1 and v_2 such that $|v_1|^2 + |v_2|^2 = 1$, one readily finds that $\det|\phi_{ij}| = 1$ and that the inverse of Φ is

$$\Phi^{-1} = \begin{pmatrix} v_1^2 & -v_2^2 & v_1v_2 \\ -v_2^{*2} & v_1^{*2} & v_1^*v_2^* \\ -2v_1v_2^* & -2v_1^*v_2 & |v_1|^2 - |v_2|^2 \end{pmatrix}. \tag{B6}$$

The vector V is then given by Eq. (B3) with

$$\int_{-\infty}^{s} ds'\,\Phi^{-1}F = -\tfrac{1}{2}\int_{-\infty}^{s} ds' \begin{pmatrix} \gamma v_1^2 - \gamma^* v_2^2 \\ \gamma^* v_1{}^{*2} - \gamma v_2^{*2} \\ -2\gamma v_1v_2^* - 2\gamma^* v_1^* v_2 \end{pmatrix}. \tag{B7}$$

A similar solution to equations of the form of Eq. (B1) was employed in Ref. 8.

The integrals in Eq. (B6) are readily evaluated when γ is introduced in the form given in Eq. (3.8). The integrals in Eqs. (B6) are then readily carried out with the help of Eqs. (4.12).

For the nonlinear Schrödinger equation γ is given by Eq. (3.21b) and one finds

$$\int_{-\infty}^{s} ds'\,\Phi^{-1}F = \begin{pmatrix} i\psi - \frac{1}{2}Gv_1^*v_2 \\ -i\psi^* - \frac{1}{2}Gv_1v_2^* \\ 4\tau_0 - \frac{1}{2}G(|v_2|^2 - |v_1|^2) \end{pmatrix}, \tag{B8}$$

where

$$G = 8\tau_0(|v_2|^2 - |v_1|^2)\big|_{s=-\infty}. \tag{B9}$$

If one now chooses $c_1 = c_2 = 0$ and $c_3 = \frac{1}{2}G$, then $h = i\psi$ and $g = 4\tau_0$. Equation (6.2) reduces to Eq. (6.4).

For the modified Korteweg—de Vries equation, γ is given by Eq. (3.23b), and one finds

$$\int_{-\infty}^{s} ds'\,\Phi^{-1}F = \begin{pmatrix} i\tau_0\psi - \frac{1}{2}\psi_s - \frac{1}{2}Gv_1^*v_2 \\ -i\tau_0\psi^* - \frac{1}{2}\psi_s^* - \frac{1}{2}Gv_1v_2^* \\ -\frac{1}{2}|\psi|^2 + 3\tau_0^2 - \frac{1}{2}G(|v_2|^2 - |v_1|^2) \end{pmatrix} \tag{B10}$$

where

$$G = 6\tau_0^2(|v_2|^2 - |v_1|^2)\big|_{s=-\infty}. \tag{B11}$$

If one chooses $c_1 = c_2 = 0$ and $c_3 = \frac{1}{2}G$, then $h = i\tau_0\psi - \frac{1}{2}\psi_s$ and $g = -\frac{1}{2}|\psi|^2 + 3\tau_0^2$. Equation (6.2) reduces to Eq. (6.5).

It does not appear that the corresponding analysis can be carried out for the sine—Gordon equation. However, for steady state motion of a curve of constant torsion, the calculation is feasible. Writing the steady state solution in the form $\sigma(as+t/a)$, one notes that $\gamma=(i/\tau_0)\times\sin\sigma=i/(\tau_0a^2)\sigma_{ss}$. Then with the help of Eqs. (5.3), one obtains

$$\int_{-\infty}^{s} ds'\Phi^{-1}F=\frac{i}{2\tau_0a^2}\begin{pmatrix}-\sigma_s+2i\tau_0Gv_1{}^*v_2\\ \sigma_s+2i\tau_0Gv_1v_2{}^*\\ 2i\tau_0+2i\tau_0G(|v_1|^2-|v_2|^2)\end{pmatrix}\tag{B12}$$

where

$$G=(|v_1|^2-|v_2|^2)|_{s=-\infty}.\tag{B13}$$

If one now chooses $c_1=c_2=0$ and $c_3=-G/a^2$, then $h=-i\sigma_s/2\tau_0a^2$ and $g=-1/a^2$. Noting that $\kappa=\sigma_s$, one finds that Eq. (6.2) reduces to Eq. (6.6).

[1] N. J. Zabusky, in *Nonlinear Partial Differential Equations*, edited by W. Ames (Academic, New York, 1967), pp. 223—58.

[2] C.S. Gardner, J.M. Greene, M.D. Kruskal, and R.M. Miura, Phys. Rev. Lett. **19**, 1095 (1967).

[3] R.M. Miura, C.S. Gardner, and M.D. Kruskal, J. Math. Phys. **9**, 1204 (1968).

[4] P.D. Lax, Commun. Pure Appl. Math. **21**, 467 (1968).

[5] V.E. Zakharov and L.D. Fadeev, Functs. Anal. Ego Pritozhen. **5**, 18 (1971) [Funct. Anal. Appl. **5**, 280 (1971)].

[6] V.E. Zakharov and A.B. Shabat, Zh. Eksp. Teor, Fiz. **61**, 118 (1971) [Sov. Phys. JETP **34**, 62 (1972)].

[7] M.J. Ablowitz, D.J. Kaup, A.C. Newell, and H. Segur, Phys. Rev. Lett. **31**, 125 (1973).

[8] M.J. Ablowitz, D.J. Kaup, A.C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).

[9] H.D. Wahlquist and F.B. Estabrook, Phys. Rev. Lett. **31**, 1386 (1973); J. Math. Phys. **16**, 1 (1975).

[10] G.L. Lamb, Jr., J. Math Phys. **15**, 2157 (1974).

[11] D.W. McLaughlin, J. Math. Phys. **16**, 96 (1975).

[12] R. Hirota, in *Bäcklund Transformations*, edited by R. Miura Vol. 515 of Lecture Notes in Mathematics, edited by A. Dold and B. Eckmann (Springer-Verlag, New York, 1974), pp. 40—68.

[13] H. Rund, Ref. 12, pp. 199—226.

[14] F. Calogero and A. Degasperis, Nuovo Cimento B **32**, 201 (1976).

[15] F. Lund and T. Regge, Phys. Rev. D **14**, 1524 (1976).

[16] S. Kumei, J. Math. Phys. **16**, 2461 (1975); **18**, 256 (1977).

[17] H.D. Wahlquist and F.B. Estabrook, J. Math. Phys. **17**, 1293 (1976).

[18] R. Hermann, Phys. Rev. Lett. **36**, 835 (1976).

[19] H. Hasimoto, J. Fluid Mech. **51**, 477 (1972).

[20] R. Hirota, J. Math. Phys. **14**, 805 (1973).

[21] G.L. Lamb, Jr., Phys. Rev. A **9**, 422 (1974).

[22] L.P. Eisenhart, *A Treatise on the Differential Geometry of Curves and Surfaces* (Dover, New York, 1960), Sec. 13-15.

[23] R.M. Miura, J. Math. Phys. **9**, 1202 (1968).

[24] G.L. Lamb, Jr., Phys. Rev. Lett. **37**, 235 (1976).

[25] G.L. Lamb, Jr., Rev. Mod. Phys. **43**, 99 (1971), Eq. (5.10).

[26] R. Hoppe, J. Math. **60**, 182 (1862).

[27] J.L. Goldberg and A.J. Schwartz, *Systems of Ordinary Differential Equations* (Harper and Row, New York, 1972), p. 123.

Spinorial structures and electromagnetic hyperheavens

J. D. Finley, III[a)]

University of New Mexico, Albuquerque, New Mexico 87131

J. F. Plebański[b)]

Centro de Investigación y de Estudios Avanzados del I. P. N., México 14, D. F., Mexico

(Received 24 January 1977)

Following the results of Garcia, Plebański, and Robinson, and the spinorial technique of Finley and Plebański, complex space-times with a null string (therefore, one-sidedly degenerate) are investigated in the presence of the cosmological constant and the electromagnetic field. The structure derived by Garcia, Plebański, and Robinson thus acquires a more lucid and concise presentation. The computational procedure is improved considerably, permitting the inclusion of an explicit outline of all calculations involved.

1. INTRODUCTION

This article is a further step in the series initiated by Plebański and Robinson,[1] where it was established that, in a complex space—time, the algebraic degeneration from (at least) one side of the conformal curvature—that is, e.g., the self-dual part of the conformal curvature tensor has a multiple Penrose's spinor—implies that the solutions to Einstein's equations in vacuum are described in terms of *one* function which must fulfill an equation of the second order and second degree (the hyperheavenly equation). Those authors also outlined the basic idea of a generalization to the presence of a (complex) electromagnetic field in the space. The results of Ref. 1 were reformulated by the present authors (without the electromagnetic field, but with an allowed cosmological constant) by using spinorial techniques.[2] (Hereafter we refer to this paper as I). Because of the intrinsic alignment of the spinorial parametrization with the very nature of the geometry under consideration, and its compactness, some rather nontrivial computations have been presented in detail.

Recently, Garcia, Plebański, and Robinson[3] have given a complete description of the structure of a complex one-sidedly degenerate electrovac space—time with cosmological constant (where the electromagnetic field is required to be correlated with the null, totally geodesic 2-surfaces which are characteristic of complex minimally degenerate space—times[4]), reducing all the Einstein—Maxwell equations to two differential conditions on two arbitrary functions. (This structure is therefore considerably more involved than the case of pure vacuum.) The objective of this present study is to give a spinorial derivation of the structure in this situation, and to give a complete calculation of the results, following the notation and techniques of I.

In Sec. 2 we list the assumptions, summarize the notation, and then describe the results by giving the representation of the (complex) electromagnetic and gravitational fields through their respective (scalar) potentials and the residual equations which they must satisfy. In Sec. 3 we outline the explicit derivation of these results, which were made possible because of the automatic convenient, sequential grouping of the equations according to their helicity (made possible by the choice of tetrad and the spinorial notation) into trivially integrable triplets or singlets.

2. ASSUMPTIONS AND THE RESULTS

We study the combined Einstein—Maxwell equations with cosmological constant, λ, analytically continued to complex space—time. The 2-form $F=\frac{1}{2}F_{\mu\nu}dx^\mu\wedge dx^\nu$ which describes the electromagnetic field may be written in terms of its spinor components[5]:

$$\omega\equiv\tfrac{1}{2}(F_{\mu\nu}+{}^*F_{\mu\nu})\,dx^\mu\wedge dx^\nu\equiv 2f_{AB}S^{AB},$$

the self-dual part,

$$\tilde{\omega}\equiv\tfrac{1}{2}(F_{\mu\nu}-{}^*F_{\mu\nu})\,dx^\mu\wedge dx^\nu\equiv 2f_{\dot{A}\dot{B}}S^{\dot{A}\dot{B}},$$

the anti-self-dual part.

$$(2.1)$$

The Einstein—Maxwell equations are then written as the set

$$d\omega=0=d\tilde{\omega}, \tag{2.2a}$$

$$C_{AB\dot{C}\dot{D}}=-8f_{AB}f_{\dot{C}\dot{D}},\quad R=-4\lambda, \tag{2.2b}$$

where the second line is just the usual Einstein equations with electromagnetic source, written in their equivalent spinor form.[6]

In order to solve these equations we require, as in the vacuum case, that the space—time admit a congruence of null strings, i.e., null surfaces described by the 2-form

$$\Sigma\equiv du\wedge dv, \tag{2.3}$$

with u and v independent variables constant on each surface. Moreover, in addition we assume that the electromagnetic field is aligned so that

$$F\,\lrcorner\,\Sigma\equiv{}^*(F\wedge{}^*\Sigma)=0. \tag{2.4}$$

It turns out that, in a complex space—time, these assumptions are sufficient to allow us to completely characterize all solutions of our set of 18 coupled partial differential equations in terms of *two* key functions, which are subject to a pair of residual constraint equations. We will now state how the gravitational and electromagnetic structures may be described in terms of

[a)]Work supported in part by the Fomento Educacional, A.C., México 5, D.F., Mexico, and by the Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México 14, D.F., Mexico.

[b)]On leave of absence from the University of Warsaw, Warsaw, Poland.

these key functions and some parameters (constant on each null surface in our congruence) which can be interpreted as "constants of integration."

Capitalizing on the existence of our congruence we use especially adapted coordinates and tetrads: In addition to u and v as coordinates, we choose a pair, x and y, of longitudinal variables which vary along the surfaces. Following I (which see for more motivation) we interpret our coordinates as a pair of spinors

$$q_{\dot{A}} \equiv \begin{pmatrix} u \\ v \end{pmatrix}, \quad p^{\dot{A}} \equiv \begin{pmatrix} -y \\ -x \end{pmatrix}, \tag{2.5}$$

and choose a (complex, null) tetrad and the dual basis for the tangent space as

$$e_{\dot{A}} \equiv \phi^{-2} dq_{\dot{A}} \equiv \begin{pmatrix} e^3 \\ e^1 \end{pmatrix}, \quad E^{\dot{A}} \equiv -dp^{\dot{A}} + Q^{\dot{A}\dot{B}} dq_{\dot{B}} \equiv \begin{pmatrix} e^4 \\ e^2 \end{pmatrix}, \tag{2.6}$$

$$\eth^{\dot{A}} \equiv \phi^2 \left(\frac{\partial}{\partial q_{\dot{A}}} + Q^{\dot{A}\dot{B}} \partial_{\dot{B}} \right) \equiv \begin{pmatrix} \partial_3 \\ \partial_1 \end{pmatrix}, \quad -\partial_{\dot{A}} \equiv -\frac{\partial}{\partial p^{\dot{A}}} = \begin{pmatrix} \partial_y \\ \partial_x \end{pmatrix}, \tag{2.7}$$

where the metric is to be taken as[7]

$$g = 2e^1 \underset{s}{\otimes} e^2 + 2e^3 \underset{s}{\otimes} e^4 = 2\phi^{-2}(-dq_{\dot{A}} \underset{s}{\otimes} dp^{\dot{A}} + Q^{\dot{A}\dot{B}} dq_{\dot{A}} \underset{s}{\otimes} dq_{\dot{B}}), \tag{2.8}$$

while ϕ and $Q^{\dot{A}\dot{B}}$ are sufficiently smooth functions of all the coordinates. When ϕ is linear in $p^{\dot{A}}$, as will be shown, it and the symmetric second rank spinor $Q^{\dot{A}\dot{B}}$ determine the character of the metric as a double Kerr–Schild metric.[8] Note that, unless stated to the contrary, all our spinorial objects (connections, electromagnetic field, Riemann tensor, etc.) transform in the usual way under the tetradial gauge group, according to the type of their indices.[9]

The Einstein—Maxwell equations and Eq. (2.4) then require that $\phi = J_{\dot{A}} p^{\dot{A}} + \kappa$, where $J_{\dot{A}}$ and κ may be chosen to be constant.[10] There are then two distinct cases: If $J_{\dot{A}} = 0$, we call this case I, and renormalize κ to 1; otherwise we call the solution case II.[11] In Sec. 3 we will discuss the limiting process between these two cases, but here we give explicit formulas for both cases. We then, further, find that the electromagnetic fields must have the form

$$\omega = d\{(\epsilon p^{\dot{A}} + \gamma^{\dot{A}})\, dq_{\dot{A}}\}, \quad \tilde{\omega} = +d\{(\partial_{\dot{A}} H)\, dq^{\dot{A}}\}, \tag{2.9a}$$

with the requirement that

$$\eth_{\dot{A}} \partial^{\dot{A}} H = 0, \qquad 2.9b)$$

while the structural functions for the metric are just given by

$$\text{II: } \phi = J_{\dot{A}} p^{\dot{A}} + \kappa, \quad Q^{\dot{A}\dot{B}} = -\phi^3 \partial^{(\dot{A}} \phi^{-2} \partial^{\dot{B})} W + \tfrac{1}{2}(\mu \phi^3 + \lambda/6) K^{\dot{A}} K^{\dot{B}} - 2\epsilon \phi^3 \partial^{(\dot{A}} G^{\dot{B})}, \tag{2.10}$$

$$\text{I: } \phi = 1, \quad Q^{\dot{A}\dot{B}} = -\partial^{\dot{A}} \partial^{\dot{B}} \Theta + \tfrac{2}{3}(F^{(\dot{A}} + \tfrac{1}{2}\lambda p^{(\dot{A}}) p^{\dot{B})} - 2\epsilon \partial^{(\dot{A}} G^{\dot{B})},$$

where ϵ, μ, $\gamma^{\dot{A}}$, $F^{\dot{A}}$ are functions of $q_{\dot{B}}$ only, and, in case II, $K^{\dot{A}}$ is orthogonal to $J_{\dot{A}}$,

$$\delta^{\dot{A}}{}_{\dot{B}} = \frac{1}{\tau}(K^{\dot{A}} J_{\dot{B}} - J^{\dot{A}} K_{\dot{B}}), \quad \tau \equiv K^{\dot{A}} J_{\dot{A}}, \tag{2.11}$$

while W or Θ (in the two cases) and H are key functions subject to the pair of residual dynamic equations, where it was necessary, to describe the gravitational structure, to write the equations in terms of $G^{\dot{A}}$, a first integral of H such that

$$\phi H \equiv \partial_{\dot{A}} \phi^2 G^{\dot{A}}. \tag{2.12a}$$

We want to emphasize, however, that writing H in terms of the two components of $G^{\dot{A}}$ is just a matter of convenience, rather than an increase in the number of key functions, since *any* $G^{\dot{A}}$ satisfying Eqs. (2.9)—(2.13) will do. In particular, if we define C and D by

$$J^{\dot{A}} \partial_{\dot{A}} D \equiv \phi H \equiv K^{\dot{A}} \partial_{\dot{A}} C, \tag{2.12b}$$

then *either* of the choices $G^{\dot{A}} = \phi^{-2} K^{\dot{A}} C$, or $G^{\dot{A}} = \phi^{-2} J^{\dot{A}} D$ is an acceptable choice of $G^{\dot{A}}$, demonstrating that there is only one degree of freedom in $G^{\dot{A}}$, corresponding to, say, C, with the other component being just a gauge freedom. Since both of these choices can be useful in particular cases, we will continue using $G^{\dot{A}}$ in the rest of the formulas.

We must now give the pair of equations to which the Einstein—Maxwell equations have been reduced. They are most simply expressed in terms of two auxiliary quantities, which are simply abbreviations for combinations of W and $G_{\dot{A}}$:

$$\begin{aligned} \text{II. } B^{\dot{A}} &\equiv \phi^{-2} \partial^{\dot{A}} W + 2\epsilon G^{\dot{A}}, \\ \epsilon\chi &= \phi^{-1} \eth^{\dot{C}} B_{\dot{C}} + \tfrac{1}{2}\phi^4 (\partial^{\dot{C}} B^{\dot{D}}) \partial_{(\dot{C}} B_{\dot{D})} + (\phi J^{\dot{B}} B_{\dot{B}})^2 + (\mu \phi^3 - \lambda/3) K_{\dot{A}} B^{\dot{A}} \\ &\quad + \frac{K^{\dot{A}} p_{\dot{A}}}{2\tau^2} \{K^{\dot{B}} p_{\dot{B}} J^{\dot{C}} - (\phi + \kappa) K^{\dot{C}}\} \mu_{,\dot{C}} - \tfrac{1}{2}(p^{\dot{A}} \epsilon_{,\dot{A}} + \delta) H \\ &\quad - N_{\dot{A}} p^{\dot{A}}, \end{aligned} \tag{2.13a}$$

$$\begin{aligned} \text{I: } B^{\dot{A}} &= \partial^{\dot{A}} \Theta + 2\epsilon G^{\dot{A}}, \\ \epsilon\chi &= \eth^{\dot{A}} B_{\dot{A}} + \tfrac{1}{2}(\partial^{\dot{A}} B^{\dot{B}}) \partial_{(\dot{A}} B_{\dot{B})} \\ &\quad - (F^{\dot{A}} + \lambda p^{\dot{A}}) B_{\dot{A}} - \lambda\Theta + (F^{\dot{A}} p_{\dot{A}})^2/18 - (F^{\dot{A},\dot{B}} p_{\dot{A}} p_{\dot{B}})/6 \\ &\quad - \tfrac{1}{2}(p^{\dot{A}} \epsilon_{,\dot{A}} + \delta) H - N_{\dot{A}} p^{\dot{A}}. \end{aligned} \tag{2.13b}$$

The residual dynamical equations then take the form:

$$\text{II: } \partial_{\dot{A}} \chi = \phi^{-2} \eth_{\dot{A}} H - 2H J^{\dot{B}} \partial_{(\dot{A}} \phi^2 B_{\dot{B})} + 3\mu \phi^2 (H K_{\dot{A}}/\tau - B_{\dot{A}}/\epsilon), \tag{2.14a}$$

$$\text{I: } \partial_{\dot{A}} \chi = \eth_{\dot{A}} H + (\lambda p_{\dot{A}} + F_{\dot{A}}) H - 2\lambda G_{\dot{A}}, \tag{2.14b}$$

where we have introduced the notations

$$\epsilon_{,\dot{A}} \equiv \frac{\partial \epsilon}{\partial q^{\dot{A}}} \quad \text{and} \quad \delta \equiv \gamma^{\dot{A}}{}_{,\dot{A}}, \tag{2.15}$$

and $N_{\dot{A}}$ is another arbitrary function of $q_{\dot{B}}$ only. It is of importance to point out that our pair of dynamical equations just forms *one* spinor-valued equation, which links the gravitational and electromagnetic key functions.

It may appear that either set of the above equations is rather complicated. However, it is to be remembered that they are the distillation of some 18 coupled Einstein—Maxwell equations. Under only the assumptions that our complex space—time admit a congruence of null strings, and the constraint $E \lrcorner \Sigma = 0$, the entire set of dynamical equations has been integrated to only this rather plausible pair of residual dynamical equations.

The result which we have here is also very general in that *all* the results of the real-valued Einstein–Maxwell theory of degenerate solutions are contained in real slices of this result.[11] We also note that if the right electromagnetic field (anti-self-dual part only) vanishes, then H, and $G_{\dot{B}}$, vanish, Eqs. (2.14a) require that $\epsilon\chi = \xi + 3\mu W$, ξ an arbitrary function of $q_{\dot{B}}$ only, and Eqs. (2.13) reduce to the hyperheavenly equation in vacuum. [See I, Eqs. (3.14)–(3.15).]

We may also list here the components of the conformal tensor, in terms of the key functions:

$$\text{II:}\quad C_{\dot{A}\mathring{B}\mathring{C}\mathring{D}} = \phi^3\partial_{(\dot{A}}\partial_{\mathring{B}}\partial_{\dot{C}}\phi^2\left\{B_{\dot{D})} - \frac{\mu}{2\tau^2}K_{\dot{D})}K_{\dot{R}}p^{\dot{R}}\right\}, \tag{2.16}$$

$$\text{I:}\quad C_{\mathring{A}\mathring{B}\dot{C}\mathring{D}} = -\,\partial_{(\dot{A}}\partial_{\dot{B}}\partial_{\dot{C}}B_{\dot{D})}, \tag{2.17}$$

$$\text{II.}\quad C^{(5)} = 0 = C^{(4)},$$

$$C^{(3)} = -\,2\phi^3(\mu - \epsilon J^{\dot{A}}\partial_{\mathring{A}}H),$$

$$C^{(2)} = 2\phi^5\{N_{\dot{A}}J^{\dot{A}} - (p^{\mathring{A}} + \kappa K^{\dot{A}}/2\tau)\mu_{,\dot{A}} + \tfrac{1}{2}J^{\dot{B}}\partial_{\mathring{B}}(p^{\mathring{C}}\epsilon_{,\mathring{C}} + \delta)H + \tfrac{1}{2}\phi\epsilon_{,\dot{B}}\partial^{\dot{B}}H + \epsilon J^{\dot{C}}H_{,\dot{C}}\},$$

$$\begin{aligned}C^{(1)} = 2\phi^7\Bigg\{&\left[\frac{1}{\tau}\left(\mu\phi^3 - \frac{\lambda}{3}\;K^{\dot{A}} - \phi\frac{\partial}{\partial q_{\dot{A}}} - \epsilon\phi^3\partial^{\mathring{A}}H\right]\right.\\
&\times\left[N_{\mathring{A}} + \frac{1}{2\tau}p_{\mathring{A}}K^{\dot{R}}\mu_{,\dot{R}} + \tfrac{1}{2}\nu\partial_{\mathring{A}}H - \tfrac{1}{2}H\epsilon_{,\mathring{A}}\right]\\
&+ J^{\dot{C}}[N_A p^{\mathring{A}} + \epsilon\chi + \tfrac{1}{2}\nu H]_{,\dot{C}} + \phi^2 J_{\mathring{A}}[2N_{\dot{A}} + H\epsilon_{,\mathring{A}} - \nu\partial_{\mathring{A}}H]\\
&\times J_{\dot{B}}B^{\dot{B}} + \tfrac{1}{2}\phi\nu\partial_{\dot{A}}H^{,\dot{A}}\\
&- \frac{\psi}{2\tau}J^{\dot{C}}p^{\dot{D}}\mu_{,\dot{C}\dot{D}} - \phi^2\left[2p_{(\dot{B}}J_{\mathring{R})}B^{\mathring{B}} + \frac{\kappa}{\tau}J_{\dot{R}}K_{\dot{B}}B^{\dot{B}}\right.\\
&\left. + \frac{\epsilon}{\tau^2}\phi^2\psi J_{\dot{R}}K^{\dot{A}}\partial_{\dot{A}}\phi^{-1}H - \epsilon Hp_{\dot{R}}\right]\mu^{,\dot{R}}\\
&- \phi^2 J^{\dot{A}}\epsilon H[2N_{\dot{A}} - 3H\epsilon_{,\dot{A}} + 3\nu\partial_{\dot{A}}H] + \phi^5 J^{\dot{A}}B_{\dot{B},\mathring{A}}\partial^{\dot{B}}\phi^{-2}\epsilon H\;,\end{aligned}$$

where $\nu \equiv p^{\dot{C}}\epsilon_{,\dot{C}} + \delta$, (2.18)

$$\text{I.}\quad C^{(5)} = 0 = C^{(4)},$$

$$C^{(3)} = -\,2\lambda/3,$$

$$C^{(2)} = -\,F^{\dot{C}}{}_{,\dot{C}} + \epsilon_{,\dot{B}}\partial^{\mathring{B}}H,$$

$$C^{(1)} = [F^{\mathring{B}} + \lambda p^{\dot{B}} - (\epsilon\partial^{\mathring{B}}H) - \eth^{\dot{B}}]\{2N_{\dot{B}} + F^{\dot{C}}{}_{,\dot{C}}p_{\dot{B}} + \nu\partial_{\dot{B}}H - H\epsilon_{,\dot{B}}\}. \tag{2.19}$$

The entire structure of the solution given above is covariant under the group of transformations which relabels the surfaces in the congruence—$q'^{\dot{R}} = q'^{\dot{R}}(q^{\dot{A}})$—and simultaneously maintains the form of the tetrad (see I). Using it one may always, for example, choose μ to be constant,[1,2] choose $N_{\dot{A}}K^{\dot{A}} = 0$, etc. In particular μ, when it is gauged to be constant, may be seen to have the interpretation of left mass, $m + in$ where n is the NUT parameter, by virtue of its role in determining the left invariant of the conformal curvature. The electromagnetic parameter ϵ also must have the role of left charge, $e + ig$ with e and g being the electric and magnetic changes respectively, by virtue of its role in the left invariant of the electromagnetic field tensor.[12]

We will also point out briefly that one may have all degenerate Petrov types of space–times in both cases, although if $\lambda = 0$ in case I, we are left only with types [III]⊗[Anything] and [N]⊗[Anything]. Also of course both cases reduce to the situation [–]⊗[Anything], which are called heavens,[13] which motivates the name hyperheavens for the present family of spaces. With respect to the electromagnetic field present, we can have an algebraically general situation. Or, when $\epsilon = 0$ (which simplifies the structural equations enormously) the left field (self-dual part) is algebraically null since

$$\mathcal{J} \equiv 8f_{AB}f^{AB} = -\,\epsilon^2\phi^4, \tag{2.20}$$

while if

$$\tilde{\mathcal{J}} \equiv 8f_{\dot{A}\dot{B}}f^{\dot{A}\dot{B}} = \tfrac{1}{2}\phi^4(\partial^{\dot{A}}\partial^{\dot{B}}H)\,\partial_{\dot{A}}\partial_{\dot{B}}H, \tag{2.21}$$

were to vanish, then the right field is algebraically null.

3. PROOF OF RESULTS

We want to give here a fairly complete outline of the method used to reduce the full set of 18 Einstein–Maxwell equations to the one spinor-valued equation given in Sec. 2. We will utilize the same spinor technology as in I, which should be seen for complete details of formalism. However, for the convenience of the reader we summarize some of it here.

In particular, we utilize the spinor-valued coordinates and tetrad described in Eqs. (2.5)–(2.8). We may then proceed to determine the form of the solutions to Maxwell's equations without sources, Eqs. (2.2a). The alignment constraint which we force on the heavenly part of the electromagnetic field—Eq. (2.4)—when reexpressed in spinor language [see Eqs. (2.1)] just becomes $f_{11} = 0$. This is easily seen by looking at the form of the bases for 2-forms in these coordinates:

$$S^{11} = E^{\dot{A}}\wedge E_{\dot{A}},\quad S^{12} = e_{\dot{A}}\wedge E^{\dot{A}},$$

$$S^{22} = e^{\mathring{A}}\wedge e_{\dot{A}} = 2\phi^{-4}\Sigma,\quad S^{\dot{A}\dot{B}} = 2e^{(\dot{A}}\wedge E^{\dot{B})}. \tag{3.1}$$

The equations $d\omega = 0$ then become

$$\partial_{\dot{A}}\phi^{-2}f_{12} = 0,\quad \partial_{\dot{A}}\phi^{-4}f_{22} + (\phi^{-2}f_{12})_{,\mathring{A}} = 0, \tag{3.2}$$

which have the immediate solutions

$$\epsilon \equiv 4\phi^{-2}f_{12},\quad 4\phi^{-4}f_{22} = p^{\dot{A}}\epsilon_{,\dot{A}} + \delta, \tag{3.3}$$

with ϵ and δ functions of $q_{\mathring{A}}$ only, and the factors of 4 have been chosen for later convenience.

Now the equations $d\tilde{\omega} = 0$ may be written in our spinorial form as

$$(\partial_{\dot{C}}\phi^{-2}f^{\dot{A}\dot{C}})\,dq_{\dot{A}}\wedge E_{\dot{B}}\wedge E^{\dot{B}} + (\eth_{\dot{C}}\phi^{-2}f^{\dot{B}\dot{C}} - \phi^{-2}f_{\dot{A}\dot{C}}\partial^{B}Q^{\dot{A}\dot{C}})\,dq^{\dot{D}}\wedge dq_{\mathring{D}}\wedge E_{\mathring{B}} = 0. \tag{3.4}$$

The two sets in the parentheses must of course vanish separately; the first—$\partial_{\dot{C}}\phi^{-2}f^{\dot{A}\dot{C}} = 0$—implies, since $f^{\dot{A}\dot{C}}$ is symmetric (see Appendix of I), the existence of a scalar function H such that

$$f^{\mathring{A}\mathring{C}} = +\tfrac{1}{4}\phi^2\partial^{\mathring{A}}\partial^{\dot{C}}H. \tag{3.5a}$$

When this is inserted into the second pair they just imply the restriction on H that $\partial^{\dot{B}}\eth_{\dot{C}}\partial^{\dot{C}}H = 0$. This implies the existence of $\alpha = \alpha(q_{\dot{A}})$ only, such that $\eth_{\dot{C}}\partial^{\dot{C}}H = \alpha$. However, if we let $H \to H + p_{\dot{A}}L^{\dot{A}}$, with $\alpha = L^{A}{}_{,A}$, then we find that H may always be gauged so that the restriction is just[14]

$$\eth_{\dot{A}}\partial^{\dot{A}}H=0. \tag{3.5b}$$

Now, we proceed to Einstein's equations with this electromagnetic field as source. We utilize the general form of the Ricci tensor components in these coordinates and this tetrad, given in I, inserting these into Eqs. (2.2b) to obtain

$$\begin{aligned}
C_{11\dot{A}\dot{B}}&=-\phi^{-1}\partial_{\dot{A}}\partial_{\dot{B}}\phi=0,\quad R=-2\phi^5\partial_{\dot{A}}\partial_{\dot{B}}\phi^{-3}Q^{\mathring{A}\dot{B}}=-4\lambda,\\
C_{12\dot{A}\dot{B}}&=-\tfrac{1}{2}\phi^4\partial^{\dot{C}}\phi^{-2}\partial_{(\dot{A}}Q_{\dot{B})\dot{C}}=-\tfrac{1}{2}\epsilon\phi^4\partial_{\dot{A}}\partial_{\dot{B}}H, \qquad (3.6)\\
C_{22\dot{A}\dot{B}}&=-\phi^6\partial_{(\dot{A}}\phi^{-4}\eth^{\dot{C}}Q_{\dot{B})\dot{C}}-\phi^3J^{\dot{C}}Q_{\dot{A}\dot{B},\dot{C}}\\
&=-\tfrac{1}{2}\phi^6(p^{\dot{A}}\epsilon_{,\dot{A}}+\delta)\partial_{\dot{A}}\partial_{\dot{B}}H.
\end{aligned}$$

The first set just has the solution already indicated in Sec. 2,

$$\phi=J_{\mathring{A}}p^{\mathring{A}}+\kappa, \tag{3.7}$$

where we may choose $J_{\dot{A}}$ and κ to be constant.[10] The equation for the Ricci scalar is of course the same as in I, since there is no contribution from the electromagnetic field. In Appendix B of I it is shown that the simplest gauge for the solution of this (singlet) equation implies the existence of a spinor $A^{\dot{B}}$ such that

$$Q^{\mathring{A}\mathring{B}}=\phi^3\partial^{(\dot{A}}A^{\dot{B})}+\frac{\lambda}{6\tau^2}K^{\mathring{A}}K^{\dot{B}}. \tag{3.8}$$

Inserting this form of $Q^{\dot{A}\dot{B}}$ into the second triple of Eqs. (3.6), we obtain (by exactly the same procedure as in Appendix A of I) the simple equations

$$\phi^4\partial_{\dot{A}}\partial_{\dot{B}}(\Lambda_0+2\epsilon H)=0,\quad \Lambda_0\equiv\phi^{-1}\partial_{\dot{C}}\phi^2A^{\dot{C}}. \tag{3.9}$$

The solution is seen to be

$$\Lambda_0+2\epsilon H=2H_{\mathring{A}}p^{\mathring{A}}+2\xi,$$

where $H_{\dot{A}}$ and ξ are functions of $q_{\dot{B}}$ only. But, as in Eq. (3.9) of I, for simplicity and without any loss of generality, we regauge $A^{\mathring{A}}$ so as to have the form

$$\phi^{-1}\partial_{\dot{C}}\phi^2A^{\dot{C}}+2\epsilon H=\Lambda_0+2\epsilon H=2\frac{\mu}{\tau}\psi\equiv 2J^{\dot{A}}H_{\dot{A}}\psi/\tau, \tag{3.10}$$

where μ is a function of $q_{\dot{B}}$ only, and

$$\psi\equiv K^{\dot{A}}p_{\dot{A}}, \tag{3.11}$$

is the "complementary" linear combination of the $p^{\dot{A}}$'s to ϕ. Inserting H in the form defined by Eq. (2.12a), this equation is easily integrated, as in I, to give

$$A^{\dot{A}}=\phi^{-2}\partial^{\dot{A}}W+\frac{\mu}{\tau^2}\psi K^{\dot{A}}-2\epsilon G^{\dot{A}}, \tag{3.12}$$

where W is an arbitrary function of all coordinate variables. We also note again that there is only one relevant degree of freedom in $G^{\dot{A}}$, as pointed out in the discussion following Eq. (2.12a). Inserting this in Eq. (3.8) gives us our form for the spinor

$$Q^{\dot{A}\dot{B}}=-\phi^3\partial^{(\mathring{A}}\phi^{-2}\partial^{\dot{B})}W-2\epsilon\phi^3\partial^{(\dot{A}}G^{\mathring{B})}+\frac{1}{\tau}\left(\mu\phi^3+\frac{\lambda}{6}\right)K^{\mathring{A}}K^{\dot{B}}, \tag{3.13}$$

which is to be inserted into the last triple of Einstein equations in Eq. (3.6). One quickly obtains those equations in the form

$$\begin{aligned}
&-\phi^6\partial_{(\dot{A}}X_{\mathring{B})}=0, &(3.14)\\
&X_{\dot{B}}=\phi^{-2}Q^{\dot{C}\mathring{D}}\partial_{\dot{D}}Q_{\dot{B}\dot{C}}-\tfrac{1}{2}\phi\partial_{(\mathring{B}}A_{\dot{C})},^{\mathring{C}}-J^{\dot{C}}A_{\dot{B},\mathring{C}}\\
&\qquad-\tfrac{1}{2}(p^{\dot{C}}\epsilon_{,\dot{C}}+\delta)\,\partial_{\dot{B}}H+\tfrac{1}{2}H\epsilon_{,\dot{B}}+\alpha p_{\dot{B}},
\end{aligned}$$

where the arbitrary α will be determined conveniently later. By explicit calculation one sees that

$$\begin{aligned}
&\phi^{-2}Q^{\dot{C}\mathring{D}}\partial_{\dot{D}}Q_{\mathring{B}\mathring{C}}=\partial_{\mathring{B}}\Delta-Q_{\mathring{B}\dot{C}}\partial_{\mathring{S}}\phi^{-2}Q^{\mathring{C}\mathring{S}},\\
&\Delta\equiv\tfrac{1}{2}\phi^{-2}Q^{\dot{R}\dot{S}}Q_{\mathring{R}\mathring{S}},
\end{aligned}$$

while

$$\begin{aligned}
-\phi\partial_{(\mathring{B}}A_{\mathring{C})},^{\mathring{C}}-J^{\mathring{C}}A_{\mathring{B},\mathring{C}}&=-\phi\partial_{\mathring{B}}A_{\dot{C}},^{\mathring{C}}+\phi\partial_{[\dot{B}}A_{\dot{C}]},^{\mathring{C}}\\
&\quad-J^{\mathring{C}}A_{\mathring{B},\dot{C}}\\
&=-\partial_{\mathring{B}}\phi A_{\mathring{C}},^{\mathring{C}}+J_{\mathring{D}}A^{\dot{D}}{}_{,\mathring{B}}\\
&\quad+\tfrac{1}{2}\phi\partial_{\dot{C}}A^{\mathring{C}}{}_{,\mathring{B}},\\
&=-\partial_{\mathring{B}}\phi A_{\mathring{C}},^{\mathring{C}}+\left(\frac{\mu\psi}{\tau}-\epsilon H\right)_{,\dot{B}},
\end{aligned}$$

where Eq. (3.10) has been used to obtain the last equality. By choosing $\alpha=-(K^{\mathring{A}}\mu_{,\dot{A}})/2\tau$, we can write most of the terms in $X_{\mathring{B}}$ as gradients:

$$\begin{aligned}
X_{\dot{B}}=-Y_{\dot{B}}+\partial_{\dot{B}}\biggl(&\Delta-\phi A_{\dot{C}},^{\mathring{C}}-\frac{K_{\dot{A}}\mu_{,\dot{C}}p^{\mathring{A}}p^{\mathring{C}}}{2\tau}\\
&-\tfrac{1}{2}(p^{\mathring{C}}\epsilon_{,\dot{C}}+\delta)H\biggr),
\end{aligned} \tag{3.15}$$

with

$$Y_{\dot{B}}=Q_{\mathring{B}\dot{C}}\partial_{\dot{S}}\phi^{-2}Q^{\dot{C}\dot{S}}+\epsilon H_{,\dot{B}}.$$

From Eq. (3.13), one finds that

$$\begin{aligned}
\partial_{\dot{S}}\phi^{-2}Q^{\mathring{C}\mathring{S}}=2\phi^{-3}J^{\dot{C}}J_{\dot{S}}\partial^{\dot{S}}W-2\epsilon\partial_{\dot{A}}\phi\partial^{(\mathring{A}}G^{\mathring{C})}\\
+(\mu-\lambda\phi^{-3}/3)K^{\mathring{C}}/\tau.
\end{aligned}$$

We also see that

$$\begin{aligned}
2\partial_{\dot{A}}\phi\partial^{(\mathring{A}}G^{\mathring{C})}&=\partial^{\dot{C}}\phi\partial_{\dot{A}}G^{\mathring{A}}+J^{\dot{C}}\partial_{\dot{A}}G^{\mathring{A}}+2J_{\dot{A}}\partial^{(\mathring{A}}G^{\mathring{C})}\\
&=\partial^{\dot{C}}H+J^{\dot{C}}\partial_{\dot{A}}G^{\mathring{A}}+2J_{\dot{A}}\partial^{[\dot{A}}G^{\dot{C}]}\\
&=\partial^{\mathring{C}}H+2J^{\dot{C}}\partial_{\dot{A}}G^{\mathring{A}},
\end{aligned}$$

from which we obtain

$$\begin{aligned}
Y_{\mathring{B}}=2\phi^{-3}Q_{\dot{B}\mathring{C}}J^{\mathring{C}}J_{\dot{S}}\partial^{\mathring{S}}W+(\mu-\lambda\phi^{-3}/3)Q_{\dot{B}\dot{C}}K^{\mathring{C}}/\tau\\
+\phi^{-2}\epsilon\,\eth_{\mathring{B}}H-2\phi^{-1}\epsilon Q_{\mathring{B}\dot{C}}J^{\mathring{C}}(H-J_{\dot{A}}G^{\mathring{A}}).
\end{aligned} \tag{3.16}$$

It is now time to utilize the constraint which H must satisfy: the wavelike equation specified by Eq. (2.9b). Commuting $\partial_{\mathring{A}}$ and $\eth^{\dot{A}}$ give us

$$0=\partial^{\mathring{B}}[\phi^{-2}\eth_{\mathring{B}}H+H\partial^{\mathring{C}}Q_{\dot{B}\mathring{C}}]-H\partial^{\mathring{B}}\partial^{\dot{C}}Q_{\dot{B}\dot{C}}.$$

Utilizing Eq. (3.13) reduces this to

$$0=\partial^{\dot{B}}[\phi^{-2}\eth_{\mathring{B}}H+H\partial^{\dot{C}}Q_{\dot{B}\mathring{C}}+3\phi\epsilon H^2J_{\mathring{B}}-6\mu\phi^2G_{\mathring{B}}],$$

which implies that there is a scalar χ satisfying

$$\phi^{-2}\eth_{\mathring{B}}H+H\partial^{\dot{C}}Q_{\mathring{B}\mathring{C}}+3\phi\epsilon H^2J_{\mathring{B}}-6\mu\phi^2G_{\dot{B}}+\partial_{\mathring{B}}\sigma=\partial_{\dot{B}}\chi,$$

where σ is just an explicit indication of the fact that χ is arbitrary to within some choice of σ. We intend to choose σ in such a way that the nonlinear terms in the final equations are simplified: In particular, we here choose $\sigma=-\tfrac{1}{2}\epsilon\phi^2H^2+3(\mu/\epsilon)W$, which gives Eqs. (2.14a) as the determining equations for χ. However, we may now use this equation to eliminate the $\phi^{-2}\epsilon\partial_{\dot{B}}H$ term in Eq. (3.16) for $Y_{\dot{B}}$ in favor of $\partial_{\dot{B}}\chi$. Also we notice that the first two terms in our equation for $Y_{\dot{B}}$ can be written:

$$
\begin{aligned}
&2\phi^{-3}Q_{\dot{B}\dot{C}}J^{\dot{C}}J_{\dot{S}}\partial^{\dot{S}}W+[\mu-(\lambda/3)\phi^{-3}]Q_{\dot{B}\dot{C}}K^{\dot{C}}/\tau\\
&\quad=-\partial_{\dot{B}}\left[(\phi^{-1}J^{\dot{C}}\partial_{\dot{C}}W)^2+\frac{1}{\tau}\left(\mu\phi^4\partial_{\dot{C}}\phi^{-3}-\tfrac{1}{3}\lambda\phi^{-2}\partial_{\dot{C}}\right)K^{\dot{C}}W\right]\\
&\qquad+4\epsilon(\partial_{(\dot{B}}G_{\dot{C})})J^{\dot{C}}J_{\dot{S}}\partial^{\dot{S}}W+\frac{\epsilon}{\tau}\left(\mu\phi^3-\frac{\lambda}{3}\right)K^{\dot{C}}\partial_{(\dot{B}}G_{\dot{C})}. \qquad (3.17)
\end{aligned}
$$

Inserting this result into Eq. (3.16), also, we can collect terms and find that

$$
\begin{aligned}
Y_{\dot{B}}=&-\partial_{\dot{B}}[(\phi^{-1}J_{\dot{C}}B^{\dot{C}})^2+2\frac{\epsilon}{\tau}\left(\mu\phi^3-\frac{\lambda}{3}\right)K_{\dot{C}}G^{\dot{C}}-\epsilon\chi\\
&+\frac{\mu}{\tau}\phi^4K^{\dot{C}}\partial_{\dot{C}}\phi^{-3}W-\frac{\lambda}{3\tau}\phi^{-2}K^{\dot{C}}\partial_{\dot{C}}W\\
&+\epsilon\phi^2H(\epsilon H-2J_{\dot{C}}B^{\dot{C}}).
\end{aligned}
$$

This result, inserted into Eq. (3.15), tells us that $X_{\dot{B}}$ is a gradient! That is, utilizing Eq. (3.14), this last triple is also reduced to

$$
\begin{aligned}
&\partial_{\dot{A}}\partial_{\dot{B}}\Lambda_-=0,\\
&\Lambda_-=\Delta+(\phi^{-1}J_{\dot{A}}B^{\dot{A}})^2-\phi A^{\dot{A}}{}_{,\dot{A}}-\frac{1}{2\tau}K_{\dot{A}}\mu_{,\dot{B}}p^{\dot{A}}p^{\dot{B}}\\
&\qquad+\frac{2\epsilon}{\tau}\left(\mu\phi^3-\frac{\lambda}{3}\right)K_{\dot{A}}G^{\dot{A}}+\frac{\mu}{\tau}\phi^4K^{\dot{C}}\partial_{\dot{C}}\,\phi^{-3}W-\epsilon\chi\\
&\qquad-\frac{\lambda}{3\tau}\,\phi^{-2}K^{\dot{C}}\partial_{\dot{C}}W+\epsilon\phi^2H(\epsilon H-2J_{\dot{C}}B^{\dot{C}})-\tfrac{1}{2}(p^{\dot{C}}\epsilon_{,\dot{C}}+\delta)H.
\end{aligned}
\qquad (3.18)
$$

Now, however, we re-express Δ:

$$
\begin{aligned}
\Delta&\equiv\tfrac{1}{2}\phi^{-2}Q^{\dot{A}\dot{B}}Q_{\dot{A}\dot{B}}\\
&=-\tfrac{1}{2}\phi^4(\partial^{(\dot{A}}A^{\dot{B})})\partial_{\dot{A}}A_{\dot{B}}-\phi\frac{\lambda}{6\tau^2}K^{\dot{A}}K^{\dot{B}}\partial_{\dot{A}}A_{\dot{B}}\\
&=-\tfrac{1}{2}\phi^4(\partial^{\dot{A}}A^{\dot{B}})\partial_{\dot{A}}A_{\dot{B}}-\tfrac{1}{4}\phi^4(\partial^{\dot{A}}A_{\dot{A}})^2-\phi\frac{\lambda}{6\tau^2}K^{\dot{A}}\partial_{\dot{B}}K^{\dot{B}}A_{\dot{A}}.
\end{aligned}
$$

At this point we insert Eqs. (3.12) and (2.12a), collect terms, and obtain

$$
\begin{aligned}
\Delta+(\phi J_{\dot{A}}B^{\dot{A}})^2=&-\tfrac{1}{2}\phi^4(\partial^{\dot{A}}B^{\dot{B}})\partial_{\dot{A}}B_{\dot{B}}\\
&+\frac{\mu}{\tau^2}K^{\dot{A}}\partial_{\dot{A}}K^{\dot{B}}B_{\dot{B}}-\epsilon\phi^2H(\epsilon H-2J_{\dot{C}}B^{\dot{C}}),
\end{aligned}
$$

which, inserted into Eq. (3.18), gives the final form for Λ_-. However, Eq. (3.18) tells us of the existence of $N_{\dot{A}}$ and ξ, functions of $q_{\dot{B}}$ only, such that

$$\Lambda_-=N_{\dot{A}}p^{\dot{A}}+\xi. \qquad (3.19)$$

But, since χ is only defined modulo a function independent of $p^{\dot{A}}$, we absorb ξ into χ and use Eq. (3.19) as a definition of $\epsilon\chi$, which is just Eq. (2.13a), if one replaces $\phi B^{\dot{A}}{}_{,\dot{A}}$ with its equivalent $\phi^{-1}\eth_{\dot{A}}B^{\dot{A}}+\phi Q_{\dot{A}\dot{B}}\partial^{\dot{B}}B^{\dot{A}}$ [from Eq. (2.7)], to be inserted into Eqs. (2.14a), . which are then the final dynamical equations left: Every solution, for W and H, of Eqs. (2.14a) determines a solution of type II of the complete set of Einstein—Maxwell equations.

To calculate the values which the conformal tensor components take when these constraints are taken into account, we just insert the expressions for $Q^{\dot{A}\dot{B}}$ into the general expressions given in I. However, to facilitate the computation for $C^{(1)}$ and $C^{(2)}$ we note that the expression for $X_{\dot{A}}$ can be solved for $\eth^{\dot{C}}Q_{\dot{B}\dot{C}}$,

$$
\begin{aligned}
\phi^{-4}\eth^{\dot{C}}Q_{\dot{B}\dot{C}}=&N_{\dot{B}}+J^{\dot{C}}A_{\dot{B},\dot{C}}+\tfrac{1}{2}(p^{\dot{C}}\epsilon_{\dot{C}}+\delta)\,\partial_{\dot{A}}H\\
&+K^{\dot{R}}\mu_{,\dot{R}}p_{\dot{B}}/2\tau-\tfrac{1}{2}H\epsilon_{,\dot{B}}.
\end{aligned}
\qquad (3.20)
$$

The expressions for $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$, $C^{(3)}$, and $C^{(2)}$ are then obtained in a quite straightforward way. However, the expression for $C^{(1)}$ requires the insertion of Eq. (2.14a) two different times to simplify the expression.

Lastly, we want to point out how to determine the analogous equations for type I spaces. These are obtained by a limiting process in which $J_A\rightarrow 0$, $\kappa\rightarrow 1$. (In this process, of course, $K_{\dot{A}}, \mu$ and ψ also vanish.) In order to insure that all important equations have a (finite) limit, we must insist that W diverges but that[15]

$$\Theta\equiv W-\mu\psi^2\phi(\phi+2\kappa)/6\tau^2, \qquad (3.21)$$

is finite in this limit.[16] (We could, of course, have used this Θ as a key function, instead of W, originally, but our choice of W gives the simplest form to $Q^{\dot{A}\dot{B}}$ and χ in type II.) If we now give names, as before, to certain limits,

$$F^{\dot{A}}\equiv\lim_{J_{\dot{C}}\rightarrow 0}\mu K^{\dot{A}}/\tau, \qquad (3.22)$$

we obtain the limit

$$A^{\dot{B}}=-\partial^{\dot{B}}\Theta+\tfrac{2}{3}p^{\dot{B}}p_{\dot{C}}F^{\dot{C}}-2\epsilon G^{\dot{B}}. \qquad (3.23)$$

Then, by repeating the process described in Appendix B of I for adding λ to a type I solution, we obtain the expression given in Eq. (2.10) for type I. Similarly one finds the expressions in Eqs. (2.13b) and (2.14b) for χ and the hyperheavenly equation for type I, through this limiting process, taking H (and $G^{\dot{A}}$) as having finite limits.

4. CONCLUSIONS

We consider this paper as a further technical step toward better understanding of the analytic structure of the complex Einstein equations. We believe that the problem of the integration of these equations (which is surely much more straightforward on a complex manifold) and the problem of determining all real slices are of a very different nature. Therefore, we have decomposed the entire, rather complicated problem into these two separate parts. For progress on the question of finding real slices of complex solutions see Refs. 11 and 17. As a preliminary step toward a fuller understanding of the physical meaning of the techniques involved in our work on integrating the complete equations, the seven-parametric real solution of the Einstein—Maxwell equations of Plebański and Demiański is put into the formalism of case II hyperheavens,[12] explicitly giving all key functions of this space—time in terms of physical constants and real coordinates with well established physical interpretation.[18]

Thus, although the results of the present paper are, at the moment, technical ones, we consider them as some steps toward the genuine goal, which is fuller mastery of the physically interesting integral manifolds of the Einstein—Maxwell equations.[19] We believe that these spinorial structures will certainly play an important role in the (accumulating) progress in this field.

[1] J. F. Plebański and I. Robinson, Phys. Rev. Lett. **37**, 493 (1976). See also the proceedings of the Cinquieme Colloque International sur les Methodes de la Theorie des Groupes in Physique, Montreal July, 1976; to be published by Academic Press, and a detailed discussion in the Proceedings of the Symposium on Asymptotic Structure of Space—Time, University of Cincinnati, Ohio, June, 1976, F.P. Esposito and L. Witten (Eds.) (Plenum, New York, 1977).

[2] J. D. Finley, III and J. F. Plebański, J. Math. Phys. **17**, 2207 (1976). This formalism was first used, in pure heavens—$C_{\dot{A}\dot{B}\dot{C}\dot{D}}=0$, $\Gamma_{\dot{A}\dot{B}}=0$—by C. Boyer and J. F. Plebański, J. Math. Phys. **18**, 1022 (1977).

[3] A. Garcia, J. F. Plebański, and I. Robinson, "Null Strings and Complex Einstein—Maxwell Fields with Cosmological Constant," to be published in Intl. J. Theor. Physics.

[4] J. F. Plebański and S. Hacyan, J. Math. Phys. **16**, 2403 (1975).

[5] Our definition of duality is so arranged that $\omega = {}^{**}\omega$, for an arbitrary p-form: If

$$\omega = (p\,!)^{-1}\omega_{\mu_1\cdots\mu_p}, dx^{\mu_1}\wedge\cdots\wedge dx^{\mu_p},$$

then, with $p+p'=4$,

$$^{*}\omega = \exp[(i\pi/2)(pp'-1)](p\,!p'\,!)^{-1}|\det(g_{\mu\nu})|^{1/2}$$

$$\times\epsilon^{\lambda_1\cdots\lambda_p}{}_{\mu_1\cdots\mu'_p}\ \omega_{\lambda_1\cdots\lambda_p}dx^{\mu_1}\cdots dx^{\mu_{p'}}.$$

Also note that the S^{AB} and $S^{\dot{A}\dot{B}}$ are spinor components of the bases of self-dual and anti-self-dual 2-forms respectively; see Ref. 6 or 9.

[6] See I (Ref. 2) for more information on this aspect of spinor notation. Also we point out here that this form of the null tetrad formalism and real spinor applications to general relativity may be easily reviewed in G.C. Debney, R.P. Kerr, and A. Schild, J. Math. Phys. **10**, 1842 (1969) or J. F. Plebański, "Spinors, Tetrads and Forms," unpublished monograph from Centro de Invest. y Estd. Avanz. del I. P. N. Apdo. Postal 14-740, Mexico 14, D. F. See also Ref. 9.

[7] We use $\otimes_s$ for the symmetrized tensor product: $A\otimes_s B \equiv \frac{1}{2}[A\otimes B + B\otimes A]$.

[8] J. F. Plebański and A. Schild, "Complex Relativity and Double KS Metrics," Nuovo Cimento B **35**, 35 (1976).

[9] A concise summary of the spinorial form of Riemannian structures is given in this format in J. F. Plebański, J. Math. Phys. **16**, 2396 (1975). See also I. We also note, as in I, that the spinor indices are to be manipulated with a skew spinor, such as the customary $\epsilon^{\dot{A}\dot{B}}$. However, we note that the usual chain rules for the calculus require that these rules be *different* in the tangent and cotangent (spinor) spaces: $p_{\dot{A}} = \epsilon^{\dot{B}\dot{A}}p_{\dot{B}}, p^{\dot{A}} = \epsilon^{\dot{B}\dot{A}}p_B$, but $\partial/\partial p_{\dot{A}} \equiv \partial^{\dot{A}} = \epsilon^{\dot{A}\dot{B}}\partial_{\dot{B}}$, $\partial_{\dot{A}} = \epsilon_{\dot{B}\dot{A}}\partial^{\dot{B}}$.

[10] In Refs. 1 and 2 a $q_{\dot{A}}$ dependent group of transformations which leave the form of the metric invariant is discussed and the transformation properties, under the group, of all pertinent quantities is discussed. These transformations amount to a relabeling scheme for the various null strings. It is shown there that one may always make a choice of coordinates $q_{\dot{A}}$ so that $J_{\dot{A}}$ and κ are constant. Because of the ubiquitous nature of $J_{\dot{A}}$, this choice is extremely useful; therefore, we adopt it here.

[11] K. Rózga, thesis, Institute of Theoretical Physics, University of Warsaw, Warsaw, Poland, March, 1976, to be published in Rep. Math. Phys.

[12] See the paper by A. Garcia and J. F. Plebański, "Seven Parametric Type D Solution of Einstein—Maxwell Equations in the basic Left-Degenerate Representation," to be published in Gen. Rel. Grav., in which the (real) seven-parametric solution of the Einstein—Maxwell equations, of type D, of J. F. Plebański and M. Demiański, Ann. Phys. (N.Y.) **98**, 98 (1976) is recast into the structure of a case II electromagnetic hyperheaven.

[13] E. T. Newman, GR7 Conference, Tel Aviv, 1974.

[14] H is somehow reminiscent of the Hertz—Debye potentials. The general massless (irreducible) spinor field equations have solutions of much the same nature. See J. D. Finley, III and J. F. Plebański, J. Math. Phys. **17**, 585 (1976) for the special case of pure heaven.

[15] This Θ differs from the Θ defined in Eq. (4.2) of I by an additive factor which is finite in the limit in question. This different choice has been made so as to effect the "natural spinorial gauge" for Θ for type I, referred to at the end of Appendix B of I.

[16] This is the relation between Θ and W for zero value of λ. We have used the procedure of setting $\lambda=0$ in the type II equations, taking the limit to type I, and then adding back the appropriate λ terms because this is considerably simpler than actually regauging (see Appendix B of I) so as to keep finite all λ terms in the limit $J_{\dot{A}}\to 0$.

[17] E. Flaherty, *Hermitian and Kählerian Geometry in Relativity* (Springer, Berlin, 1976).

[18] For example, one might see R.H. Boyer and R.W. Lindquist, J. Math. Phys. **8**, 265 (1967) or W. Kinnersley and M. Walker, Phys. Rev. D **2**, 1359 (1970). A very complete account is given in J. F. Plebański and M. Demiański, Ann. Phys. (N.Y.) **98**, 98 (1976).

[19] See also a recent preprint by E. Flaherty on "Einstein—Maxwell Pseudo-Kählerian Spacetimes," which form a large class of weak heavens (in our terminology), and are therefore strongly related to this current investigation.

A class of inhomogeneous perfect fluid cosmologies

D. A. Szafron and J. Wainwright

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada

(Received 9 December 1976)

We present a new class of inhomogeneous and anisotropic cosmologies with perfect fluid matter content. The subclass of solutions with vanishing pressure are members of the class of inhomogeneous pressurefree cosmologies recently found by Szekeres. In addition, the Robertson–Walker solution with flat space sections and equation of state $p = (\gamma - 1)\mu$, is contained as a special case. The relation between our solutions and this Robertson–Walker solution in the limit of large cosmological time is studied in detail.

1. INTRODUCTION

This paper presents a new class of exact solutions of the Einstein field equations with a perfect fluid source. The fluid has zero acceleration and vorticity but nonzero shear and expansion.[1] In addition, the spacelike hypersurfaces orthogonal to the fluid flow are in general not the orbits of a group of isometries. The solutions thus have relevance as *inhomogeneous* and *anisotropic* cosmological models. We discuss this interpretation and in particular consider their relationship to the Robertson–Walker models in the limit of large cosmological time.

The following known solutions are included in the general class:

(1) the general Robertson–Walker (henceforth abbreviated to R–W) solution with flat space sections and equation of state of the form $p = (\gamma - 1)\mu$, $1 \leqslant \gamma \leqslant 2$,
(2) an inhomogeneous and anisotropic solution found by Stephani,[2]
(3) a subclass of the inhomogeneous and anisotropic dust solutions found by Szekeres.[3]

In Sec. 2, we present the solutions and describe their relationship to the Stephani and Szekeres solutions. The kinematics of the fluid, the properties of the Weyl tensor, and the existence of groups of isometries are discussed in Sec. 3. In the final section we study the asymptotic behavior of the solutions as the cosmological time tends to infinity. Many of the calculations in the paper were performed using the Newman–Penrose[4] formalism. Most of the paper can, however, be read without a knowledge of this formalism.

2. THE SOLUTIONS

The line element is given by

$$ds^2 = dt^2 - Q(t)^{4/3}(dx^2 + dy^2) - Q(t)^{-2/3}H^2 dz^2, \qquad (2.1)$$

where

$$Q(t) = C_1 t^{1-q} + C_2 t^q,$$

$$H = (X + C_1 F)t^{1-q} + (Y + C_2 F)t^q,$$

and q, C_1, C_2 are constants. The functions F, X, and Y are defined by[5]

$$F = a(z)x + b(z)y + \tfrac{1}{2}(1 - 2q)C(z)(x^2 + y^2),$$

$$X = B(z) + C(z)\int Q(t)^{-1/3} t^q \, dt,$$

$$Y = A(z) - C(z)\int Q(t)^{-1/3} t^{1-q} \, dt,$$

The metric thus depends on five arbitrary functions of z, namely $A(z)$, $B(z), C(z), a(z), b(z)$ and on three parameters, q, C_1, C_2. At least one of the arbitrary functions must not vanish identically, and at least one of C_1 and C_2 must be nonzero. Not all of these functions and parameters are essential. Any one of the arbitrary functions which is not identically zero may be set to be a constant by redefining the z coordinate. In addition, either C_1 or C_2, if positive, may be set equal to 1 by changing the scale of the x and y coordinates.

This line element satisfies the Einstein field equations

$$G_{ij} = -8\pi[(\mu + p)u_i u_j - p g_{ij}], \qquad (2.2)$$

with the fluid velocity u, density μ, and pressure p given by

$$u = \frac{\partial}{\partial t}, \qquad (2.3)$$

$$8\pi\mu = \tfrac{4}{3}H^{-1}Q(t)^{-1}\left[Q'(t)\frac{\partial H}{\partial t} + \tfrac{3}{2}(2q - 1)C(z)Q(t)^{2/3}\right], \qquad (2.4)$$

$$8\pi p = \tfrac{4}{3}q(1 - q)t^{-2}. \qquad (2.5)$$

Until restrictions are imposed on the density and pressure, the parameter q can assume all real values. However the solutions with $q \leqslant \frac{1}{2}$ are equivalent to the solutions with $q \geqslant \frac{1}{2}$, since the equations which define the solution are invariant under the substitutions $q \longleftrightarrow 1 - q$, $C_1 \longleftrightarrow C_2$, $A(z) \longleftrightarrow B(z)$, and $C(z) \longleftrightarrow -C(z)$. We will henceforth, without loss of generality, restrict q to the range

$$q \geqslant \tfrac{1}{2}.$$

The general dust solution in the above class is given by $q = 1$. If $C_2 \neq 0$, we can use the coordinate freedom to set $C_2 = 1$, $C_1 = 0$, $B(z) = \epsilon$, where $\epsilon = 0$ or ± 1, and if $C_2 = 0$, we can set $C_1 = 1$, $A(z) = \epsilon$, where $\epsilon = 0$ or ± 1. The resulting solutions are the solutions of Szekeres, mentioned in the Introduction. The Stephani solution referred to earlier, is obtained when $q = \frac{3}{4}$, $C_1 = 0$, and $C_2 = 1$.

3. PROPERTIES OF THE SOLUTIONS

The solutions of Sec. 2 can be conveniently studied using the Newman–Penrose[4] formalism. A suitable null tetrad for the line element (2.1) is defined by the following equations:

$$k_a dx^a = 2^{-1/2}[dt - Q(t)^{-1/3} H dz],$$
$$n_a dx^a = 2^{-1/2}[dt + Q(t)^{-1/3} H dz], \qquad (3.1)$$
$$m_a dx^a = 2^{-1/2} Q(t)^{2/3}(dx + idy).$$

With this choice of null tetrad the fluid velocity vector (2) is given by

$$u^a = 2^{-1/2}(k^a + n^a). \qquad (3.2)$$

The field equations (2.2) then assume the form

$$\phi_{01} = \phi_{02} = \phi_{12} = 0,$$
$$\phi_{00} = 2\phi_{11} = \phi_{22} = 2\pi(\mu + p),$$
$$(\phi_{11} - 3\Lambda) = 4\pi p$$

[see Wainwright[6]]. Using this form of the field equations and the general formulas given in the Appendix, it is straightforward to verify that the metric (2.1) satisfies the field equations (2.2), with μ and p given by (2.4) and (2.5).

As regards the kinematic quantities[1] of the fluid, it is an immediate consequence of (2.1) and (2.3) that the acceleration $\dot{u}$ and the vorticity vector ω are zero. The rate of shear tensor and the expansion scalar can be calculated using the formulas given by Wainwright[7] [see Appendix 1] and the expressions for the spin coefficients of the null tetrad (3.1) given in the Appendix. One finds that the rate of shear is

$$\sigma_{ab} = \chi[v_a v_b - m_{(a}\bar{m}_{b)}], \quad v_a = 2^{-1/2}(k_a - n_a),$$

where

$$\chi = -\tfrac{2}{3}(1 - 2q)H^{-1}Q(t)^{-1}(C_2 X - C_1 Y), \qquad (3.3)$$

and that the expansion is

$$\theta = 2Q'(t)/Q(t) - 3\chi/2. \qquad (3.4)$$

It thus follows that the fluid has nonzero shear provided that $C_2 X - C_1 Y \neq 0$ and $q \neq \frac{1}{2}$. It can further be shown that the expansion never vanishes identically.

The tetrad components of the Weyl tensor can be calculated using the formulas in the Appendix. One finds that

$$\psi_0 = \psi_1 = 0, \quad \psi_3 = \psi_4 = 0,$$
$$\psi_2 = \tfrac{2}{9}(1 - 2q)H^{-1}Q(t)^{-2}[Q'(t)(C_2 X - C_1 Y) \qquad (3.5)$$
$$- \tfrac{3}{2}C(z)Q(t)^{5/3}].$$

This implies that *the spacetimes are of Petrov type D or are conformally flat*, and that in the type D spacetimes the null tetrad vectors k and n define the repeated principal null directions of the Weyl tensor. Unlike the case of type D vacuum spacetimes, *the repeated principal null directions[8] are in general not tangent to null geodesics*. The subclass in which these directions are tangent to null geodesics (i.e., the spin coefficients κ, ν are zero) is defined by $a(z) = b(z) = C(z) = 0$, as follows from the formulas in the Appendix.

The expressions (3.5) enable one to calculate the scalar polynomial curvature invariants of the Weyl tensor. It can be shown using the formula for the Weyl tensor in terms of the null tetrad vectors (see for example, Trim and Wainwright,[9] p. 545) that

$$C_{abcd}C^{abcd} = 48\psi_2^2, \quad C_{abcd}{}^*C^{abcd} = 0,$$
$$C_{abcd}C^{cd}{}_{rs}C^{rsab} = 96\psi_2^3, \quad C_{abcd}{}^*C^{cd}{}_{rs}C^{rsab} = 0.$$

Hence the behavior of these invariants can be studied using the expression for ψ_2.

It follows from (3.5) that the solutions are conformally flat if and only if

$$q = \tfrac{1}{2} \text{ or } C_2 B(z) - C_1 A(z) = 0 = C(z).$$

Since the fluid acceleration and vorticity are zero [and the shear is necessarily zero, see (3.3)] it follows (see, for example, Ref. 1, p. 135) that the solutions in this case belong to the R–W class of solutions. The spatial coordinates are however nonstandard ones since the metric has the form

$$ds^2 = dt^2 - Q(t)^{4/3}\{dx^2 + dy^2 + [f(z) + a(z)x + b(z)y]^2 dz^2\}, \qquad (3.6)$$

where $f(z)$ is related to $A(z)$, $B(z)$, C_1, and C_2. As pointed out by Bonnor and Tomimura,[10] the space sections t = const for a metric of this form are flat. The expressions (2.4) and (2.5) for the pressure and density reduce to

$$8\pi\mu = \tfrac{4}{3}[Q'(t)/Q(t)]^2, \quad 8\pi p = \tfrac{4}{3}q(1 - q)t^{-2}.$$

In the special case $C_1 = 0$, it follows that $p = (q^{-1} - 1)\mu$. Hence *the solution of Sec. 2 contains the R–W solution with flat space sections and equation of state of the form* $p = (\gamma - 1)\mu$. The requirement $1 \leq \gamma \leq 2$ restricts q according to $\frac{1}{2} \leq q \leq 1$.

It can be shown that the general solution of Sec. 2 admits no Killing vector fields. There are however a variety of special cases which admit nontrivial local isometry groups. We list some of the possibilities in Table I, with $C(z) = 0$ and $B(z) > 0$ for simplicity. We choose the z coordinate so that $B(z) = 1$, and require $A(z) \neq 0$ so that the solutions are not conformally flat.

Remarks: 1. The solutions of class 5 are examples of spatially homogeneous cosmologies of Bianchi type I,[11]

TABLE I. Killing vector fields admitted by metrics with $C(z) = 0$, $B(z) = 1$. The dimension of the isometry group and its isotropy subgroup are r and s, respectively.

	Restrictions on A, a, b	r	s	KVF's
1	None	0	0	—
2	a, b nonzero constants	1	0	$\frac{b\partial}{\partial x} - \frac{a\partial}{\partial y}$
3	A, a, b nonzero constants	2	0	$\frac{b\partial}{\partial x} - \frac{a\partial}{\partial y}, \frac{\partial}{\partial z}$
4	$a = b = 0$	3	1	$\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{y\partial}{\partial x} - \frac{x\partial}{\partial y}$
5	$a = b = 0, \quad A = \text{const} \neq 0$	4	1	As in 4, plus $\frac{\partial}{\partial z}$

which are also locally rotationally symmetric. These solutions with $p \neq 0$ do not appear to have been given previously.[12]

2. The solutions of class 4 are the simplest models which are not spatially homogeneous. They are generalizations of one of the dust solutions given by Ellis,[13] and appear to be new. They occur in class II of the classification scheme of locally rotationally symmetric perfect fluid solutions of Stewart and Ellis.[14]

4. ASYMPTOTIC BEHAVIOR

We are interested in those solutions of Sec. 2 which are not R–W solutions (i.e., $\psi_2 \neq 0$) but which in some sense approach a R–W solution as the cosmological time t tends to infinity. We also demand that $p \geq 0$, $p \neq 0$. It follows from (2.5) and (3.5) that q must be restricted according to

$$\tfrac{1}{2} < q < 1. \tag{4.1}$$

We permit the coordinates to assume the values

$$t_0 < t < \infty,$$
$$-\infty < x, y, z < \infty,$$

and consider two types[15] of limits as $t \to \infty$, defined as follows:

Definition 1: A function $F(x, y, z, t)$ is said to have *a pointwise limit L as $t \to \infty$ along the fluid flow lines* if for each fixed x, y, z,

$$\lim_{t\to\infty} F = L.$$

Definition 2: A function $F(x, y, z, t)$ is said to have *a uniform limit L as $t \to \infty$ along the fluid flow lines* if

$$\lim_{t\to\infty}\Bigl(\operatorname*{lub}_{t=\text{const}} |F(x, y, z, t) - L|\Bigr) = 0.$$

In both cases L is a constant.

For example, the function $F = 1 + (ax + by + cz)/t$, with a, b, c constant, has $L = 1$ as a pointwise limit, but not as a uniform limit, as $t \to \infty$.

We first consider the asymptotic behavior of μ and p. If $C(z) \neq 0$, it follows from (2.4) that

$$8\pi\mu \approx \begin{cases} \frac{4}{3}(q-1)(q+3)t^{-2}, & \text{if } C_2 \neq 0, \\ \frac{4}{3}q(q-4)t^{-2}, & \text{if } C_2 = 0, \end{cases}$$

as $t \to \infty$. This implies that $\mu < 0$ for t sufficiently large. We thus restrict our considerations in the remainder of this section to solutions with

$$C(z) \equiv 0. \tag{4.2}$$

When (4.2) holds, it follows from (2.4) that

$$8\pi\mu \approx \begin{cases} \frac{4}{3}q^2 t^{-2}, & \text{if } C_2 \neq 0, \\ \frac{4}{3}q(1-q)t^{-2}, & \text{if } C_2 = 0, \end{cases}$$

as $t \to \infty$, for fixed x, y, z. Thus

$$\lim_{t\to\infty}(p/\mu) = \begin{cases} \gamma - 1, \quad \gamma = 1/q, & \text{if } C_2 \neq 0, \\ 1, & \text{if } C_2 = 0, \end{cases} \tag{4.3}$$

in the pointwise sense. As regards the expansion and rate of shear of the fluid, one finds, using (3.3) and (3.4), that

$$\theta \approx \begin{cases} 2qt^{-1}, & \text{if } C_2 \neq 0, \\ t^{-1}, & \text{if } C_2 = 0, \end{cases}$$

and

$$\chi \approx \begin{cases} 2(1-2q)(C_1 A - C_2 B)t^{-2q}/[3C_2(A + C_2 F)], & \text{if } C_2 \neq 0, \\ \frac{2}{3}(1-2q)t^{-1}, & \text{if } C_2 = 0, \end{cases}$$

where $\chi^2 = 2\sigma_{ab}\sigma^{ab}/3$. It thus follows that

$$\lim_{t\to\infty}(8\pi\mu/\theta^2) = \begin{cases} \frac{1}{3}, & \text{if } C_2 \neq 0, \\ \frac{4}{3}q(1-q), & \text{if } C_2 = 0, \end{cases} \tag{4.4}$$

in the pointwise sense. We note that this dimensionless ratio[16] equals $\frac{1}{3}$ for all t in the expanding R–W solutions[17] with $k = 0$. Secondly we can conclude that

$$\lim_{t\to\infty}(\sigma_{ab}\sigma^{ab}/\theta^2) = \begin{cases} 0, & \text{if } C_2 \neq 0, \\ 2(1-2q)^2/3, & \text{if } C_2 = 0, \end{cases} \tag{4.5}$$

in the pointwise sense. This suggests[18] that when $C_2 \neq 0$, the model approaches isotropy along each fluid worldline as $t \to \infty$. This is borne out by the behavior of the line element (2.1) subject to (4.2). One finds that as $t \to \infty$ this line element approaches a R–W form similar to (3.6) if $C_2 \neq 0$, while if $C_2 = 0$, it approaches the nonisotropic form

$$ds^2 = dt^2 - t^{4(1-q)/3}(dx^2 + dy^2) - t^{-2(1-4q)/3}dz^2,$$

provided we use the coordinate freedom to set $C_1 = 1$, $A(z) = 1$.

As regards the Weyl tensor, it follows from (3.3)–(3.5) and (4.2) that

$$\psi_2 = -\tfrac{1}{3}\chi Q'(t)/Q(t). \tag{4.6}$$

Thus $\lim_{t\to\infty}\psi_2 = 0$, in the pointwise sense, independently of whether C_2 is zero. In order to distinguish the cases $C_2 \neq 0$, $C_2 = 0$ as regards the rate at which $\psi_2 \to 0$, we consider the dimensionless ratio ψ_2/θ^2. It follows from (3.4), (4.5), and (4.6) that[19]

$$\lim_{t\to\infty}(\psi_2/\theta^2) = \begin{cases} 0, & \text{if } C_2 \neq 0, \\ 2(1-q)(2q-1)/3, & \text{if } C_2 = 0. \end{cases} \tag{4.7}$$

The limits (4.5) and (4.7) relate to the possible *isotropy* of the solution as $t \to \infty$. To obtain invariant information concerning the possible *homogeneity* of the solution as $t \to \infty$, we consider the spatial derivatives of the density. We need not consider the pressure since it is homogeneous for all t. A plausible necessary condition for asymptotic spatial homogeneity is that *the spatial gradient of the density μ approaches zero faster than its time derivative, as $t \to \infty$*, i.e., that

$$\lim_{t\to\infty}(-\mu_{,a}\mu_{,b}h^{ab})^{1/2}/\dot{\mu} = 0, \tag{4.8}$$

at least in the pointwise sense. Here $h^{ab} = g^{ab} - u^a u^b$ denotes the projection tensor into the 3-spaces orthogonal to u^a. The time derivative $\dot{\mu}$ can be eliminated, since one of the contracted Bianchi identities reads $\dot{\mu} = -(\mu + p)\theta$ in the case of a perfect fluid [see, for example, Ref. 1, p. 117]. Thus, in a spacetime in which the limit (4.4) is nonzero, condition (4.8) is equivalent to

$$\lim_{t\to\infty} (-\mu_{,a}\mu_{,b}h^{ab})^{1/2}/\theta^3 = 0. \tag{4.9}$$

For the present class of solutions with $C_2 \neq 0$, it is straightforward to verify that

$$\lim_{t\to\infty} (-\mu_{,a}\mu_{,b}h^{ab})^{1/2}/\theta^3 = \begin{cases} 0, & \text{if } \frac{3}{4} < q < 1, \\ f(x,y,z) & \text{if } q = \frac{3}{4}, \\ \pm\infty, & \text{if } \frac{1}{2} < q < \frac{3}{4}, \end{cases} \tag{4.10}$$

where $f(x,y,z)$ is determined by $A(z)$, $B(z)$, F, C_1, and C_2. This implies in conjunction with (4.3) that *the necessary condition for asymptotic homogeneity (4.9) is fulfilled if and only if the limiting equation of state $p = (\gamma - 1)\mu$ satisfies $1 < \gamma < \frac{4}{3}$.*

The limits (4.3), (4.4), (4.5), (4.7), and (4.10) are pointwise, i.e., they describe the behavior along a single fluid worldline. It is easily verified that *if the functions a(z) and b(z) do not both vanish identically, then these limits are not uniform*, no matter how the functions $A(z)$, $B(z)$, and the constants q, C_1, C_2 are chosen. The reason for this is that when $a(z)$ and $b(z)$ are not identically zero, the function F, whose domain is $\mathbf{R}_3$, assumes all real values. This implies that H and $\partial H/\partial t$ have zeros on each spatial section $t = \text{const}$ and hence that the scalars $|\chi|$, $|\theta|$, $|\psi_2|$, and $|\mu_{,a}\mu_{,b}h^{ab}|$ are unbounded on each spatial section. In addition the density μ assumes *all* real values[20] on the spatial sections $t = \text{const}$ no matter how large t is.

Thus, although the solutions with $C(z) = 0$, $C_2 \neq 0$ approach a Friedman solution pointwise (i.e., locally) as $t \to \infty$, these solutions do not approach arbitrarily closely to a Friedman solution in the large, as $t \to \infty$, if $a(z)$ and $b(z)$ are not both identically zero.

In order to exhibit a class of inhomogeneous, anisotropic models which approach a Friedman model uniformly as $t \to \infty$, we consider the solutions subject to (4.2) and

$$a(z) \equiv 0, \quad b(z) \equiv 0, \quad C_1 = 0.$$

For the sake of clarity, we write out the metric and the various scalars in this special case. After using the coordinate freedom to set $C_2 = \pm 1$, we obtain

$$ds^2 = dt^2 - t^{4q/3}(dx^2 + dy^2) - t^{-2q/3}H^2dz^2, \tag{4.11}$$

with

$$H = B(z)t^{1-q} + A(z)t^q,$$

$$8\pi\mu = \tfrac{4}{3}q\,t^{-2}[B(z)(1-q)t^{1-q} + A(z)q\,t^q]/H,$$

$$8\pi p = \tfrac{4}{3}q(1-q)t^{-2},$$

$$\chi = -\tfrac{2}{3}(1-2q)t^{-q}B(z)/H,$$

$$\theta = 2qt^{-1} - 3\chi/2,$$

$$\psi_2 = -qt^{-1}\chi/3.$$

The coordinate ranges are $t > 0$, $-\infty < x, y, z < \infty$. We choose $A(z)$ and $B(z)$ to be nonnegative bounded functions of class C^1, with bounded derivatives on $\mathbf{R}$, and with $B(z) \neq 0$ and

$$A(z) \geq k > 0,$$

for some constant k and all real z. The metric is then analytic, and the pressure and density satisfy $0 < p < \mu$ over the whole coordinate range. In addition the density is bounded for $t \geq t_0 > 0$. The following limits hold uniformly as $t \to \infty$:

$$p/\mu \to \gamma - 1, \quad \text{with } \gamma = 1/q,$$

$$8\pi\mu/\theta^2 \to \tfrac{1}{3},$$

$$\sigma_{ab}\sigma^{ab}/\theta^2 \to 0,$$

$$\psi_2/\theta^2 \to 0,$$

for all q subject to $\frac{1}{2} < q < 1$, as can be verified using the definition. If in addition $\frac{3}{4} < q < 1$, the homogeneity condition

$$(-\mu_{,a}\mu_{,b}h^{ab})^{1/2}/\theta^3 \to 0 \tag{4.12}$$

holds uniformly as $t \to \infty$. Finally, since we demand that $A(z)$ be bounded below away from zero, we can redefine the z coordinate so that $A(z) = 1$. Then the limiting form of the metric as $t \to \infty$ is

$$ds^2 = dt^2 - t^{4q/3}(dx^2 + dy^2 + dz^2).$$

On account of the above properties, *we tentatively regard the solution defined by (4.11), with q subject to $\frac{3}{4} < q < 1$, as representing a class of inhomogeneous and anisotropic cosmologies which at large time approximate arbitrarily closely the R–W solutions with flat space sections and equation of state $p = (\gamma - 1)\mu$, $1 < \gamma < \frac{4}{3}$, where $\gamma = 1/q$.*

The value of q also affects the nature of the singularity at $t = 0$. The limiting form of the metric at $t = 0$ is

$$ds^2 = dt^2 - t^{4q/3}(dx^2 + dy^2) - t^{2(1-4q/3)}B(z)^2dz^2.$$

Thus the spacetime is highly anisotropic as $t \to 0^+$, and for $\frac{3}{4} < q < 1$, the singularity is of the cigar type.[21]

It can also be shown that the homogeneity condition (4.12) holds uniformly as $t \to 0^+$, for all q subject to $\frac{1}{2} < q < 1$. Thus *the solutions defined by (4.11) emerge from an anisotropic but homogeneous state at $t = 0$, pass through an inhomogeneous era, and finally, if $\frac{3}{4} < q < 1$, approximate an R–W solution.* The nature of the inhomogeneities is determined by the choice of the arbitrary function $B(z)$. Since the only freedom is a function of the single variable z, the inhomogeneities are necessarily layered in sheets in any spatial section $t = \text{const}$.

ACKNOWLEDGMENTS

We would like to thank Dr. C. B. Collins for a number of comments and suggestions concerning Sec. 4. This work was supported in part by a research grant (J.W.) and a postgraduate scholarship (D.S.) from the National Research Council of Canada.

APPENDIX

We list the Newman–Penrose spin coefficients and components of the Weyl and Ricci tensors relative to the null tetrad (3.1). The calculation was performed using a computer program written in the symbolic manipulation language CAMAL (see Campbell and Wainwright[22]). For convenience we introduce complex coordinates $\zeta, \bar{\zeta}$ defined by $\zeta = x + iy$, $\bar{\zeta} = x - iy$, so that $m_a dx^a = 2^{-1/2} Q^{2/3} d\zeta$, $\bar{m}_a dx^a = 2^{-1/2} Q^{2/3} d\bar{\zeta}$.

Spin coefficients:

$$\kappa = -2^{-1/2} H_{\bar{\zeta}} H^{-1} Q^{-2/3}$$

$$\rho = -2^{1/2} Q' Q^{-1}/3$$

$$\epsilon = \rho/4 + 2^{-3/2} H_t/H$$

$$\tau = -\kappa, \quad \pi = \bar{\kappa}, \quad \nu = -\bar{\kappa}, \quad \mu = -\rho, \quad \gamma = -\epsilon,$$

$$\alpha = \beta = 0 = \lambda = \sigma$$

Weyl tensor components:

$$\psi_0 = -H_{\bar{\zeta}\bar{\zeta}} H^{-1} Q^{-4/3}$$

$$\psi_1 = \tfrac{1}{2}(H_{\bar{\zeta}}/Q)_t H^{-1} Q^{1/3}$$

$$\psi_2 = -\tfrac{1}{18} H^{-1}[3H_{tt} - 4H_t Q'/Q + 6H_{\zeta\bar{\zeta}} Q^{-4/3}] + \tfrac{1}{18}[3Q''/Q - 4(Q'/Q)^2]$$

$$\psi_3 = \bar{\psi}_1, \quad \psi_4 = \bar{\psi}_0$$

Ricci tensor components:

$$\Lambda = \tfrac{1}{36} H^{-1}[3H_{tt} + 2H_t Q'/Q - 12H_{\zeta\bar{\zeta}} Q^{-4/3}] + \tfrac{1}{12}(Q''/Q),$$

$$\phi_{00} = \tfrac{1}{3} H^{-1}[H_t Q'/Q - 3H_{\zeta\bar{\zeta}} Q^{-4/3}] - \tfrac{1}{3}(Q''/Q),$$

$$\phi_{11} = \tfrac{1}{12} H^{-1}[-3H_{tt} + 2H_t Q'/Q] + \tfrac{1}{12}(Q''/Q),$$

$$\phi_{10} = \psi_3, \quad \phi_{20} = \psi_4, \quad \phi_{12} = \psi_1, \quad \phi_{22} = \phi_{00}.$$

[1] For this terminology, see for example, G. F. R. Ellis, in *General Relativity and Cosmology, Proceedings of the International School of Physics Enrico Fermi*, edited by R. K. Sachs (Academic, New York, 1971), pp. 109–14.

[2] H. Stephani, Commun. Math. Phys. **9**, 53 (1968).

[3] P. Szekeres, Commun. Math. Phys. **41**, 55 (1975).

[4] E. T. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).

[5] The integrals appearing in these formulas may be expressed in terms of Gauss hypergeometric functions. We thank Dr. M. L. Glasser for this information.

[6] J. Wainwright, Commun. Math. Phys. **17**, 42 (1970).

[7] J. Wainwright, J. Math. Phys. **18**, 672 (1977).

[8] For this and related terminology we refer the reader to F. A. E. Pirani, in *Lectures on General Relativity*, *Brandeis Summer Institute in Theoretical Physics* (Prentice Hall, Englewood Cliffs, New Jersey, 1964), Vol. 1. See Chapters 3 and 4.

[9] D. W. Trim and J. Wainwright, J. Math. Phys. **15**, 535 (1974).

[10] W. B. Bonnor and N. Tomimura, Mon. Not. R. Astron. Soc. **175**, 85 (1976).

[11] For this terminology see G. F. R. Ellis and M. A. H. MacCallum, Commun. Math. Phys. **12**, 108 (1969).

[12] The dust solutions in this class are contained, for example, in O. Heckmann and E. Schücking, in *Gravitation*: *An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962). Set $\alpha = \pi/6$ in the solution given on p. 444.

[13] G. F. R. Ellis, J. Math. Phys. **8**, 1171 (1967). The solution in question is given on p. 1183, Eq. (4.38). We note that the expression for X should read $X = [1 + C(x^1)x^0]^{-1/3}$.

[14] J. M. Stewart and G. F. R. Ellis, J. Math. Phys. **9**, 1972 (1968).

[15] Bonnor and Tomimura, Ref. 10, studied the asymptotic behavior of the Szekeres solutions, but restricted their considerations to pointwise limits.

[16] We are using geometrized units (i.e., $c = 1$, $G = 1$) so that length, time, and mass can be measured in units of length. It follows that the physical and geometrical scalars considered in this section have units as follows: $\theta \sim L^{-1}$, $\mu \sim L^{-2}$, $p \sim L^{-2}$, $p \sim L^{-2}$, $\psi_2 \sim L^{-2}$, $\sigma_{ab}\sigma^{ab} \sim L^{-2}$, $(-\mu_{,a}\mu_{,b}h^{ab})^{1/2} \sim L^{-3}$. Thus the ratios p/μ, μ/θ^2, $\sigma^{ab}\sigma_{ab}/\theta^2$, ψ_2/θ^2, and $(-\mu_{,a}\mu_{,b}h^{ab})^{1/2}/\theta^3$ are dimensionless.

[17] See Ref. 1, p. 139.

[18] M. A. H. MacCallum, Commun. Math. Phys. **20**, 57 (1971). See pp. 64–5. We note that when $C_2 \neq 0$ the integrated rate of shear tends to zero as $t \to \infty$, since the eigenvalues of σ_{ab} are $-\chi$, $\frac{1}{2}\chi$, $\frac{1}{2}\chi$ and the integral $\int_{t_0}^{\infty} \chi\, dt$ is finite.

[19] Note that although $\psi_2 \to 0$ in both cases, the limiting metric, regarded as a solution of the field equations, satisfies $\psi_2 = 0$ only when $\psi_2/\theta^2 \to 0$. A similar situation holds for the rate of shear σ_{ab} of the fluid.

[20] If the x and y coordinates are restricted to be nonnegative, however, the arbitrary functions can be chosen so that $\mu > 0$ for all possible t. We note that in those Szekeres solutions which approach R–W solutions asymptotically in the pointwise sense [case *PI* with $\beta = 0$ and case *HI* with $\beta = 0$ in Ref. 10] the density is also unbounded both positively and negatively on the spatial sections $t = \text{const}$, no matter how large t is.

[21] See Ref. 18, p. 64, for this terminology.

[22] S. J. Campbell and J. Wainwright, preprint, University of Waterloo, 1976, to appear in Gen. Rel. Grav.

Inhomogeneous cosmologies: New exact solutions and their evolution

D. A. Szafron

Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada

(Received 28 February 1977)

An algorithm is presented for determining all solutions of the Einstein field equations representing a perfect fluid with metric of the form $ds^2 = dt^2 - e^{2\alpha}dz^2 - e^{2\beta}(dx^2 + dy^2)$ and fluid flow vector $\mathbf{u} = \partial/\partial t$. The entire class of solutions is then invariantly characterized. These new solutions generalize Szekeres' inhomogeneous cosmological models containing dust. A subclass of these solutions is studied in detail and it is interesting that some of these models approach isotropy but not homogeneity for large cosmological times.

1. INTRODUCTION

This paper investigates all solutions to the Einstein field equations for a perfect fluid

$$G_{ab} = \kappa T_{ab} = \kappa[(\mu + p)u_a u_b - p g_{ab}] \tag{1.1}$$

for which the metric tensor can be written in the form

$$\begin{aligned} &ds^2 = dt^2 - e^{2\alpha}dz^2 - e^{2\beta}(dx^2 + dy^2), \\ &\qquad \alpha = \alpha(t,x,y,z), \quad \beta = \beta(t,x,y,z), \end{aligned} \tag{1.2}$$

and for which the fluid flow vector is $\mathbf{u} = \partial/\partial t$. The quanquantity μ represents the total (relativistic) energy density measured by an observer moving with the fluid and p is the isotropic pressure. The present work generalizes that of Szekeres[1] who recently found all solutions of this type for the special case of dust ($p = 0$).

In Sec. 2 the solutions are divided into two classes labeled I and II. The field equations in class I are reduced to one second order linear ordinary differential equation. The field equations in class II are reduced to two decoupled second order linear ordinary differential equations. In each class, the spacelike hypersurfaces orthogonal to the fluid flow are in general *not* the orbits of a local group of isometries (i.e., *these models are inhomogeneous*). Some class II solutions have recently been obtained explicitly and studied by Szafron and Wainwright.[2] Finally, an invariant characterization of both classes is presented generalizing that of Wainwright[3] who has considered the dust ($p = 0$) case.

In Sec. 3 some class I solutions are exhibited explicitly and studied in detail. The possible Killing vector fields are listed, the kinematical quantities of the fluid are calculated and the behavior of the solutions is investigated at large cosmological times. Although these solutions are both anisotropic and inhomogeneous, they all approach isotropy for large t; some however, do *not* approach homogeneity in a certain well-defined sense.

2. DERIVATION OF THE SOLUTIONS

Introduce a pair of complex variables ξ and $\bar{\xi}$, defined by

$$\xi = x + iy, \quad \bar{\xi} = x - iy.$$

The field equations then reduce to[4]:

$$G^0_0 - 2G^\xi_\xi - G^1_1 = -2(\ddot{\alpha} + 2\ddot{\beta} + \dot{\alpha}^2 + 2\dot{\beta}^2) = \kappa(\mu + 3p), \tag{2.1}$$

$$\tfrac{1}{2}G^0_1 = -\dot{\beta}' + \beta'(\dot{\alpha} - \dot{\beta}) = 0, \tag{2.2}$$

$$G^1_1 = -4e^{-2\beta}\beta_{\xi\bar{\xi}} - e^{-2\alpha}\beta'^2 + 2\ddot{\beta} + 3\dot{\beta}^2 = -\kappa p, \tag{2.3}$$

$$\begin{aligned} G^\xi_\xi = {}& e^{-2\alpha}(-\beta'' + \alpha'\beta' - \beta'^2) - 2e^{-2\beta}(\alpha_{\xi\bar{\xi}} + \alpha_\xi\alpha_{\bar{\xi}}) + \ddot{\alpha} + \ddot{\beta}, \\ & + \dot{\alpha}^2 + \dot{\beta}^2 + \dot{\alpha}\dot{\beta} = -\kappa p, \end{aligned} \tag{2.4}$$

$$G^0_\xi = -\dot{\alpha}_\xi - \dot{\beta}_\xi + \alpha_\xi(\dot{\beta} - \dot{\alpha}) = 0, \tag{2.5}$$

$$e^{2\alpha}G^1_\xi = \beta'_\xi - \beta'\alpha_\xi = 0, \tag{2.6}$$

$$\tfrac{1}{2}e^{2\beta}G^{\bar{\xi}}_\xi = \alpha_{\xi\xi} + (\alpha_\xi)^2 - 2\beta_\xi\alpha_\xi = 0, \tag{2.7}$$

where $' \equiv \partial/\partial z$, $\cdot \equiv \partial/\partial t$, and $\alpha_\xi \equiv \partial\alpha/\partial\xi \equiv \frac{1}{2}(\partial\alpha/\partial x - i\partial\alpha/\partial y)$, $\alpha_{\bar{\xi}} \equiv \partial\alpha/\partial\bar{\xi} \equiv \frac{1}{2}(\partial\alpha/\partial x + i\partial\alpha/\partial y)$. Notice that Eqs. (2.5), (2.6), and (2.7) are complex, so there are in fact ten real field equations. Since $\mathbf{u} = \partial/\partial t$, the flow lines are geodesics and the contracted Bianchi identities imply that the pressure is a function of t only: $p = p(t)$. However the energy density is in general a function of all four variables: $\mu = \mu(t, z, \xi, \bar{\xi})$, i.e., no equation of state of the form $p = p(\mu)$ is imposed.

Lemma: The field equations (2.1)–(2.7) imply that

$$\dot{\beta}_\xi = 0. \tag{2.8}$$

Proof: See Appendix A.

Note that Szekeres[1] states this lemma for dust ($p = 0$), without giving the details of the proof. The function β is real so that $\dot{\beta}_\xi = 0$ if and only if $\dot{\beta}_{\bar{\xi}} = 0$. Further integration of the field equations depends on whether or not $\beta' = 0$.

A. Class I, $\beta' \neq 0$

Equation (2.8) implies that β may be written in the form

$$\beta = \log\phi(t, z) + \nu(z, \xi, \bar{\xi}). \tag{2.9}$$

Since $\beta' \neq 0$, Eq. (2.2) can now be integrated to obtain

$$\alpha = \log(h(z, \xi, \bar{\xi})\phi' + h(z, \xi\bar{\xi})\phi\nu'). \tag{2.10}$$

Equation (2.6) requires that

$$h = h(z). \tag{2.11}$$

Thus ϕ can be redefined as ϕh and ν as $\nu - \log h$ so that

β remains unchanged and

$$\alpha = \log(\phi' + \phi\nu'). \tag{2.12}$$

Equation (2.5) is now identically satisfied. Equation (2.3) implies that

$$4e^{-2\nu}\nu_{\xi\bar{\xi}} + 1 = 2\ddot{\phi}\phi + \dot{\phi}^2 + \kappa p\phi^2 = -k(z), \tag{2.13}$$

for some arbitrary real function $k(z)$. Szekeres[1] obtained this equation, using the variable $\psi = \log\phi$, for the particular case $p=0$; notice that his subsequent choice of coordinates ξ and $\bar{\xi}$ is also valid in the present context[5] so that

$$e^{-\nu} = A(z)\xi\bar{\xi} + B(z)\xi + \bar{B}(z)\bar{\xi} + C(z), \tag{2.14}$$

where $A(z)$ and $C(z)$ are real functions, and $B(z)$ is complex, with

$$AC - B\bar{B} = \tfrac{1}{4}(1 + k(z)). \tag{2.15}$$

The remaining part of Eq. (2.13) can be rewritten

$$\frac{2\ddot{\phi}}{\phi} + \frac{\dot{\phi}^2}{\phi^2} + \kappa p + \frac{k(z)}{\phi^2} = 0. \tag{2.16}$$

For each fixed value of z, this is just the propagation equation for the length scale in the Robertson–Walker (henceforth abbreviated R–W) solution. Equation (2.7) is now identically satisfied. After a lengthy but straightforward calculation, Eq. (2.4) can also be reduced to an identity. The equations which remain to be solved are Eqs. (2.16) and (2.1), for the unknowns μ, p, and ϕ. This system is indeterminate. An algorithm for generating particular exact solutions is as follows. Specify explicitly $p = p(t)$ and solve Eq. (2.16) for $\phi(t,z)$. The metric is now determined since α and β may be obtained directly from Eqs. (2.9), (2.12), and (2.14). Finally, use Eq. (2.1) as the definition of μ. Naturally, this procedure does not necessarily generate a physically reasonable fluid, satisfying fundamental energy conditions. In Sec. 3 a subclass of physically reasonable solutions is investigated.

B. Class II, $\beta' = 0$

Equation (2.8) allows β to be written

$$\beta = \log\phi(t) + \nu(\xi, \bar{\xi}). \tag{2.17}$$

Equations (2.2) and (2.6) are now identically satisfied and Eq. (2.5) can be integrated to obtain

$$\alpha = \log[\lambda(t,z) + \phi(t)\sigma(z, \xi, \bar{\xi})], \tag{2.18}$$

where λ and σ are arbitrary. Equation (2.3) then becomes

$$4e^{-2\nu}\nu_{\xi\bar{\xi}} = 2\phi\,\ddot{\phi} + \dot{\phi}^2 + \kappa p\phi^2 = -k, \quad k \text{ constant}. \tag{2.19}$$

The functions ϕ and ν may be redefined to make $k = 0$, ±1. Equation (2.19) generalizes to the case $p \neq 0$, a relationship obtained by Szekeres[1] for dust. His subsequent choice of coordinates is ξ and $\bar{\xi}$ so that

$$e^{-\nu} = \tfrac{1}{2}(1 + k\xi\bar{\xi}) \tag{2.20}$$

is also valid here. Then Eq. (2.19) assumes the form

$$\frac{2\ddot{\phi}}{\phi} + \frac{\dot{\phi}^2}{\phi^2} + \kappa p + \frac{k}{\phi^2} = 0. \tag{2.21}$$

From Eq. (2.7) it follows that

$$(e^{-\nu}\sigma)_{\xi\xi} = (e^{-\nu}\sigma)_{\bar{\xi}\bar{\xi}} = 0. \tag{2.22}$$

These can be integrated to obtain

$$\sigma = [U(z)\xi\bar{\xi} + V(z)\xi + \bar{V}(z)\bar{\xi} + W(z)]e^{\nu}, \tag{2.23}$$

where $U(z)$ and $W(z)$ are real functions and $V(z)$ is complex. A long but straightforward calculation, using Eqs. (2.20) and (2.23) reduces Eq. (2.4) to

$$\ddot{\lambda}\phi + \dot{\lambda}\dot{\phi} + \lambda\ddot{\phi} + \lambda\phi\kappa p = U(z) + kW(z). \tag{2.24}$$

The equations which remain to be solved are Eqs. (2.24), (2.21), and (2.1) for the unknowns p, μ, ϕ, and λ. This system is again indeterminate (as in class I). An algorithm for generating exact solutions in class II is as follows. Specify explicitly $p = p(t)$ and solve Eq. (2.21) for $\phi(t,z)$. Then substitute $p(t)$ and $\phi(t,z)$ into Eq. (2.24) and solve for $\lambda(t,z)$. Calculate α and β from Eqs. (2.17) and (2.18). Finally calculate μ from Eq. (2.1). Again the solutions generated do not necessarily represent a physically reasonable fluid. The solutions in this class with $p = bt^{-2}$, $0 < b < \frac{1}{3}$, and $k=0$ have been studied by Szafron and Wainwright.[2]

A form of Eq. (2.16) or (2.21) which is useful for generating exact solutions is obtained by the substitution $\phi = G^{2/3}$. Then the equation becomes

$$\ddot{G} + \tfrac{3}{4}(\kappa p - k)G = 0. \tag{2.25}$$

For example, in the case $k=0$, if $\kappa p = \frac{4}{3}q^2t^{-r}$, $r \neq 2$, the real and imaginary parts of the function $\mathcal{G}$ given below, are solutions[6] of this DE,

$$\mathcal{G} \equiv \begin{cases} t(t^{1-2/(2n+1)}D)^{n+1}H, & \text{for } n \geqslant 0, \\ (t^{1-2/(2n+1)}D)^{-n}H, & \text{for } n < 0, \end{cases} \tag{2.26}$$

with

$$r = \frac{4n}{2n+1},$$

$$D \equiv \frac{d}{dt}, \qquad H \equiv g\cos[(2n+1)qt^{-(2n+1)}] + hi\sin[(2n+1)qt^{-(2n+1)}].$$

In this solution g, h are constants when Eq. (2.25) is regarded as Eq. (2.21) and g, h are arbitrary functions of z when Eq. (2.25) is regarded as Eq. (2.16). Further solutions to Eq. (2.25) may be found in Kamke[7] for various forms of the function $p = p(t)$ and values of k, but these solutions usually involve functions (such as orthogonal polynomials) more complicated than those of Eq. (2.26).

In concluding this section, it should be noted that the characterization of the Szekeres dust solutions given by Wainwright[3] can be extended to the present classes of solutions. The characterization depends on the Weyl tensor of the spacetime and the rate of shear tensor of the fluid.

Theorem: The solutions in classes I and II with nonzero Weyl tensor comprise all solutions of the Einstein field equations (1.1) with the flow vector **u** geodesic and hypersurface orthogonal (irrotational), which satisfy the following properties:

(1) The Weyl tensor is algebraically special of type $\{22\}$, i.e. type D, and the fluid velocity lies in the 2-space spanned by the repeated principal null directions.

(2) Any vector which is orthogonal to both of the repeated principal null directions is an eigenvector of the shear tensor σ_{ab}.

(3) The 2-spaces generated by the repeated principal null directions admit orthogonal 2-surfaces.

Proof: The proof is identical to that of Wainwright,[3] since the only use made of the condition $p=0$ in his proof is to guarantee that the flow is geodisic.

It is of interest to note that the repeated principal null directions are in general not tangent to geodesics, unlike the case of vacuum solutions of type $\{22\}$.

3. SOME SOLUTIONS FROM CLASS I

Let $k(z)=0$ and define p explicitly by

$$\kappa p=\tfrac{4}{3}q(1-q)t^{-2}, \quad \text{where } q=\text{const.} \tag{3.1}$$

Equation (2.16) can now be integrated to yield $\phi(t,z)$. Then Eqs. (2.9), (2.12), and (2.14) give α and β so that the line element can be written

$$ds^2=dt^2-D^2(dx^2+dy^2)-F^2D_z^2dz^2,$$

where

$$\begin{aligned} &D(t,x,y,z)=G^{2/3}(t,z)/F(x,y,z),\\ &G(t,z)=g(z)t^{1-q}+h(z)t^{q},\\ &F(x,y,z)=a(z)(x^2+y^2)+b(z)x+d(z)\,y+c(z),\end{aligned} \tag{3.2}$$

with $a(z)$, $b(z)$, $c(z)$, $d(z)$, $g(z)$, and $h(z)$ are arbitrary functions of z subject to

$$ac-b^2-d^2=\tfrac{1}{4}.$$

Equation (2.1) implies that

$$\kappa\mu=\frac{4G_t(2G_{tz}F-3G_tF_z)}{3G(2G_zF-3GF_z)}, \quad G_t\equiv\frac{\partial G}{\partial t}, \quad G_z\equiv\frac{\partial G}{\partial z}. \tag{3.3}$$

The parameter q can assume all real values; however the solutions with $q\leqslant\frac{1}{2}$ are equivalent to the solutions with $q\geqslant\frac{1}{2}$, since the equations which define solution (3.2) are invariant under the substitutions $q\rightarrow 1-q$, $g(z)\rightarrow h(z)$. Henceforth, without loss of generality, q is restricted to the range $q\geqslant\frac{1}{2}$. In addition, the requirement $q\leqslant 1$ is imposed, to ensure that the pressure is nonnegative. The general dust solution is given by $q=1$ and is contained in Szekeres.[1]

The properties of this solution can be conveniently studied using the Newman–Penrose[8] formalism. Introduce a null tetrad for the line element (3.2) defined by

$$\begin{aligned} &\mathbf{k}=2^{-1/2}(dt-FD_zdz),\\ &\mathbf{n}=2^{-1/2}(dt+FD_zdz),\\ &\mathbf{m}=-2^{-1/2}Dd\bar{\xi}.\end{aligned} \tag{3.4}$$

Notice that the fluid velocity vector can be written

$$u^a=2^{-1/2}(k^a+n^a). \tag{3.5}$$

Consider first the kinematic quantities of the fluid. It is clear that both the acceleration $\dot{\mathbf{u}}$ and the vorticity vector ω are zero. The rate of shear scalar σ and the volume expansion scalar θ, can be calculated using the formulas given by Wainwright[3] and the expressions for the spin coefficients of the null tetrad (3.4) given in Appendix B of the present paper. One finds that

$$\sigma=\frac{2F(1-2q)(gh_z-hg_z)}{\sqrt{3}\,G(2G_zF-3GF_z)} \tag{3.6}$$

and

$$\theta=\frac{2G_t}{G}-\frac{\sqrt{3}\,\sigma}{\sqrt{2}}. \tag{3.7}$$

It thus follows that the fluid has nonzero shear if $gh_z-g_zh\neq 0$ and $q\neq\frac{1}{2}$.

The geometry of the solutions is considered next. It can be shown that the Riemann tensor of each $\{t=\text{const}\}$ hypersurface is identically zero. The behavior of the Weyl tensor can be studied by calculating its tetrad components[9] ψ_0, ψ_1, ψ_2, ψ_3, and ψ_4. One finds that

$$\psi_0=\psi_1=0, \quad \psi_3=\psi_4=0,$$

and

$$\psi_2=\frac{4(1-2q)FG_t(hg_z-gh_z)}{9G^2(2G_zF-2GF_z)}. \tag{3.8}$$

Thus all the solutions are type $\{22\}$. They are conformally flat if and only if

$$q=\tfrac{1}{2} \quad \text{or} \quad hg_z-gh_z=0. \tag{3.9}$$

In the conformally flat case the fluid acceleration and vorticity are zero [and the shear is necessarily zero by (3.6)] and since the spatial sections $\{t=\text{const}\}$ are flat it follows that these solutions belong to the $k=0$, R–W class of solutions.[10] The spatial coordinates are however, not the standard ones. The expressions for the pressure (3.1) and the density (3.3) reduce to

$$\kappa\mu=\frac{4G_t^2}{3G^2} \quad \text{and} \quad \kappa p=\frac{4q(1-q)t^{-2}}{3}. \tag{3.10}$$

In the special case $g=0$, it follows that $p=(q^{-1}-1)\mu$ so this solution contains the $(k=0)$ R–W solutions with barotropic equation of state $p=(\gamma-1)\mu$. The restriction $\frac{1}{2}\leqslant q\leqslant 1$ restricts γ to satisfy the inequalities $1\leqslant\gamma\leqslant 2$.

If Eq. (3.9) does not hold, the solutions are not conformally flat and they are spatially inhomogeneous. Then all Killing vector fields may be written in the form $X(\partial/\partial x)-Y(\partial/\partial y)$ where X and Y are given by Table I.

Remarks: The three Killing vector fields in the first case are those of spherical symmetry in a nonstandard coordinate system. The case $\lambda=\epsilon=0$ and $\delta\neq 0$ is not listed in the table since the substitution $x\rightarrow y$, $b\rightarrow d$ makes it identical to the second case $\lambda=\delta=0$ and $\epsilon\neq 0$.

Solutions are sought which are not R–W (i.e., $\psi_2\neq 0$) but which approach a R–W solution for large cosmological time t. Since the behavior of the dust solutions $(q=1)$ has already been studied by Bonnor and Tominura,[11] q is henceforth restricted by the inequality

$$\tfrac{1}{2}<q<1. \tag{3.11}$$

The coordinates are permitted to assume values

TABLE I. Killing vector fields admitted by the metric (3.2) when it is not conformally flat. Define $\delta = b_z a - a_z b$, $\epsilon = a_z d - d_z a$, and $\lambda = ac_z - a_z c$. Let r denote the dimension of the isometry group.

Case	r	X	Y
$\delta = \epsilon = \lambda = 0$		$y + (d/2a)$	$x + (b/2a)$
a, b, c, d constants	3	$y^2 - x^2 - (b/a)x - (c/a)$	$2xy + (b/a)y$
		$2xy + (d/a)x$	$x^2 - y^2 - (d/a)y - (c/a)$
$\lambda = \delta = 0,\ \epsilon \neq 0$	1	$y^2 - x^2 - (b/a)x - (c/a)$	$2xy + (b/a)y$
$\lambda = 0,\ \delta\epsilon \neq 0$	1	$y^2 - x^2 + [(bd_z - db_z)/\beta]x$	$(\alpha/\beta)(y^2 - x^2) + [(b_z d - bd_z)/\beta]y$
		$-2(\alpha/\beta)xy - (c/a)$	$+ 2xy + (c\alpha/a\beta)$
$\lambda \neq 0$	1	$\frac{1}{2\lambda}\{(y^2 - x^2)\beta + 2\lambda y + (db_z - bd_z)$	$\frac{1}{2\lambda}\{(x^2 - y^2)\alpha + 2\lambda x - (db_z - bd_z) - 2\beta xy$
		$+ 2\alpha xy + (dc_z - cd_z)\}$	$+ (bc_z - b_z c)\}$

$$t_0 < t < \infty, \quad -\infty < x, y, z < \infty. \tag{3.12}$$

Two types of limits are considered,[12] *pointwise limits* L, where $\lim_{t\to\infty} F = L$ and *uniform limits* where $\lim_{t\to\infty}\{\mathrm{lub}_{t=\mathrm{const}} |F(x, y, z, t) - L|\} = 0$.

First consider pointwise limits. It follows from Eq. (3.3) that

$$\kappa\mu \approx 4qt^{-2/3} \tag{3.13}$$

as $t \to \infty$, for fixed x, y, z. Thus in the pointwise sense

$$\lim_{t\to\infty} (p|\mu) \approx \gamma - 1, \quad \gamma = q^{-1}. \tag{3.14}$$

It follows from Eqs. (3.6) and (3.7) that

$$\theta \approx 2qt^{-1} \tag{3.15}$$

and

$$\sigma \approx \frac{2F(1 - 2q)(gh_z - hg_z)t^{-2q}}{\sqrt{3}\, h(2h_z F - 3hF_z)}. \tag{3.16}$$

Equations (3.6) and (3.8) imply that

$$\psi_2 = -2\sqrt{3}\, \sigma G_t / G. \tag{3.17}$$

Now the following three limits may be calculated using Eqs. (3.13), (3.16), (3.17), and (3.15):

$$\lim_{t\to\infty} \frac{\kappa\mu}{\theta^2} = \frac{1}{3}, \quad \lim_{t\to\infty} \frac{\sigma}{\theta} = \lim_{t\to\infty} \frac{\psi_2}{\theta^2} = 0. \tag{3.18}$$

These limits suggest that the models approach isotropy along each fluid line as $t \to \infty$. The metric itself approaches the isotropic form

$$ds^2 = dt^2 - t^{4q/3} h^{4/3} F^{-2}[dx^2 + dy^2 + \tfrac{1}{9}\{2Fh_z - 3hF_z\}^2 h^{-2} dz^2]. \tag{3.19}$$

Although these limits relate to the isotropy of the solutions, the homogeneity can be studied[13] by calculating $\lim_{t\to\infty}(-\kappa^2 \mu_{,a} \mu_{,b} h^{ab})^{1/2}/\theta^3$. In fact the calculation yields

$$\lim_{t\to\infty}(-\kappa^2\mu_{,a}\mu_{,b}h^{ab})^{1/2}/\theta^3 = \begin{cases} 0, & \text{if } \frac{3}{4} < q < 1, \\ f(x,y,z) & \text{if } q = \frac{3}{4}, \\ \pm\infty & \text{if } \frac{1}{2} < q < \frac{3}{4}, \end{cases} \tag{3.20}$$

where $f(x, y, z)$ is determined by $g(z)$, $h(z)$, and $F(x, y, z)$. Thus solutions with $\frac{3}{4} < q < 1$ approach isotropy and homogeneity in a pointwise fashion. It is interesting to observe that solutions with $\frac{1}{2} < q \leq \frac{3}{4}$ approach isotropy *but not homogeneity*.

The limits calculated so far have been pointwise. These limits are not uniform however, since $|\psi_2|$, $|\theta|$, $|\sigma|$, and $|\mu|$ are unbounded on each spatial section.

There are models however in which these limits are uniform. One such model is given by $F_z = 0$ and F, G, G_z are bounded functions of class C^1 with

$$F(x, y) \geq k > 0, \quad g(z) \geq l > 0, \quad \text{and} \quad g_z(z) \geq m > 0 \tag{3.21}$$

for some constants k, l, and m, for all x, y, z. Notice that such a k can be found. Equation (3.2) implies that

$$F(x,y) = a[(x + b/2a)^2 + (y + d/2a)^2] + 3c/4 + 1/16a, \tag{3.22}$$

so let $a, c > 0$. Then all the previous pointwise limits (3.18) and (3.20) hold uniformly. Local equations of state of the form $\epsilon = \epsilon(p, \rho)$ exist for this special case, where the specific energy density ϵ is defined by $\mu = \rho(1 + \epsilon)$ and ρ is the particle rest mass density. This is true since ρ can be determined from[14]

$$\dot{\rho} + \rho\theta = 0. \tag{3.23}$$

But $\theta = \theta(t, z)$ implies ρ can be chosen as $\rho = \rho(t, z)$, which since $\rho_z \neq 0$ can be solved for z to yield $z = z(\rho, t)$. Also $p = p(t)$ can be inverted since p is not constant. Thus μ, which is a function of t and z, can be expressed in terms of p and ρ. Finally ϵ, which is a function of μ and p, can be expressed as $\epsilon = \epsilon(p, \rho)$.

The behavior of the general solutions (3.2) can also be studied at $t \to 0$; in fact

$$\lim_{t\to 0} \frac{\sigma}{\theta} = \lim_{t\to 0} \frac{\psi_2}{\theta^2} = 0, \quad \lim_{t\to 0} \frac{\kappa\mu}{\theta^2} = \frac{1}{3},$$

$$\lim_{t\to 0} \frac{p}{\mu} = \gamma - 1, \quad \gamma = 1/(1 - q),$$

$$ds^2 \approx dt^2 - t^{4(1-q)/3} g^{4/3} F^{-2}[dx^2 + dy^2 + \tfrac{1}{9}\{2Fg_z - 3gF_z\}^2 g^{-2} dz^2],$$

$$\lim_{t\to 0} \frac{(-\kappa^2 \mu_{,a} \mu_{,b} h^{ab})^{1/2}}{\theta^3} = 0.$$

Thus near the singularity the solutions are isotropic and homogeneous but with an unphysical equation of state, since $\frac{1}{2} < q < 1$ implies $2 < \gamma < \infty$. Again for uniform limits it is sufficient to demand (3.22).

ACKNOWLEDGMENTS

I would like to thank Dr. C. B. Collins and Dr. R. G. McLenaghan for comments and suggestions. Special thanks are due to Dr. J. Wainwright for his help with the computer programs used in this paper and for several useful discussions. This work was supported by a postgraduate scholarship awarded by the National Research Council of Canada.

APPENDIX A

Lemma: The field equations (2.1)–(2.7) imply that $\dot{\beta}_\xi = 0$.

Proof: Assume that $\dot{\beta}_\xi \neq 0$. Equations (2.2), (2.5)–(2.7) have zeros on the right-hand side so they are identical to the dust ($p = 0$) case. In fact Szekeres[15] has shown that these equations, together with the assumption $\dot{\beta}_\xi \neq 0$, imply that coordinates t, x, y, z exist with

$$\alpha = \alpha(t, x), \quad \beta = \beta(t, x). \tag{A1}$$

Equation (2.7) now becomes

$$\alpha_{xx} + \alpha_x^2 - \beta_x \alpha_x = 0, \tag{A2}$$

which can be integrated to give

$$\beta = \tfrac{1}{2}[\alpha + \ln(K\alpha_x)], \quad K = K(t), \quad K \neq 0. \tag{A3}$$

Define a new variable $\epsilon(t, x)$ by

$$\alpha = \ln(K\epsilon). \tag{A4}$$

Substitution of (A4) into (A3) yields

$$\beta = \tfrac{1}{2}[\ln(K\epsilon) + \ln(\epsilon^{-1}K\epsilon_x)]. \tag{A5}$$

Subsitution of (A4) and (A5) into Eq. (2.5) results in

$$[\epsilon\{\ln(\epsilon_x)\}^{\bullet}]_x = 0 \tag{A6}$$

which can be integrated to read

$$\dot{\epsilon} = L\ln(M\epsilon), \quad L = L(t), \quad M = M(t), \quad M \neq 0. \tag{A7}$$

Notice that $\dot{\beta}_\xi \neq 0$ implies $\dot{\beta}_x \neq 0$, and consequently

$$L \neq 0. \tag{A8}$$

Equation (2.4) can be used with the aid of Eqs. (A4), (A5), and (A7), to calculate ϵ_x. In fact Eq. (2.4) becomes

$$\epsilon_{xx} = \epsilon_x\{\text{terms involving } \epsilon, K, L, M, \dot{K}, \dot{L}, \dot{M}\}. \tag{A9}$$

Perform a coordinate transformation $x' = x'(x, t)$, $t' = = t'(x, t)$ defined by

$$x' = \epsilon(x, t), \quad t' = t. \tag{A10}$$

The chain rule then yields

$$\epsilon_{xx} = \frac{\partial\epsilon_x}{\partial\epsilon}\epsilon_x, \tag{A11}$$

so (A9) becomes

$$\frac{d\epsilon_x}{d\epsilon} = \{\text{terms involving } \epsilon, K, L, M, \dot{K}, \dot{L}, \dot{M}\}. \tag{A12}$$

This can be integrated to yield ϵ_x. The resulting expression for ϵ_x can now be differentiated with respect to t and set equal to the x derivative of $\dot{\epsilon}$ obtained from (A7). The resulting integrability condition gives $L = 0$ which violates Eq. (A8). Thus the assumption $\dot{\beta}_\xi \neq 0$ leads to a contradiction. It therefore follows that

$$\dot{\beta}_\xi = 0. \tag{A13}$$

APPENDIX B

The Newman–Penrose spin coefficients relative to the null tetrad (3.4) in the $\xi, \bar{\xi}$ coordinates are

$$\kappa = \bar{\pi} = -\bar{\nu} = -\tau = 2^{-1/2}[F_{\bar{\xi}}F^{-1}D^{-1} + D_{z\bar{\xi}}D^{-1}D_z^{-1}],$$

$$\gamma = -\epsilon = -2^{-3/2}D_{tz}D^{-1},$$

$$\beta = -\alpha = 2^{-1/2}D_{\bar{\xi}}D^{-2},$$

$$\rho = -2^{-1/2}[D_tD^{-1} + F^{-1}D^{-1}],$$

$$\mu = 2^{1/2}[D_tD^{-1} - F^{-1}D^{-1}],$$

$$\sigma = \lambda = 0.$$

These expressions were obtained using a Camal computer program (see Campbell and Wainwright[16]).

[1] P. Szekeres, Commun. Math. Phys. **41**, 55 (1975). The coordinates (x, y, z) in the present paper replace the coordinates (y, z, r) of Szekeres.

[2] D. A. Szafron and J. Wainwright, "A Class of Inhomogeneous Perfect Fluid Cosmologies," J. Math. Phys. **18**, 1668 (1977).

[3] J. Wainwright, J. Math. Phys. **18**, 672 (1977).

[4] These equations were checked using the Camel computer programs of S. J. Campbell, M. Math. Thesis, University of Waterloo (1976).

[5] The metric $e^{2\nu}d\xi d\bar{\xi}$ is a surface of constant curvature $1 + k(z)$. See Ref. 1, Appendix.

[6] E. Kamke, *Differentialgleichungen Losungsmethoden und Losungen* (Akademische Verlagsgesellschaft Geest Portig K.-H., Leipzig, 1961), Vol. 1, pp. 401–2.

[7] See Ref. 6.

[8] E. T. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).

[9] See Ref. 2.

[10] See, for example, G. F. R. Ellis, "General Relativity and Cosmology," *Proceedings of the International School of Physics Enrico Fermi*, *Course XLVII*, edited by R. K. Sachs (Academic, New York, 1971).

[11] W. B. Bonnor and N. Tomimura, Mon. Not. R. Astron. Soc. **175**, 85 (1976).

[12] For definitions of these limits see Ref. 2.

[13] See Ref. 2.

[14] See Ref. 10, page 115.

[15] See Ref. 1, page 62.

[16] S. J. Campbell and J. Wainwright, "Symbolic Computation and the Newman–Penrose Formalism," preprint, University of Waterloo (1976).

The decomposition of the tensor product of representations of the symmetric group

Susan Schindler

Department of Mathematics, Baruch College, City University of New York, New York 10010

R. Mirman

155 East 34th Street, New York, New York 10016

(Received 12 July 1976)

The decomposition of the tensor product of two irreducible representations of the symmetric group is studied, providing a foundation for the calculation of the Clebsch–Gordon coefficients. Realizations on spaces of polynomials are emphasized. This leads to tensor coupling coefficients. An iterative formula for the calculation of the tensor coupling coefficients is derived, and symmetry constructions are discussed and shown to lead to a reduction in the calculation needed. A well-defined procedure for the construction of the Clebsch–Gordan coefficients from the tensor coupling coefficients, which works for any multiplicity, is obtained. It may be used for any finite group and, with some modifications, for any compact group.

I. INTRODUCTION

One of the major topics in the theory of representations of a group is the reduction into a direct sum of the tensor product of irreducible representations. Important in itself, it is fundamental in physical and mathematical applications. It involves, basically, the determination of two classes of numbers. The first is the number of times each irreducible representation appears as a summand in the direct sum (the multiplicities). The second is the coefficient of each basis vector in the direct sum (the Clebsch—Gordan coefficients).

Here we consider these latter numbers and related coefficients for the symmetric group, S_n. The significance of this group and its Clebsch—Gordan decomposition and coefficients, in the quantum mechanical theory of systems of several particles, with their intrinsic symmetry, is well known. In addition, much of the theory of representations of Lie groups is built on that of the symmetric groups. In particular, the decomposition of the tensor product of representations of S_n appears to be closely related to the problem of determining noncanonical decompositions of representations of the unitary groups. These groups, of course, have physical significance and they are also important in the theory of special functions.

It is also likely that the procedure used in decomposing the symmetric group can be useful in finding the Clebsch—Gordan decompositions of other groups.

This work has three aspects: a review, derivations, and tables. The tables will be given in a later paper.[1]

For easier accessibility, and to establish notation, we present a review of the symmetric group and its representations. Thus, we cover, for the most part in summary form, material discussed in more depth elsewhere. We try to provide references which are well enough defined so the interested reader can easily find more detailed treatments. The most important of these are in Rutherford,[2] Boerner,[3] and Hamermesh.[4] The last is the only other discussion of which we know on the subject of this paper. A large part of this paper is an expansion of the material found there.

Other important references on the symmetric group are Weyl,[5,6] Robinson,[7] and Littlewood.[8]

There are cases where we rederive well known results because it is useful to show how they are found in our notation. Where the discussion is self-contained, standard, and easily found, we do not feel it is necessary to give a specific reference.

The computations are based on an iterative formula which we will derive. It is this which makes the calculations feasible.

We also discuss the meaning of the Clebsch—Gordan decomposition for tensors and introduce the concept of tensor coupling coefficients, whose relationship to the Clebsch—Gordan coefficients we study. We expect that the values of the tensor coupling coefficients, which can be obtained from our formulas and tables, will be needed in some of the applications of this work suggested above.

The tensor coupling coefficients appear in the reduction, into an appropriate direct sum, of the tensor product, with itself, of the left regular representation of S_n. The Clebsch—Gordan coefficients give a reduction, into a direct sum, of the tensor product of irreducible components of this representation.

The tensor coupling coefficients arise from our explicit use of polynomial spaces as carrier spaces of representations.

The question of multiplicity adds a major complication to the determination of the Clebsch—Gordan coefficients. Methods for computing the coefficients in some cases with multiplicity greater than one have been discussed elsewhere. However, these are heuristic in nature and there is no guarantee they will always work. We will describe, systematically and rigorously, a detailed procedure for finding the coefficients in general; it will work for any finite group. We expect that, with certain modifications, the procedure works for any compact group as well.

There have recently been several papers discussing the Clebsch—Gordan decompositions of various finite groups.[9] The approaches used in these, however, are different from ours. We have tried to keep this paper self-contained.

In the next section, we review the symmetric group and its group ring, which is used to give the representations. The following section discusses two classes of representations, the seminormal and orthogonal representations, stating explicitly the rules giving their matrix elements. Then we discuss the decomposition of the tensor product of vectors and the corresponding tensor coupling coefficients. Section V consists of the definition of the Clebsch—Gordan coefficients, and Sec. VI is concerned with the relations between these two sets of coefficients. The tensor coupling coefficients are calculated by means of an iterative formula which is presented and derived in Sec. VII. Symmetries for the two sets of coefficients are investigated in Sec. VIII. The coefficients also obey certain other relations, given in Sec. IX. To illustrate our methods and results, we present several examples in Sec. X. Finally, Sec. XI contains some concluding remarks.

II. THE SYMMETRIC GROUP AND ITS GROUP RING

The symmetric group on n objects, S_n, is the set of all permutations of $\{1,2,\ldots,n\}$. We define the product of two such permutations σ and τ as

$$(\sigma\tau)(j)=\sigma(\tau j),\quad j=1,2,\ldots,n, \tag{II.1}$$

that is we first apply τ and to the result we apply σ. The order of S_n is $n!$

A transposition is a permutation which interchanges two numbers and leaves the others fixed. Every transposition is its own inverse. A neighboring transposition, τ_j, is one interchanging $j-1$ and j $(j=2,\ldots,n)$. Any permutation can be written as a product of transpositions (and, in fact, of neighboring transpositions). The sign of a permutation is $+1$ if it is the product of an even number of transpositions, -1 otherwise. (The sign is well defined, even though the transpositions occurring in the product and the number of such transpositions are not unique.)

To find the irreducible representations of S_n, we will use its group ring, O_n. It consists of all linear combinations of permutations which we take as acting on products of noncommuting variables $x_1, x_2, \ldots, x_n$. That is, $h\in O_n$ if

$$h=\sum_{\tau\in S_n}\xi(\tau)\tau, \tag{II.2}$$

where the $\xi(\tau)$'s are scalars and $\tau(x_1x_2\cdots x_n)=x_{\tau(1)}x_{\tau(2)}\cdots x_{\tau(n)}$. (We use lower case italic letters for elements of O_n, Greek for those of S_n.) Note that S_n can be considered as a subset of O_n. O_n is a vector space and the elements of S_n form a basis for it. O_n is the smallest algebra which contains S_n.

We note that $\sigma\in S_n$, $h\in O_n$ imply $\sigma h\in O_n$, i.e., O_n is closed under left multiplication by elements of S_n. Thus, O_n is the space of a representation of S_n, given by left multiplication. It is the left regular representation of S_n. The irreducible components of this representation can be used to decompose any finite dimensional representation of S_n: Any finite dimensional representation of S_n is equivalent to a direct sum of irreducible components of the left regular representation.[10]

The remarks of the last paragraph are true for the group ring of any finite group.

We must now determine the irreducible components of the left regular representation of S_n; again, the method applies to any finite group. We will outline the construction; the details and proofs can be found in Boerner.[10]

We first construct the minimal ideals of O_n. A subset, J, of O_n is a left (right) ideal if it is a ring and $h'\in O_n$, $h\in J$ imply $h'h\in J$ $(hh'\in J)$. We note that ideals of O_n are also algebras because O_n contains scalars since it contains the identity permutation. A two-sided ideal is a subring which is simultaneously a left and right ideal. A minimal left (right) ideal is a left (right) ideal which properly contains no left (right) ideal. Clearly, a subspace of O_n is an invariant subspace of the left regular representation if and only if it is a minimal left ideal.

Each minimal left ideal J of O_n is of the form

$$J=\{he: h\in O_n\}=O_ne,$$

where $e\in O_n$ is an idempotent, i.e., $e^2=e$, and is primitive in the sense that no equation $e=e'+e''$ (for e', e'' nonzero idempotents) holds.

The set of minimal left ideals can be partitioned into classes, such that the members of each fixed class yield equivalent representations. The number of distinct ideals in an equivalence class equals the dimension of any of them. The direct sum of all the members of a particular equivalence class is a two-sided ideal which is simple, in the sense that it properly contains no nontrivial two-sided ideal. We let α denote a particular equivalence class and J^α the resulting two-sided ideal and we write

$$J^\alpha=O_ne_1^\alpha\oplus O_ne_2^\alpha\oplus\cdots\oplus O_ne_k^\alpha, \tag{II.3}$$

where each $O_ne_j^\alpha$ is in the equivalence class α, and $O_ne_j^\alpha$ is generated by the primitive idempotent e_j^α.

The group ring is the direct sum, over all equivalence classes, of such two-sided ideals. That is,

$$O_n=J^{\alpha_1}\oplus J^{\alpha_2}\oplus\cdots\oplus J^{\alpha_l}, \tag{II.4}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_l$ are the equivalence classes.

We note that the equivalence classes $\alpha_1, \alpha_2, \ldots, \alpha_l$ are unique but that the primitive idempotents e_s^α are not. They depend on the choice of decomposition (II.3). (However, for the seminormal and the orthogonal representations, we do get the same primitive idempotents.)

We next consider the basis vectors. These are three index objects, one giving the equivalence class, one the representation, and one the vector. One basis vector for the representation $O_ne_s^\alpha$ is the primitive idempotent e_s^α. To construct the other basis vectors we start with the following facts: Any equivalence mapping between two minimal left ideals is given by right multiplication of the elements of one of them by some $eh'e'\neq 0$, where e and e' are the primitive idempotents generating, respectively, the former ideal and its image. (We use *ehe'*, rather than *he'*, which gives the same mapping; this choice facilitates subsequent arguments.) Also, any *ehe*, *e* a primitive idempotent, is a scalar multiple of *e*.

We consider the minimal left ideals $O_n e_1^\alpha$ and $O_n e_s^\alpha$ in the class α. Since they give equivalent representations, there is an element $e_{1s}^\alpha \equiv e_1^\alpha h_s' e_s^\alpha \neq 0$. That is, right multiplication of $O_n e_1^\alpha$ by e_{1s}^α gives an equivalence mapping onto $O_n e_s^\alpha$. The inverse mapping is also given by right multiplication, by another element $e_{s1}^\alpha \equiv e_s^\alpha h_s'' e_1^\alpha \neq 0$. Thus, $e_1^\alpha = e_1^\alpha e_{1s}^\alpha e_{s1}^\alpha$. However, we also have $e_{1s}^\alpha e_{s1}^\alpha = e_1^\alpha (h_s' e_s^\alpha e_s^\alpha h_s'') e_1^\alpha = \eta e_1^\alpha$, since we have an expression of the form *ehe*. We obtain $e_1^\alpha = e_1^\alpha e_{1s}^\alpha e_{s1}^\alpha = \eta e_1$ and so $\eta = 1$ and $e_{1s}^\alpha e_{s1}^\alpha = e_1^\alpha$. Moreover, $e_{s1}^\alpha e_{1s}^\alpha = e_s^\alpha$, $e_{1s}^\alpha = (e_1^\alpha h_s') e_s^\alpha \in O_n e_s^\alpha$, and $e_{s1}^\alpha \in O_n e_1^\alpha$. Let $e_{rs}^\alpha \equiv e_{r1}^\alpha e_{1s}^\alpha \in O_n e_s^\alpha$. Note that $e_{ss}^\alpha = e_s^\alpha$.

To show that, for fixed α and s, the vectors e_{rs}^α are independent, we use the identity[11]

$$e_k^\alpha e_{k'}^\alpha = \delta_{kk'} e_k^\alpha. \tag{II. 5}$$

Then if $\sum_r \eta(r) e_{rs}^\alpha = 0$, it follows that

$$\begin{aligned} 0 &= e_{r_0}^\alpha \sum_r \eta(r) e_{rs}^\alpha \\ &= \sum_r \eta(r) e_{r_0}^\alpha e_{r1}^\alpha e_{1s}^\alpha \\ &= \sum_r \eta(r) (e_{r_0}^\alpha e_r^\alpha) h_r'' e_1^\alpha e_{1s}^\alpha \\ &= \eta(r_0) h_r'' e_1^\alpha e_{1s}^\alpha, \end{aligned}$$

and so $\eta(r_0) = 0$ for each r_0. Thus $\{e_{rs}^\alpha\}$ is linearly independent. Since[12] the set $\{e_{rs}^\alpha\}$ spans $O_n e_s^\alpha$, $\{e_{rs}^\alpha\}$ is a basis for $O_n e_s^\alpha$.

We record the following properties:

$$e_{ss}^\alpha = e_s^\alpha, \tag{II. 6}$$

$$e_{rs}^\alpha e_{r's'}^{\alpha'} = \delta^{\alpha\alpha'} \delta_{sr'} e_{rs'}^\alpha, \tag{II. 7}$$

$$O_n = \sum_{\alpha, s} \oplus O_n e_{ss}^\alpha, \tag{II. 8}$$

$$\epsilon = \sum_{\alpha, s} e_{ss}^\alpha, \tag{II. 9}$$

where ϵ is the identity element of S_n. That the product on the left-hand side of (II. 7) is zero when $\alpha \neq \alpha'$, follows from the fact that J^α and $J^{\alpha'}$ are simple two-sided ideals.

Since we have a basis for each summand in (II. 8), the set of e_{rs}^α for all α, r, and s, forms a basis of O_n. The basis vectors e_{rs}^α, including the primitive idempotents $e_{ss}^\alpha = e_s^\alpha$, are not unique. However, the sum

$$\sum_s e_{ss}^\alpha = \epsilon^\alpha$$

is unique, and

$$e_{rs}^\alpha \epsilon^\alpha = \epsilon^\alpha e_{rs}^\alpha = e_{rs}^\alpha, \tag{II. 10}$$

for each α, r, and s. (The latter equalities simply state that ϵ^α is the identity element of the subring J^α.)

Since the set of all e_{rs}^α forms a basis for O_n, we can write

$$\sigma = \sum_{\alpha, r, s} u_{rs}^\alpha(\sigma) e_{rs}^\alpha \tag{II. 11}$$

for any $\sigma \in S_n$. By (II. 7),

$$\sigma e_{ks_0}^{\alpha_0} = \sum_{\alpha, r, s} u_{rs}^\alpha(\sigma) e_{rs}^\alpha e_{ks_0}^{\alpha_0} = \sum_r u_{rk}^{\alpha_0}(\sigma) e_{rs_0}^{\alpha_0}, \tag{II. 12}$$

so the u's are the matrices for the left regular representation. The index s_0 does not occur as a label for any entry of the matrix for σ; the matrices for minimal left ideals corresponding to the same class α are not only equivalent, they are equal.

Let us summarize the construction. To determine a representation we need objects on which the action of the group is known. The group elements acting on these objects give linear combinations of them, in which the numerical coefficients are the matrix elements. The group, of course, acts on itself by multiplication. However, the group is not a vector space, thus, it cannot be the carrier space of a representation. To construct the necessary vector space we introduce the operations of addition and scalar multiplication on the group elements, and obtain the group ring. This space is the carrier space of the left regular representation.

The irreducible components of this representation are given by the minimal left ideals of the group ring; each of these is generated by a primitive idempotent. The product of primitive idempotents giving different minimal left ideals is zero. Each such ring element gives one basis element of the corresponding minimal left ideal.

The totality of basis elements corresponding to all the minimal left ideals of the group ring is a basis of the group ring. Each group element can be expanded, uniquely, in terms of the basis, and the entries of its representation matrices are given by coefficients in the expansion.

The actual construction of the basis is the content of the next section.

III. REPRESENTATIONS OF S_n

The basis vectors e_{rs}^α are labeled by the equivalence class α, the particular representation, s, and the basis vector, r. The standard notation we use specifies the equivalence class by a frame, and the representation and the basis vectors by standard tableaux.

By a frame we mean n boxes, arranged in rows of nonincreasing lengths. We label a frame, α, by the sequence of its row lengths, m_j,

$$\alpha = (m_1, m_2, \ldots, m_n), \tag{III. 1}$$

where

$$m_1 \geq m_2 \geq \cdots \geq m_n \geq 0, \quad \sum_j m_j = n. \tag{III. 2}$$

If each of the numbers $1, 2, \ldots, n$ is inserted in a different box of the frame, so that the numbers in each row and in each column form an increasing sequence, the result is called a standard tableau. We label a standard tableau, r, by

$$r = (j_1, j_2, \ldots, j_n), \tag{III. 3}$$

where j_k is that row of the tableau r containing the number k. Let f^α be the number of standard tableaux corresponding to the frame α. Later we will see that

f^{α} is the dimension of any representation of the equivalence class labeled by the frame α.

Associated with every frame there is another, its conjugate. The frame, α, is labeled by the sequence of numbers $(m_1(\alpha),\ldots,m_n(\alpha))$ giving the lengths of its rows, or by its column lengths $(m_1'(\alpha),\ldots,m_n'(\alpha))$. Conjugation interchanges rows and columns. Using a bar for the conjugate, we have

$$m_i(\bar{\alpha}) = m_i'(\alpha) \qquad \text{(III. 4)}$$

and

$$m_i'(\bar{\alpha}) = m_i(\alpha)。 \qquad \text{(III. 5)}$$

Clearly each frame has a conjugate; a frame may be its own conjugate.

For each standard tableau corresponding to the frame α, there is a conjugate tableau going with the conjugate frame. Thus, if r is a standard tableau, we let $j_k'(r)$ denote the column of r in which the number k lies. We define $\bar{r}$, the conjugate of r, by

$$j_k(\bar{r}) = j_k'(r)。 \qquad \text{(III. 6)}$$

It follows that

$$j_k'(\bar{r}) = j_k(r). \qquad \text{(III. 7)}$$

To show there is a one-to-one correspondence between standard tableaux and their conjugates, we must show that

$$j_k(\bar{r}) \geq j_i(\bar{r})$$

and

$$j_k'(\bar{r}) \geq j_i'(\bar{r})$$

whenever $k \geq i$. But these follow immediately from (III. 6) and (III. 7) and the fact that the numbers in any row or column of r are nondecreasing.

There are three standard constructions for the states e_{rs}^{α}, they give the seminormal, orthogonal, and natural representations.

We will discuss only the first two, and are now ready to turn to their construction.

Associated with a standard tableau of frame α are three operators. Let P_r be the sum of all permutations which preserve the sets of numbers in each row of r. (Such a permutation, σ, must satisfy $j_{\sigma(k)} = j_k$ for $k = 1, 2, \ldots, n$.) A second operator Q_r is obtained by considering any permutation σ preserving the sets of numbers in each column of r, multiplied by the sign of σ. The sum of these terms is Q_r. We define $E^{(r)} = P_r Q_r$.

The primitive idempotents for the seminormal and the orthogonal representations are the same. They are defined by an iterative procedure from the E's. We let r^* denote the tableau, in the numbers $1, 2, \ldots, n-1$, obtained by deleting the box of r which contains the number n, and $\alpha^* = \alpha^*(r)$ the frame for S_{n-1} to which r^* corresponds. We also need the numbers

$$\theta^{\alpha} = n!/f^{\alpha}. \qquad \text{(III. 8)}$$

Then we define, iteratively,

$$e_r^{\alpha} = \frac{1}{\theta^{\alpha}} e^* E^{(r)} e^*,$$

$$e^* = \frac{1}{\theta^{\alpha^*}} e^{**} E^{(r^*)} e^{**},$$

$$\vdots$$

$$e^1 = \epsilon。$$

We see that e_r^{α} is determined by a sequence $r^*, r^{**}, \cdots$ of standard tableaux, and not just by r.

Each e_r^{α} turns out[13] to be a primitive idempotent. The minimal left ideals $O_n e_r^{\alpha}$ yield equivalent representations for fixed α and different r's. Different frames α give inequivalent representations.

For a fixed frame α, the vectors e_{rs}^{α} are constructed in a manner discussed in the previous section; their number is $(f^{\alpha})^2$. The dimension of O_n is $n!$, and, in fact, $\sum_{\alpha} (f^{\alpha})^2 = n!$. Hence, the set of frames does provide all the equivalence classes for the left regular representation of S_n. We shall not give the explicit constructions for the vectors e_{rs}^{α} when $r \neq s$. These constructions, as well as proof of the foregoing facts, can be found in Ref. 13. (As we will see later, the vectors e_{rs}^{α} may be obtained from the representation matrices.)

We next discuss the matrices for the seminormal representation. Let us fix an equivalence class, labeled by a frame α, and a particular representation, labeled by a standard tableau s_0 (i.e., we are looking at the minimal left ideal $O_n e_{s_0}^{\alpha}$, spanned by $e_{rs_0}^{\alpha}$). We recall, from (II. 11), that the matrices for S_n will be independent of the index s_0. Since the neighboring transpositions τ_k $(k = 2, \ldots, n)$ generate S_n, it is enough to consider the various matrices $u_{rr'}^{\alpha}(\tau_k)$.

We start by defining axial distance. Given a tableau r, we may label r by the sequence $j_k(r)$ which gives the row in which any number k lies, or by the sequence $j_k'(r)$, which gives the column in which any k lies $(k = 1, 2, \ldots, n)$. The axial distance, $\eta_r(i,k)$, for r, between the two numbers k, $i = 1, 2, \ldots, n$, is defined by

$$\eta_r(i,k) = |j_k(r) - j_i(r)| + |j_k'(r) - j_k'(r)|。 \qquad \text{(III. 9)}$$

Then, for the seminormal representation we obtain

$$u_{rr}^{\alpha}(\tau_k) = \begin{cases} +1, & \text{if} \quad j_{k-1}(r) = j_k(r), \\ -1, & \text{if} \quad j_{k-1}'(r) = j_k'(r), \end{cases} \qquad \text{(III. 10)}$$

and if the tableau r' results from interchanging k and $k-1$ in r, with $j_k(r) > j_{k-1}(r)$,

$$\begin{aligned} u_{rr}^{\alpha}(\tau_k) &= -1/\eta, \quad & u_{rr'}^{\alpha}(\tau_k) &= 1 - 1/\eta^2, \\ u_{r'r}^{\alpha}(\tau_k) &= 1, \quad & u_{r'r'}^{\alpha}(\tau_k) &= 1/\eta, \end{aligned} \qquad \text{(III. 11)}$$

where $\eta = \eta_r(k, k-1)$. All other entries of the matrix for τ_k are zero.

Using the Greek letter μ to denote the matrices in the orthogonal representation, we have

$$\mu_{rr}^{\alpha}(\tau_k) = \begin{cases} +1, & \text{if} \quad j_{k-1}(r) = j_k(r), \\ -1, & \text{if} \quad j_{k-1}'(r) = j_k'(r), \end{cases} \qquad \text{(III. 12)}$$

and, if r' results from interchanging k and $k-1$ in r, with $j_k(r) > j_{k-1}(r)$,

$$\begin{aligned} &\mu^{\alpha}_{rr}(\tau_k) = -1/\eta, && \mu^{\alpha}_{rr'}(\tau_k) = [1-(1/\eta)^2]^{1/2}, \\ &\mu^{\alpha}_{r'r}(\tau_k) = [1-(1/\eta)^2]^{1/2}, && \mu^{\alpha}_{r'r'}(\tau_k) = 1/\eta. \end{aligned} \tag{III.13}$$

For a given α, the seminormal and orthogonal representations are equivalent. In fact, the matrix for the similarity transformation involved is diagonal. Both representations are, of course, real. However, only the orthogonal representation is unitary.

We will also need another function, the tableau function, ψ. It occurs in relating the orthogonal and seminormal representations. To define it, we consider a standard tableau, s, for S_n. We let $\varphi_s(n) = 1$, if n is in the first row of s, and define it by

$$\varphi_s(n) = \prod_{i=1}^{j} (1 + 1/\zeta_i),$$

if n lies in row j, where ζ_i is the axial distance, relative to s, from the number at the end of row i to n.

We next delete the box containing n, and find $\varphi_{s^*}(n-1)$, similarly, for this new tableau. We continue this process until we reach a tableau of one box (whose φ function is one). The tableau function is then

$$\psi_s = \varphi_s(n)\varphi_{s^*}(n-1)\varphi_{s^{**}}(n-2)\cdots 1. \tag{III.14}$$

The vectors of the orthogonal representation, ξ^{α}_{rs} and those of the seminormal representation, e^{α}_{rs}, are related by

$$\xi^{\alpha}_{rs} = (\psi_r/\psi_s)^{1/2} e^{\alpha}_{rs} \tag{III.15}$$

and the matrix elements are related by

$$\mu^{\alpha}_{rs}(\tau) = (\psi_s/\psi_r)^{1/2} u^{\alpha}_{rs}(\tau). \tag{III.16}$$

We note, as shown by Rutherford[14] and, more elegantly, by Boerner,[15]

$$e^{\alpha}_{rs} = \frac{1}{\theta^{\alpha}} \sum_{\tau \in S_n} u^{\alpha}_{sr}(\tau^{-1})\tau, \tag{III.17}$$

and

$$\xi^{\alpha}_{rs} = \frac{1}{\theta^{\alpha}} \sum_{\tau \in S_n} \mu^{\alpha}_{sr}(\tau^{-1})\tau \tag{III.18}$$

$$= \frac{1}{\theta^{\alpha}} \sum_{\tau \in S_n} \mu^{\alpha}_{rs}(\tau)\tau, \tag{III.19}$$

using unitarity. Thus one can explicitly obtain the vectors of both representations from their matrices.

By (II.9), (III.17), and (III.18),

$$\epsilon = \sum_{\alpha, r} e^{\alpha}_{rr} = \sum_{\alpha, r, \sigma \in S_n} \frac{1}{\theta^{\alpha}} u^{\alpha}_{rr}(\sigma)\sigma$$

$$= \sum_{\alpha, r} \xi^{\alpha}_{rr} = \sum_{\alpha, r, \sigma \in S_n} \frac{1}{\theta^{\alpha}} \mu^{\alpha}_{rr}(\sigma)\sigma. \tag{III.20}$$

Then, using (III.20) and the linear independence of the elements of S_n,

$$\sum_{\alpha, r} \frac{1}{\theta^{\alpha}} u^{\alpha}_{rr}(\sigma) = \sum_{\alpha, r} \frac{1}{\theta^{\alpha}} \mu^{\alpha}_{rr}(\sigma) = \delta_{\sigma\epsilon}. \tag{III.21}$$

The matrices for a representation and its conjugate are related. Note that if $k = 2, \ldots, n$, and if for a standard tableau r, $j_k(r) > j_{k-1}(r)$, then $j_k(\bar{r}) < j_{k-1}(\bar{r})$. Hence if r and r' are obtainable from each other by interchanging the numbers k and $k-1$, then

$$\mu^{\alpha}_{rr}(\tau_k) = -\mu^{\bar{\alpha}}_{\bar{r}\bar{r}}(\tau_k). \tag{III.22}$$

Also, since axial distance is unchanged under conjugation

$$\mu^{\alpha}_{rr'}(\tau_k) = \mu^{\bar{\alpha}}_{\bar{r}\bar{r}'}(\tau_k). \tag{III.23}$$

If, on the other hand, k and $k-1$ lie in the same row of r they will lie in the same column of $\bar{r}$, so that relation (III.22) holds in general. If $r \neq r'$, but r' is not obtainable from r by interchanging k and $k-1$, then both sides of (III.23) are zero, and this formula holds whenever $r \neq r'$.

IV. THE TENSOR COUPLING COEFFICIENTS

Up to now, the representation space we have used is O_n, whose elements are operators. We next consider other spaces giving equivalent representations as a prelude to a concrete realization of the tensor product.

Let $x_1, x_2, \ldots, x_n$ be a set of noncommuting variables and let $V[x]$ be the set of linear combinations of all permutations of these variables, that is

$$V[x] = \operatorname{span}\{x_{\rho(1)}x_{\rho(2)}\cdots x_{\rho(n)}; \rho \in S_n\}. \tag{IV.1}$$

Elements of $V[x]$ are called tensors. We denote the basis vectors by

$$\rho(x) = x_{\rho(1)}x_{\rho(2)}\cdots x_{\rho(n)}. \tag{IV.2}$$

We specify a representation W_x, on $V[x]$ by defining

$$W_x(\sigma)(\rho(x)) = (\sigma\rho)(x), \tag{IV.3}$$

for $\sigma, \rho \in S_n$, and extending it to all of $V[x]$ making it linear. By (IV.3), we see that the representation W_x is equivalent to the left regular representation of S_n on O_n. Another basis for $V[x]$ consists of the $e^{\beta}_{rs}(x)$, for all r, s, β. (We use the notation of the seminormal representation but the present discussion applies equally well in the orthogonal case.) It is obvious that the space $V[y]$, built, in an analogous way, on another set $y_1, y_2, \ldots, y_n$ of noncommuting variables, gives an equivalent representation, W_y.

In order to define the Clebsch–Gordan series and coefficients we must introduce further spaces.

First, we state the definition of tensor product. The tensor product $V \otimes V'$, of two vector spaces V and V', is the space spanned by all products $\varphi \otimes \varphi'$, $\varphi \in V$, $\varphi' \in V'$, with

$$\varphi \otimes (\varphi_1' + \varphi_2') = (\varphi \otimes \varphi_1') + (\varphi \otimes \varphi_2'),$$

$$(\varphi_1 + \varphi_2) \otimes \varphi' = (\varphi_1 \otimes \varphi') + (\varphi_2 \otimes \varphi'),$$

and, if a is any scalar,

$$(a\varphi) \otimes \varphi' = \varphi \otimes (a\varphi') = a(\varphi \otimes \varphi').$$

If V and V' are the spaces of two representations of a group G, then the representation $U \otimes U'$, acting on $V \otimes V'$, is defined by

$$(U \otimes U')(g)(\varphi \otimes \varphi') = (U(g)\varphi) \otimes (U'(g)\varphi'),$$

for $g \in G$.

We see that the space $V[x] \otimes V[y]$, acted on by $W_x \otimes W_y$, gives a realization of the tensor product of the left regular representation with itself. Identifying the tensor product of two basis vectors, $\rho(x)$ and $\sigma(y)$ $(\rho, \sigma \in S_n)$ with the monomial $\rho(x)\sigma(y) = x_{\rho(1)}x_{\rho(2)}\cdots x_{\rho(n)} \times y_{\sigma(1)}y_{\sigma(2)}\cdots y_{\sigma(n)}$, we write the representation $W_x \otimes W_y$ as

$$\begin{aligned}[(W_x \otimes W_y)(\tau)](\rho(x)\sigma(y)) &= (\tau\rho)(x)(\tau\sigma)(y) \\ &= \tau(\rho(x)\sigma(y)),\end{aligned} \qquad \text{(IV. 4)}$$

for $\tau \in S_n$.

It is well known that if G is a compact group (as are all finite groups) then any finite-dimensional representation of G is equivalent to a direct sum of irreducible components of the left regular representation of G. "Equivalent" means there is an isomorphism, between the given representation and the direct sum, which preserves the action of G. Thus the tensor product of the left regular representation with itself is isomorphic to a direct sum of irreducible components of the group ring and the matrices for the tensor product and for the direct sum are similar.

Formally, the decomposition is an isomorphism between objects which transform as the tensor product of representations and objects which transform as the sum of representations. From a matrix point of view, this says that the representation matrix for the product is similar to a block diagonal matrix resulting from the direct sum. This is the usual approach to the direct sum decomposition problem. As we shall see later, this latter point of view gives rise to the Clebsch–Gordan coefficients.

We can examine the direct sum decomposition in another way. We will show that each element of the tensor product $V[x] \otimes V[y]$ is actually equal to a sum of elements of various spaces each giving representations equivalent to the left regular representation on O_n. Each of these spaces will also be be identifiable with a certain space of polynomials in $\{x_j\}$ and $\{y_j\}$. The terms of the sum must, of course, be unique.

We now turn to the explicit construction of the required spaces. Each one must be a subspace of $V[x] \otimes V[y]$. Moreover, each will be modeled on $V[x]$. We first define $V[x,y]$ by

$$V[x,y] = \text{span}\{\rho(xy) = \rho(x)\rho(y) \colon \rho \in S_n\}.$$

Since, for $\sigma \in S_n$,

$$\begin{aligned}[(W_x \otimes W_y)(\sigma)](\rho(xy)) &= [(W_x \otimes W_y)(\sigma)](\rho(x)\rho(y)) \\ &= (\sigma\rho)(x)(\sigma\rho)(y),\end{aligned}$$

$V[x,y]$ is an invariant subspace of $V[x] \otimes V[y]$. There must be further spaces. Obviously the dimension of $V[x] \otimes V[y]$ is $(n!)^2$, while the dimension of $V[x,y]$ is only $n!$ A clue to the nature of the other irreducible spaces we seek is that in each of the monomials which span $V[x,y]$, the indices on the x's and y's occur in the same order, while for $V[x] \otimes V[y]$ there is no such simultaneity; any ordering of the x's can be accompanied by any ordering of the y's. (An example, showing this for S_2, is given below.) We are thus led to define, for $\sigma \in S_n$,

$$V[x, \sigma y] = \text{span}\{\rho(x\sigma y) = \rho(x)(\rho\sigma y) \colon \rho \in S_n\}. \qquad \text{(IV. 5)}$$

Then $V[x,y] = V[x, \epsilon y]$. As shown for $V[x,y]$, $V[x, \sigma y]$ is an invariant subspace of $V[x] \otimes V[y]$. Moreover, each space $V[x, \sigma y]$ gives a representation equivalent to the left regular representation on O_n, since, for $\tau \in S_n$,

$$\begin{aligned}[(W_x \otimes W_y)(\tau)](\rho(x\sigma y)) &= (W_x \otimes W_y)(\tau)\rho(x)(\rho\sigma)(y) \\ &= \tau(\rho(x)(\rho\sigma)(y)) \\ &= \tau(\rho(x\sigma y)) \\ &= (\tau\rho)(x\sigma y).\end{aligned}$$

We now state and prove the required direct sum reduction.

Theorem IV.1: $V[x] \otimes V[y]$ is the direct sum of the invariant subspaces $V[x, \sigma y]$, taken over all $\sigma \in S_n$, that is,

$$V[x] \otimes V[y] = \sum_{\sigma \in S_n} \oplus V[x, \sigma y]. \qquad \text{(IV. 6)}$$

Proof of Theorem IV.1: The dimension of $V[x] \otimes V[y]$ is $(n!)^2$, that of each of the $n!$ subspaces $V[x, \sigma y]$ is $n!$ Thus we only need to show that the monomials $\rho(x\sigma y)$, for $\rho, \sigma \in S_n$, span $V[x] \otimes V[y]$. A basis for $V[x] \otimes V[y]$ consists of all $\sigma(x)\tau(y)$, $\sigma, \tau \in S_n$, and

$$\begin{aligned}\sigma(x)\tau(y) &= \sigma(x)\sigma(\sigma^{-1}\tau y) \\ &= \sigma(x)((\sigma^{-1}\tau)y)\end{aligned}$$

as required.

The direct sum decomposition can be described in terms of the basis vectors, in particular, those determined by various e's and products of e's. We see that for each σ, the set of all $e^{\beta}_{rs}(x\sigma y)$ is a basis of $V[x, \sigma y]$ and the set of $e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y)$ is a basis of $V[x] \otimes V[y]$. So what we wish to do is to express each $e^{\beta}_{rs}(x\sigma y)$ as a linear combination of various $e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y)$ and each $e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y)$ as a linear combination of vectors $e^{\beta}_{rs}(x\sigma y)$, for various σ.

We have by (III. 17),

$$\begin{aligned}e^{\beta}_{rs}(x\sigma y) &= \frac{1}{\theta^{\beta}} \sum_{\rho \in S_n} u^{\beta}_{sr}(\rho^{-1})\rho(x\sigma y) \\ &= \frac{1}{\theta^{\beta}} \sum_{\rho \in S_n} u^{\beta}_{sr}(\rho^{-1})\rho(x)\rho(\sigma y) \\ &= \frac{1}{\theta^{\beta}} \sum_{\substack{\rho \in S_n \\ p,q,\alpha \\ p',l,\alpha'}} u^{\beta}_{sr}(\rho^{-1})u^{\alpha}_{pq}(\rho)e^{\alpha}_{pq}(x)u^{\alpha'}_{p'l}(\rho)e^{\alpha'}_{p'l}(\sigma y).\end{aligned}$$

Now

$$\begin{aligned}e^{\alpha'}_{p'l}(\sigma y) &= \sum_{k,q',\alpha''} e^{\alpha'}_{p'l}u^{\alpha''}_{kq'}(\sigma)e^{\alpha''}_{kq'}(y) \\ &= \sum_{q'} u^{\alpha'}_{lq'}(\sigma)e^{\alpha'}_{p'q'}(y),\end{aligned}$$

by (II. 11) and (II. 7). Hence

$$\begin{aligned}e^{\beta}_{rs}(x\sigma y) = \frac{1}{\theta^{\beta}} \sum_{\substack{\rho \in S_{n,l} \\ p,q,\alpha \\ p',q',\alpha' \\ l}} & u^{\beta}_{sr}(\rho^{-1})u^{\alpha}_{pq}(\rho)u^{\alpha'}_{p'l}(\rho)\,u^{\alpha'}_{lq'}(\sigma) \\ & \times e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y).\end{aligned} \qquad \text{(IV. 7)}$$

Defining

$$A^{\beta}_{rs}(p,q,\alpha;p'q',\alpha')=\frac{1}{\theta^{\beta}}\sum_{\rho\in S_n} u^{\beta}_{sr}(\rho^{-1})u^{\alpha}_{pq}(\rho)u^{\alpha'}_{p'q'}(\rho), \qquad \text{(IV. 8)}$$

we have

$$e^{\beta}_{rs}(x\sigma y)=\sum_{\substack{p,q,\alpha,l\\ p',q',\alpha'}} A^{\beta}_{rs}(p,q,\alpha;p',l,\alpha')u^{\alpha'}_{lq'}(\sigma)e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y). \qquad \text{(IV. 9)}$$

The coefficients A are unique with respect to the expansion (IV. 9) because the set of $e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y)$ is linearly independent. We call the A's the tensor coupling coefficients.

Conversely, for each p,q,α,p',q',α',

$$\begin{aligned}\theta^{\alpha}\theta^{\alpha}\, & e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y)\\
&=\sum_{\tau,\rho\in S_n} u^{\alpha}_{qp}(\tau^{-1})u^{\alpha'}_{q'p'}(\rho^{-1})\tau(x)\rho(y)\\
&=\sum_{\tau,\rho\in S_n} u^{\alpha}_{qp}(\tau^{-1})u^{\alpha'}_{q'p'}(\rho^{-1})\tau(x\tau^{-1}\rho y)\\
&=\sum_{\tau,\sigma\in S_n} u^{\alpha}_{qp}(\tau^{-1})u^{\alpha'}_{q'p'}(\sigma^{-1}\tau^{-1})\tau(x\sigma y)\\
&=\sum_{\tau,\sigma\in S_n} u^{\alpha}_{qp}(\tau^{-1})u^{\alpha}_{q'k}(\sigma^{-1})u^{\alpha'}_{kp'}(\tau^{-1})\tau(x\sigma y)\\
&=\sum_{\substack{\tau,\sigma\in S_n\\ k,r,s,\beta}}^{k} u^{\alpha}_{qp}(\tau^{-1})u^{\alpha'}_{q'k}(\sigma^{-1})u^{\alpha'}_{kp'}(\tau^{-1})u^{\beta}_{rs}(\tau)e^{\beta}_{rs}(x\sigma y),\end{aligned}$$

and

$$\begin{aligned}\theta^{\alpha}\theta^{\alpha'} & e^{\alpha}_{pq}(x)e^{\alpha'}_{p'q'}(y)\\
&=\sum_{\substack{\sigma\in S_n\\ k,r,s,\beta}} \theta^{\beta}A^{\beta}_{sr}(q,p,\alpha;k,p',\alpha')u^{\alpha'}_{q'k}(\sigma^{-1})e^{\beta}_{rs}(x\sigma y). \qquad \text{(IV. 10)}\end{aligned}$$

The coefficients $\mathcal{A}$, for the orthogonal representation, are defined analogously, and we have

$$\mathcal{A}^{\beta}_{rs}(p,q,\alpha;p';q',\alpha')=\frac{1}{\theta^{\beta}}\sum_{\rho\in S_n} \mu^{\beta}_{sr}(\rho^{-1})\mu^{\alpha}_{pq}(\rho)\mu^{\alpha'}_{p'q'}(\rho), \qquad \text{(IV. 11)}$$

$$\xi^{\beta}_{rs}(x\sigma y)=\sum_{\substack{p,q,\alpha\\ p',q',\alpha',l}} \mathcal{A}^{\beta}_{rs}(p,q,\alpha;p',l\ ,\alpha')\mu^{\alpha'}_{lq'}(\sigma)\xi^{\alpha}_{pq}(x)\xi^{\alpha'}_{p'q'}(y), \qquad \text{(IV. 12)}$$

and

$$\begin{aligned}\theta^{\alpha}\theta^{\alpha'} & \xi^{\alpha}_{pq}(x)\xi^{\alpha'}_{p'q'}(y)\\
&=\sum_{\substack{\sigma\in S_n\\ k,r,s,\beta}} \theta^{\beta}\mathcal{A}^{\beta}_{sr}(q,p,\alpha;k,p',\alpha')\mu^{\alpha'}_{q'k}(\sigma^{-1})\xi^{\beta}_{rs}(x\sigma y). \qquad \text{(IV. 13)}\end{aligned}$$

Also,

$$\begin{aligned}\mathcal{A}^{\beta}_{rs}(p,q,\alpha;p',q',\alpha') &=\left(\frac{\psi^{\beta}_{r}}{\psi^{\beta}_{s}}\ \frac{\psi^{\alpha}_{q}}{\psi^{\alpha}_{p}}\ \frac{\psi^{\alpha'}_{q'}}{\psi^{\alpha'}_{p'}}\right)^{1/2}\\
&\quad\times A^{\beta}_{rs}(p,q,\alpha;p',q',\alpha'). \qquad \text{(IV. 14)}\end{aligned}$$

We consider the case of S_2 as an example; here the orthogonal and seminormal representations are the same. O_2 is two-dimensional; it is the direct sum of two irreducible and inequivalent one-dimensional representations. The symmetric representation is labeled by the frame $\alpha_2=(2,0)$:$\square\square$; it is spanned by the basis vector $\epsilon+(12)$, where ϵ is the identity of S_2 and (12) interchanges 1 and 2. The antisymmetric representation is labeled by the frame $\alpha_1=(1,1)$:$\begin{smallmatrix}\square\\ \square\end{smallmatrix}$, and is spanned by $\epsilon-(12)$. Writing

$$e^{\alpha_1}=e^{\alpha_1}_{11}=\epsilon-(12),\quad e^{\alpha_2}=e^{\alpha_2}_{11}=\epsilon+(12), \qquad \text{(IV. 15)}$$

we see that $V[x]$ is spanned by

$$e^{\alpha_1}(x)=x_1x_2-x_2x_1,\quad e^{\alpha_2}(x)=x_1x_2+x_2x_1.$$

Note that since both irreducible representations occur once and have dimension one, both subscripts, on both e's, are ones.

Then the basis vectors for $V[x]\otimes V[y]$, which is four-dimensional, are:

$$\begin{aligned}
e^{\alpha_1}(x)e^{\alpha_1}(y)&=(x_1x_2-x_2x_1)(y_1y_2-y_2y_1),\\
e^{\alpha_1}(x)e^{\alpha_2}(y)&=(x_1x_2-x_2x_1)(y_1y_2+y_2y_1),\\
e^{\alpha_2}(x)e^{\alpha_1}(y)&=(x_1x_2+x_2x_1)(y_1y_2-y_2y_1),\\
e^{\alpha_2}(x)e^{\alpha_2}(y)&=(x_1x_2+x_2x_1)(y_1y_2+y_2y_1).
\end{aligned}$$

One finds that the matrix elements for the left regular representation are

$$\begin{aligned}
&u^{\alpha_1}_{11}(\epsilon)=u^{\alpha_2}_{11}(\epsilon)=+1,\\
&u^{\alpha_1}_{11}(12)=-1,\quad u^{\alpha_2}_{11}(12)=+1.
\end{aligned}$$

The tensor coupling coefficients are given by

$$\begin{aligned}
A^{\alpha_1}_{11}(1,1,\alpha_1;1,1,\alpha_2)&=A^{\alpha_1}_{11}(1,1,\alpha_2;1,1,\alpha_1)\\
&=A^{\alpha_2}_{11}(1,1,\alpha_1;1,1,\alpha_1)\\
&=A^{\alpha_2}_{11}(1,1,\alpha_2;1,1,\alpha_2)\\
&=+1;
\end{aligned}$$

the rest are all zero.

We then have

$$\begin{aligned}
&e^{\alpha_1}(xy)=e^{\alpha_1}(x\epsilon y)=e^{\alpha_1}(x)e^{\alpha_2}(y)+e^{\alpha_2}(x)e^{\alpha_1}(y),\\
&e^{\alpha_2}(x(12)y)=-e^{\alpha_1}(x)e^{\alpha_1}(y)+e^{\alpha_2}(x)e^{\alpha_2}(y),
\end{aligned}$$

and so forth.

In the reduction of the tensor product we have

$$\begin{aligned}
e^{\alpha_1}(x)e^{\alpha_1}(y)&=\tfrac{1}{2}(e^{\alpha_2}(xy)-e^{\alpha_2}(x(12)y)),\\
e^{\alpha_1}(x)e^{\alpha_2}(y)&=\tfrac{1}{2}(e^{\alpha_1}(xy)+e^{\alpha_1}(x(12)y)),\\
e^{\alpha_2}(x)e^{\alpha_1}(y)&=\tfrac{1}{2}(e^{\alpha_1}(xy)-e^{\alpha_1}(x(12)y)),\\
e^{\alpha_2}(x)e^{\alpha_2}(y)&=\tfrac{1}{2}(e^{\alpha_2}(xy)+e^{\alpha_2}(x(12)y)).
\end{aligned}$$

Note that in each of the above sums, the first term is from $V[x,y]$, the second from $V[x,\sigma y]$.

We have given a technique for decomposing the product representation on $O_n\otimes O_n$; we have used a subspace approach. Of course, O_n does not give an irreducible representation. In fact, our method does not reduce the tensor product of two irreducible representations into a direct sum of irreducible representations. To see this, let

$$\begin{aligned}
&V^{\alpha,q}[x]=\text{span}\{e^{\alpha}_{pq}(x)\text{: any } p\},\\
&V^{\beta,s}[x,\sigma y]=\text{span}\{e_{rs}(x\sigma y)\text{: any } y\}.
\end{aligned}$$

These specify two irreducible representations. By irreducibility, for each $\sigma,\beta,\alpha,\alpha',s,q,q'$, either

$$V^{\beta,s}[x,\sigma y]\subset V^{\alpha,q}[x]\otimes V^{\alpha',q'}[y], \qquad \text{(IV. 16)}$$

or

$$V^{\beta,s}[x,\sigma y]\cap(V^{\alpha,q}[x]\otimes V^{\alpha',q'}[y])=0. \tag{IV.17}$$

If we could write, for each α,α',q,q'

$$V^{\alpha,q}[x]\otimes V^{\alpha',q'}[y]=\sum\oplus V^{\beta,s}[x,\sigma y],$$

where the direct sum is extended over all σ,β,s such that (IV.16) holds, then since $V[x]\otimes V[y]=\sum\oplus V^{\beta,s}[x,\sigma y]$, summing over all σ, β, and s, we would have, that for each σ, β, and s

$$V^{\beta,s}[x,\sigma y]\subset V^{\alpha,q}[x]\otimes V^{\alpha'q'}[y],$$

for exactly one value of each of the indices α, α', q, and q'. This, in turn, would imply that in the expansion (IV.9), for fixed σ, β, and s, exactly one value of each of the indices α, α', q, and q' can occur in the sum on the right. There are many counterexamples to this, such as the case where β is the completely symmetric representation.

After developing more machinery we will return to the question of decomposing the tensor product of irreducible representations into a direct sum of irreducible representations.

V. THE CLEBSCH-GORDAN COEFFICIENTS

Instead of constructing further equalities between basis vectors for various representations, we will view the decomposition problem more abstractly. In our second approach, we investigate an isomorphism between the tensor product of irreducible representations and a direct sum of irreducible representations. Later we will see how the two approaches merge.

The notation we will use in this discussion is that of the orthogonal representation. However, all of our remarks, except as specifically noted, are also valid in the seminormal case. Starting with two irreducible representations, labeled by frames α and α', and generated by vectors ξ^{α}_{qq} and $\xi^{\alpha'}_{q'q'}$, respectively (i.e., given by the minimal left ideals $O_n\xi^{\alpha}_{qq}$ and $O_n\xi^{\alpha'}_{q'q'}$), we form the tensor product $O_n\xi^{\alpha}_{qq}\otimes O_n\xi^{\alpha'}_{q'q'}$. It is equivalent to a direct sum $\sum\oplus O_n\xi^{\alpha_0}_{s_0s_0}$. (For a given frame α_0, any tableau s_0 can be used; all give equivalent representations.) The number of occurrences of a particular α_0 in this sum is the multiplicity $m(\alpha_0,\alpha,\alpha)=m(\alpha_0)$, which can be zero. The direct sum, more precisely, is $\sum\oplus\operatorname{span}\{\xi^{\alpha_0,\lambda}_{rs_0}\}_{r=1,\ldots,f^{\alpha_0}}$, where λ runs from 1 to $m(\alpha_0)$, and $\operatorname{span}\{\xi^{\alpha_0,\lambda}_{rs_0}\}$ gives the λth occurrence of the class α_0.

Let $C=C^{(q,q')}$ be a one-to-one linear transformation which maps the direct sum onto $O_n\xi^{\alpha}_{qq}\otimes O_n\xi^{\alpha'}_{q'q'}$ and which preserves the action of S_n. We write

$$C^{-1}(\xi^{\alpha}_{pq}\otimes\xi^{\alpha'}_{p'q'})=\sum_{\alpha_0,\lambda,r}(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\xi^{\alpha_0,\lambda}_{rs_0}.$$

For the moment, the C's depend on q, q', and s_0. Since C preserves the action of S_n,

$$C^{-1}\left(\sum_{k,k'}\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)\xi^{\alpha}_{kq}\otimes\xi^{\alpha'}_{k'q'}\right)=\sum_{\alpha_0,\lambda,r,r'}(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\mu^{\alpha_0}_{r'r}(\tau)\xi^{\alpha_0,\lambda}_{r's_0}, \tag{V.1}$$

and so

$$\sum_{\substack{k,k'\\ \tilde{\alpha}_0,\bar{\lambda},t}}\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)(C^{-1})^{\tilde{\alpha}_0,\bar{\lambda}}_{\bar{\lambda}t}(k,k')\xi^{\tilde{\alpha}_0,\bar{\lambda}}_{ts_0}=\sum_{\alpha_0,\lambda,r,r'}(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\mu^{\alpha_0}_{r'r}(\tau)\xi^{\alpha_0,\lambda}_{r's_0}. \tag{V.2}$$

The set $\{\xi^{\alpha_0,\lambda}_{r's_0}\}$ is a basis for the direct sum so that for each α_0, r', λ, p, and p',

$$\sum_{k,k'}\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)(C^{-1})^{\alpha_0}_{\lambda r'}(k,k')=\sum_{r}(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\mu^{\alpha_0}_{r'r}(\tau). \tag{V.3}$$

Applying the linear transformation C to (V.1), we also have

$$\begin{aligned}\sum_{k,k'}\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)\xi^{\alpha}_{kq}\otimes\xi^{\alpha'}_{k'q'}&=\sum_{\alpha_0,\lambda,r,r'}(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\mu^{\alpha_0}_{r'r}(\tau)C(\xi^{\alpha_0,\lambda}_{r's_0})\\&=\sum_{\substack{\alpha_0,\lambda,r,r'\\ \tilde{p},\tilde{p}'}}(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\mu^{\alpha_0}_{r'r}(\tau)(C)^{\alpha_0}_{\lambda r'}(p,p')\xi^{\alpha}_{\tilde{p}q}\otimes\xi^{\alpha'}_{\tilde{p}'q'}.\end{aligned}$$

Thus for each k,k',p,p',

$$\begin{aligned}\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)&=\sum_{\alpha_0,\lambda,r,r'}(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\mu^{\alpha_0}_{r'r}(\tau)(C)^{\alpha_0}_{\lambda r'}(k,k')\\&=\sum_{\alpha_0,\lambda,r,r'}(C)^{\alpha_0}_{\lambda r'}(k,k')\mu^{\alpha_0}_{r'r}(\tau)(C^{-1})^{\alpha_0}_{\lambda r}(p,p')\end{aligned}$$

and

$$\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)=\sum_{\substack{\alpha_0,\lambda,r\\ \alpha_0',\lambda',r'}}(C)^{\alpha_0}_{\lambda r'}(k,k')\mu^{\alpha_0}_{r'r}(\tau)\delta_{\lambda\lambda'}\delta^{\alpha_0\alpha_0'}(C^{-1})^{\alpha_0'}_{\lambda'r}(pp'). \tag{V.4}$$

The δ's in the preceding formula display the matrices as block diagonal, which is the appropriate form for a direct sum.

Different values of q and q' actually give the same (not merely similar) matrices (II.12). Thus the C's can be taken independent of q and q', which justifies our suppression of these indices. Likewise, the index s_0 is irrelevant.

Note that on the left side of formula (V.4) the rows of the matrix for the tensor product are labeled by pairs (k,k'), the columns by pairs (p,p'). To interpret the right side as the (k,k'), (p,p') entry of the product of matrices requires that we take the rows of the matrix for C as labeled by (k,k'), the columns by α_0,λ,r'. For the matrix for C^{-1}, the rows are labeled by α_0',λ',r and the columns by (p,p'). The terms $\mu^{\alpha_0}_{r'r}(\tau)\delta_{\lambda\lambda'}\delta^{\alpha_0\alpha_0'}$ form the entries of a block diagonal matrix $M(\tau)$: The rows of $M(\tau)$ are labeled by α_0,λ,r', the columns by α_0',λ',r. [The corresponding entry is $\mu^{\alpha_0}_{r'r}(\tau)\delta_{\lambda\lambda'}\delta^{\alpha_0\alpha_0'}$.] Each block is labeled by a pair α_0,λ.

As we will see later, the C's, in general, are not unique.

It is known[16] that when all representations involved are unitary, the transformation used in the reduction of the tensor product can be taken to be unitary. Thus, for the orthogonal representation, we choose the similarity transformation to be unitary. This means that

$$\sum_{\alpha_0,\lambda,r'} C^{\alpha_0}_{\lambda r'}(p,p')\overline{C^{\alpha_0}_{\lambda r'}(k,k')}=\delta_{pk}\delta_{p'k'} \qquad (\text{V.5})$$

and

$$\sum_{p,p'} C^{\alpha_0}_{\lambda r}(p,p')\overline{C^{\alpha_0'}_{\lambda' r'}(p,p')}=\delta^{\alpha_0\alpha_0'}\delta_{\lambda\lambda'}\delta_{rr'}. \qquad (\text{V.6})$$

Also, (V.4) becomes

$$\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)$$

$$=\sum_{\substack{\alpha_0,\lambda,r\\ \alpha_0',\lambda',r'}} C^{\alpha_0}_{\lambda r'}(k,k')\mu^{\alpha_0}_{r'r}(\tau)\delta^{\alpha_0\alpha_0'}\delta_{\lambda\lambda'}\times C^{\alpha_0'}_{\lambda' r}(p,p') \qquad (\text{V.7})$$

Analogous to (V.7) the reduction of the tensor product in the seminormal case takes the form

$$u^{\alpha}_{kp}(\tau)u^{\alpha'}_{k'p'}(\tau)$$

$$=\sum\left(\frac{\psi^{\alpha}_{k}\psi^{\alpha'}_{k'}}{\psi^{\alpha_0}_{r'}}\right)^{1/2} C^{\alpha_0}_{\lambda r'}(k,k')$$

$$\times u^{\alpha_0}_{r'r}(\tau)\left(\frac{\psi^{\alpha_0'}_{r}}{\psi^{\alpha}_{p}\psi^{\alpha'}_{p'}}\right)^{1/2}\overline{C^{\alpha_0'}_{\lambda r}(p,p')} \qquad (\text{V.8})$$

by (III.16).

VI. RELATIONS BETWEEN THE COEFFICIENTS

We have decomposed tensor products of representations in two ways, obtaining two sets of coefficients. Now we develop relations between these and show how to construct one set from the other, for the orthogonal case. We then return to the decomposition problem discussed in Sec. IV.

We start from an equivalent form of the mutual annihilation property (II.7),

$$\sum_{\tau}\mu^{\alpha}_{pq}(\tau)\mu^{\alpha'}_{p'q'}(\tau)=\theta^{\alpha}\delta^{\alpha\alpha'}\delta_{qq'}\delta_{pp'}\ . \qquad (\text{VI.1})$$

We multiply both sides of (V.7) by $\mu^{\alpha_0}_{rr'}(\tau)$ and sum on τ to obtain

$$\frac{1}{\theta^{\alpha_0}}\sum_{\tau\in S_n}\mu^{\alpha_0}_{rr'}(\tau)\mu^{\alpha}_{kp}(\tau)\mu^{\alpha'}_{k'p'}(\tau)$$

$$=\sum_{\lambda=1}^{m(\alpha_0)} C^{\alpha_0}_{\lambda r}(k,k')\overline{C^{\alpha_0}_{\lambda r'}(p,p')}, \qquad (\text{VI.2})$$

i.e.,

$$\mathcal{A}^{\alpha_0}_{rr'}(k,p,\alpha;k',p',\alpha')$$

$$=\sum_{\lambda=1}^{m(\alpha_0)} C^{\alpha_0}_{\lambda r}(k,k')\overline{C^{\alpha_0}_{\lambda r'}(p,p')}. \qquad (\text{VI.3})$$

If we take the Clebsch—Gordan coefficients to be unitary, we multiply (VI.3) by $C^{\alpha_0}_{\lambda' r'}(p,p')$, sum on (p,p'), using (V.6) to obtain

$$\sum_{p,p'}\mathcal{A}^{\alpha_0}_{rr'}(k,p,\alpha;k',p',\alpha')C^{\alpha_0}_{\lambda' r'}(p,p')$$

$$=C^{\alpha_0}_{\lambda' r}(k,k'). \qquad (\text{VI.4})$$

Note that if we multiply both sides of (VI.3) by $\mu^{\alpha_0}_{rr'}(\bar{\tau})=\mu^{\alpha_0}_{r'r}(\bar{\tau}^{-1})$ (by unitarity), sum on r and r', and use (III.21), we obtain (V.7). Thus, in the orthogonal case, when the Clebsch—Gordan matrix is taken to be unitary, (V.7) and (VI.3) are equivalent.

If $r=r'$, (VI.4) resembles an eigenvector equation. To so regard it, we must first specify, more precisely, the way in which the $\mathcal{A}$'s may be treated as matrices and the C's as vectors. Our observation will then be used to construct the Clebsch—Gordan coefficients. We will regard $\mathcal{A}^{\alpha_0}_{rr'}(k,p,\alpha;k',p',\alpha')$ as an entry of the matrix $\mathcal{A}^{\alpha_0}_{rr'}$, whose rows are labeled by pairs (k,k') and whose columns are labeled by pairs (p,p'). Each set $\{C^{\alpha_0}_{\lambda' r'}(p,p')\}$, for fixed α_0,λ',r' forms a vector whose coordinates are labeled by pairs (p,p'). Relation (V.6) is the assertion that the set of such vectors is orthonormal. Hence the vectors $C^{\alpha_0}_{\lambda' r'}$ are linearly independent. Then, from (VI.4), the rank of the matrix is at least equal to the multiplicity $m(\alpha_0)$. To see that equality holds, let Y be a vector orthogonal to each vector $C^{\alpha_0}_{\lambda r'}$, i.e., suppose

$$\sum_{p,p'}Y(p,p')\overline{C^{\alpha_0}_{\lambda' r'}(p,p')}=0.$$

For each λ',

$$\sum_{p,p'}\mathcal{A}^{\alpha_0}_{rr'}(k,p,\alpha;k',p',\alpha')Y(p,p')$$

$$=\sum_{p,p'}C^{\alpha_0}_{\lambda r}(k,k')\overline{C^{\alpha_0}_{\lambda r'}(p,p')}Y(p,p')=0.$$

Thus Y is in the null space of $\mathcal{A}^{\alpha_0}_{rr'}$ and the rank of $\mathcal{A}^{\alpha_0}_{rr'}$ must be exactly $m(\alpha_0)$.

As a result, the coefficients $C^{\alpha_0}_{\lambda r_0}(p,p')$ can be interpreted as giving vectors which span the eigenspace, for eigenvalue one, of the matrix $\mathcal{A}^{\alpha_0}_{r_0r_0}$. Since any vector orthogonal to these vectors is in the null space of $\mathcal{A}^{\alpha_0}_{r_0r_0}$, this matrix gives an orthogonal projection (i.e., a self-adjoint projection). We now use the fact that the columns of the matrix of a projection form vectors which span the range to determine the Clebsch—Gordan coefficients.

We begin by picking $m(\alpha_0)$ linearly independent columns of $\mathcal{A}^{\alpha_0}_{r_0r_0}$, since $m(\alpha_0)$ is the dimension of the range. We can use the Gram—Schmidt process to select these columns and orthonormalize them. The result is a set of $m(\alpha_0)$ orthonormal eigenvectors, of eigenvalue one, for the matrix $\mathcal{A}^{\alpha_0}_{r_0r_0}$. Their components are the coefficients $C^{\alpha_0}_{\lambda r_0}(p,p')$, $\lambda=1,\ldots,m(\alpha_0)$. Note that the entries of the matrix $\mathcal{A}^{\alpha_0}_{r_0r_0}$ are real so that the C's we obtain are also real. We now have one of the columns of the Clebsch—Gordan matrix in each of the $m(\alpha_0)$ sets of columns corresponding to α_0; the columns in question are labeled by α_0 and $\lambda=1,\ldots,m(\alpha_0)$. We now wish to find the entries in the other columns corresponding to α_0. The above process can only be used to find one column for each fixed α_0 and λ; we cannot repeat it for different values of r and expect (VI.3) to hold in general. To obtain the remaining coefficients $C^{\alpha_0}_{\lambda r}(k,k')$ $(r\neq r_0)$, either we can use the identity (VI.4) or we can solve the Eqs. (VI.3) with $r'=r_0$. The second approach uses a system of $f^{\alpha}\times f^{\alpha'}$ nonhomogeneous equations, each labeled by a pair (p,p'). The rows of the coefficient matrix are thus labeled by a pair (p,p') and the columns by $\lambda=1,\ldots,m(\alpha_0)$. We have already seen that there are $m(\alpha_0)$ independent columns, so $m(\alpha_0)$ is the rank of this matrix. Thus, of the $f^{\alpha}\times f^{\alpha'}$ equations we need only take $m(\alpha_0)$

$$\mathcal{A}^{\alpha_0}_{rr_0}(k,p_j,\alpha;k',p_j',\alpha')$$

$$=\sum_{\lambda=1}^{m(\alpha_0)}\overline{C^{\alpha_0}_{\lambda r_0}(p_j,p_j')}C^{\alpha_0}_{\lambda r}(k,k'), \qquad \text{(VI.5)}$$

$j=1,\ldots,m(\alpha_0)$.

The pairs (p_j,p_j') must be chosen so that the rows $C^{\alpha_0}_{\lambda r_0}(p_j,p_j')$ (of the coefficient matrix of the system) are linearly independent, i.e., so that any relation of the form

$$\sum_{j=1}^{m(\alpha_0)}\eta_j C^{\alpha_0}_{\lambda r_0}(p_j,p_j')=0 \qquad \text{(VI.6)}$$

for all λ, implies each $\eta_j = 0$. If relation (VI.6) holds, we multiply both sides by $\overline{C^{\alpha_0}_{\lambda r_0}(q,q')}$ and sum on λ, using (VI.3), to obtain

$$\sum_{j=1}^{m(\alpha_0)}\eta_j\mathcal{A}^{\alpha_0}_{r_0r_0}(p_j,q,\alpha;p_j',q',\alpha')=0,$$

for each pair (q,q'). Hence, if we choose pairs (p_j,p_j'), $j=1,\ldots,m(\alpha_0)$, making the rows $\mathcal{A}^{\alpha_0}_{r_0r_0}(p_j,q,\alpha;p_j',q',\alpha')$ independent, then the corresponding rows $C^{\alpha_0}_{\lambda r_0}(p_j,p_j')$ are also independent. Because the rank of $\mathcal{A}^{\alpha_0}_{r_0r_0}$ is $m(\alpha_0)$, this method of selection gives the correct number of pairs. Since it is self-adjoint and has only real-valued entries, the matrix $\mathcal{A}^{\alpha_0}_{r_0r_0}$ is symmetric. Thus, the labels (p_j,p_j'), found in the process of orthonormalizing its columns, are the labels of the linearly independent rows of $\mathcal{A}^{\alpha_0}_{r_0r_0}$.

Thus, in summary, we find the coefficients as follows. We fix an r_0 and find $m(\alpha_0)$ independent columns of $\mathcal{A}^{\alpha_0}_{r_0r_0}$, label them by the pairs (p_j,p_j'), $j=1,\ldots,m(\alpha_0)$ and orthonormalize them. The components, labeled by (k,k'), of the resulting vectors, are the coefficients $C^{\alpha_0}_{\lambda r_0}(k,k')$. We then compute the inverse of the $m(\alpha_0)\times m(\alpha_0)$ coefficient matrix $C^{\alpha_0}_{\lambda r_0}(p_j,p_j')$ and use it to solve, uniquely, the system (VI.5) for the coefficients $C^{\alpha_0}_{\lambda r}(k,k')$ for $r\neq r_0$. Clearly, we obtain real-valued coefficients.

We see that our construction requires the calculation of the following sets of tensor coupling coefficients: $\mathcal{A}^{\alpha_0}_{r_0r_0}(p,k,\alpha;p',k',\alpha')$ for some fixed r_0, for all (k,k') and enough pairs (p,p') to give $m(\alpha_0)$ independent columns [the number of pairs (p,p') that must be tested is not known in advance], and the coefficients $\mathcal{A}^{\alpha_0}_{rr_0}(k,p,\alpha;k',p',\alpha')$ for all r,k, and k' and those pairs (p,p') which have just been found.

If the multiplicity, $m(\alpha_0)$, is one, the range of the projection given by $\mathcal{A}^{\alpha_0}_{r_0r_0}$ is one-dimensional, so that to compute $C^{\alpha_0}_{r_0}$ we need use only one column. We may pick any column, which we label by a pair (p_0,p_0'), which is not identically zero. Since the real-valued matrix $\mathcal{A}^{\alpha_0}_{r_0r_0}$ gives a self-adjoint projection, this condition means that the diagonal entry $\mathcal{A}^{\alpha_0}_{r_0r_0}(p_0,p_0,\alpha;p_0',p_0',\alpha')$ is strictly positive. That is,

$$\sum_{p,p'}\mathcal{A}^{\alpha_0}_{r_0r_0}(p,p_0,\alpha;p',p_0',\alpha)^2$$

$$=\sum_{p,p'}\mathcal{A}^{\alpha_0}_{r_0r_0}(p_0,p,\alpha;p_0',p',\alpha')$$

$$\times\mathcal{A}^{\alpha_0}_{r_0r_0}(p,p_0,\alpha;p',p_0',\alpha')$$

$$=\mathcal{A}^{\alpha_0}_{r_0r_0}(p_0,p_0,\alpha;p_0',p_0',\alpha'). \qquad \text{(VI.7)}$$

We then see that the square of the length of a column of $\mathcal{A}^{\alpha_0}_{r_0r_0}$ equals its diagonal element. The only part of the Gram–Schmidt procedure that is necessary in the present case is the normalization. This gives

$$C^{\alpha_0}_{r_0}(k,k')=\frac{\mathcal{A}^{\alpha_0}_{r_0r_0}(k,p_0,\alpha;k',p_0',\alpha')}{[\mathcal{A}^{\alpha_0}_{r_0r_0}(p_0,p_0,\alpha;p_0',p_0',\alpha')]^{1/2}}\ . \qquad \text{(VI.8)}$$

We note that

$$C^{\alpha_0}_{r_0}(p_0,p_0')=[\mathcal{A}^{\alpha_0}_{r_0r_0}(p_0,p_0,\alpha;p_0',p_0',\alpha')]^{1/2}. \qquad \text{(VI.9)}$$

To find $C^{\alpha_0}_{r}(k,k')$, for $r\neq r_0$, we note that the system (VI.5) reduces to one equation, whose solution is given by

$$C^{\alpha_0}_{r}(k,k')=\frac{\mathcal{A}^{\alpha_0}_{rr_0}(k,p_0,\alpha;k',p_0',\alpha')}{C^{\alpha_0}_{r_0}(p_0,p_0')}\,. \qquad \text{(VI.10)}$$

Since $C^{\alpha_0}_{r_0}(k,k')$ also satisfies this equation, (VI.10) may be used to compute all the coefficients, once the overall sign has been fixed.

Since the range of the projection $\mathcal{A}^{\alpha_0}_{r_0r_0}$ has dimension one, and since we require that the length of the vector chosen from this space be one, the C's are determined up to a phase. We take the C's to be real, so that only a sign remains to be fixed and this is done in Eq. (VI.9).

We must now prove that the coefficients we have constructed are actually Clebsch–Gordan coefficients, that is, that they satisfy (V.6) and (VI.3).

From the definition of the $\mathcal{A}$'s and the unitarity of the μ's, we have

$$\sum_{q,q'}\mathcal{A}^{\alpha_0}_{rr_0}(p,q,\alpha;p',q',\alpha')\mathcal{A}^{\alpha_0}_{sr_0}(k,q,\alpha;k',q',\alpha')$$

$$=\mathcal{A}^{\alpha_0}_{rs}(p,k,\alpha;p',k',\alpha'). \qquad \text{(VI.11)}$$

Moreover, by (VI.1) and the unitarity of the μ's, the definition (IV.11) of the $\mathcal{A}$'s gives the stronger result

$$\sum_{p,p'}\mathcal{A}^{\alpha_0}_{rs}(p,k,\alpha;p',k',\alpha')\mathcal{A}^{\alpha_0'}_{r's'}(p,q,\alpha;p',q',\alpha')$$

$$=\delta^{\alpha_0\alpha_0'}\delta_{rr'}\mathcal{A}^{\alpha_0}_{ss'}(k,q,\alpha;k',q',\alpha'). \qquad \text{(VI.12)}$$

Applying (VI.11), noting that our C's satisfy the system (VI.3) with $r'=r_0$, and using the orthonormality of the vectors $C^{\alpha_0}_{\lambda r_0}$ $[\lambda=1,\ldots,m(\alpha_0)]$, we have

$$\mathcal{A}^{\alpha_0}_{rs}(p,k,\alpha;p',k',\alpha')$$

$$=\sum_{\lambda,\rho}\sum_{q,q'}C^{\alpha_0}_{\lambda r}(p,p')C^{\alpha_0}_{\lambda r_0}(q,q')C^{\alpha_0}_{\rho s}(k,k')C^{\alpha_0}_{\rho r_0}(q,q')$$

$$=\sum_{\lambda}C^{\alpha_0}_{\lambda r}(p,p')C^{\alpha_0}_{\lambda s}(k,k'),$$

and our coefficients satisfy (VI.3), for all r and s.

Next, using (VI.12) with $r=s=r_0$ and $r'=s'=r_0'$, we see that the eigenvectors $C^{\alpha_0}_{\lambda r_0}$ and $C^{\alpha_0'}_{\lambda' r_0'}$ belong to (symmetric) projections whose product is zero if $\alpha_0'\neq\alpha_0$, and so for $\alpha_0'\neq\alpha_0$, these eigenvectors must be orthogonal. We must complete the argument that our construction yields orthonormal vectors for $r,s\neq r_0$; we use the fact that for $r=r_0$, we have an orthonormal set. Thus we may multiply (VI. 3) by $C^{\alpha_0}_{\lambda r_0}(p,p')$ and sum on (p,p') to obtain

$$\sum_{p,p'} \mathcal{A}^{\alpha_0}_{rr_0}(k,p,\alpha;k',p',\alpha')C^{\alpha_0}_{\lambda r_0}(p,p')$$

$$=C^{\alpha_0}_{\lambda r}(k,k').$$

Therefore,

$$\sum_{p,p'} C^{\alpha_0}_{\lambda r}(p,p')C^{\alpha'_0}_{\lambda' r'}(p,p')$$

$$=\sum_{p,p'}\mathcal{A}^{\alpha_0}_{rr_0}(p,k,\alpha;p',k',\alpha')$$

$$\times \mathcal{A}^{\alpha_0}_{r'r_0}(p,q,\alpha;p',q',\alpha')C^{\alpha_0}_{\lambda r_0}(k,k')C^{\alpha'_0}_{\lambda' r_0}(q,q')$$

$$=\delta^{\alpha_0\alpha'_0}\delta_{rr'}\sum_{\substack{k,k'\\q,q'}} \mathcal{A}^{\alpha_c}_{r_0r_0}(k,q,\alpha;k',q',\alpha')$$

$$\times C^{\alpha_0}_{\lambda r_0}(k,k')C^{\alpha_0}_{\lambda' r_0}(q,q')$$

$$=\delta^{\alpha_0\alpha'_0}\delta_{rr'}\sum_{k,k'} C^{\alpha_0}_{\lambda r_0}(k,k')C^{\alpha_0}_{\lambda' r_0}(k,k')$$

$$=\delta^{\alpha_0\alpha'_0}\delta_{\lambda\lambda'}\delta_{rr'},$$

and we have the required orthonormality.

We next consider the extent to which the Clebsch—Gordan coefficients are unique. Starting with the coefficients provided by our construction, we can obtain further sets of Clebsch—Gordan coefficients in the following manner: We let T be a unitary matrix with entries denoted by t's, whose rows and columns are labeled by the triplets β, λ, r. We also suppose the t's satisfy

$$t_{\beta,\lambda,r;\beta',\lambda',r'}=\begin{cases}0 & \text{if } \beta\neq\beta' \text{ or if } r\neq r',\\ t^{\beta}_{\lambda\lambda'} & \text{otherwise.}\end{cases} \tag{VI.13}$$

That is, the matrix T is block diagonal, with blocks labeled by classes β, and its entries depend only on the label given by the multiplicity index λ. Then

$$\sum_{\rho} t^{\beta}_{\mu\rho}\bar{t}^{\beta}_{\nu\rho}=\delta_{\mu\nu}\ . \tag{VI.14}$$

Thus if

$$D^{\beta}_{\lambda r}(p,p')=\sum_{\rho} t_{\lambda\rho}C^{\alpha_0}_{\rho r}(p,p'), \tag{VI.15}$$

the D's form a second set of Clebsch—Gordan coefficients as can be seen by direct substitution in (V.6) and (VI. 3).

On the other hand, suppose we have a second set of Clebsch—Gordan coefficients, $D^{\alpha_0}_{\lambda r}(p,p')$. We shall show they are related to the C's by a matrix T of the above form. First, since the coefficients $C^{\alpha_0}_{\lambda r}(p,p')$ and $D^{\alpha_0}_{\lambda r}(p,p')$ form two orthonormal sets of eigenvectors, corresponding to eigenvalue one, for the matrix $\mathcal{A}^{\alpha_0}_{rr}$, they are related by a unitary matrix, that is

$$D^{\alpha_0}_{\lambda r}(p,p')=\sum_{\lambda'} t^{\alpha_0}_{\lambda\lambda'}(r)C^{\alpha_0}_{\lambda' r}(p,p'), \tag{VI.16}$$

where

$$\sum_{\rho} t^{\alpha_0}_{\mu\rho}(r)\overline{t^{\alpha_0}_{\nu\rho}}(r)=\delta_{\mu\nu}, \tag{VI.17}$$

for each r. Moreover, we can see that the t's are independent of r: Relation (VI. 3) must hold for the D's as well as for the C's, and so

$$\sum_{\lambda} C^{\alpha_0}_{\lambda r}(p,p')C^{\alpha_0}_{\lambda s}(q,q')$$

$$=\sum_{\rho} D^{\alpha_0}_{\rho r}(p,p')\overline{D^{\alpha_0}_{\rho s}(q,q')}$$

$$=\sum_{\rho,\mu,\nu} t^{\alpha_0}_{\rho\mu}(r)\overline{t^{\alpha_0}_{\rho\nu}(s)}C^{\alpha_0}_{\mu r}(p,p')C^{\alpha_0}_{\nu s}(q,q').$$

Using the orthonormality condition (V.6) on the C's, twice, gives

$$\delta_{\mu\nu}=\sum_{\rho} t^{\alpha_0}_{\rho\mu}(r)\overline{t^{\alpha_0}_{\rho\nu}(s).}$$

Thus the inverse of the transpose of $t^{\alpha_0}(r)$ is $\overline{t^{\alpha_0}(s)}$. But the transpose of $t^{\alpha_0}(r)$ is also unitary so that $\overline{t^{\alpha_0}(s)}=\overline{t^{\alpha_0}(r)}$, and the t^{α_0}'s are indeed independent of r.

We have indicated how any matrix T, satisfying (VI. 14) allows us to construct a second set of Clebsch—Gordan coefficients from a previously obtained set. This method involves taking linear combinations where the sum is only over the multiplicity indices. Conversely, we have seen that any two sets of Clebsch—Gordan coefficients are related by (VI. 15) for a matrix T satisfying (VI. 14). We note that if the multiplicity is one, then (VI. 15) reduces to a change of phase.

In Sec. IV we expressed a space for the tensor product of the group ring with itself as a direct sum of spaces, on each of which S_n acts irreducibly. The carrier space consists of polynomials. Our approach did not actually give the decomposition of the tensor product of two irreducible representations. With the machinery of this and the preceding section, we again address ourselves to this question.

Initially we work in $O_n\otimes O_n$. We let

$$V^{\alpha,q}=\text{span}\{\xi^{\alpha}_{pq}\text{: any } p\},$$

and

$$g^{\alpha_0}_{\lambda r}(q,\alpha;q',\alpha')$$

$$=g^{\alpha_0}_{\lambda r}=\sum_{k,k'} C^{\alpha_0}_{\lambda r}(k,k')\xi^{\alpha}_{kq}\otimes\xi^{\alpha'}_{k'q'}, \tag{VI.18}$$

$g^{\alpha_0}_{\lambda r}\in V^{\alpha,q}\otimes V^{\alpha'q'}$. By orthonormality, $\{g^{\alpha_0}_{\lambda r}\}$ is linearly independent, for each α, α', q, and q'. Also, the number of $g^{\alpha_0}_{\lambda r}$'s is $\sum_{\alpha_0} m(\alpha_0)f^{\alpha_0}=f^{\alpha}\cdot f^{\alpha'}$, so the set $\{g^{\alpha_0}_{\lambda r}\}$ forms a basis for $V^{\alpha,q}\otimes V^{\alpha',q'}$. Let

$$V^{\alpha_0}_{\lambda}=\text{span}\{g^{\alpha_0}_{\lambda r}\text{: any } r\}. \tag{VI.19}$$

We have shown that $V^{\alpha,q}\otimes V^{\alpha',q'}=\sum_{\alpha_0,\lambda}\oplus V^{\alpha_0}_{\lambda}$, we must show that this decomposition reduces $\mu^{\alpha}\otimes\mu^{\alpha'}$. Thus, fix α_0 and λ. We have

$$\tau(g^{\alpha_0}_{\lambda r})=\sum_{k,k'} C^{\alpha_0}_{\lambda r}(k,k')\tau(\xi_{kq}\otimes\xi_{k'q'})$$

$$=\sum_{\substack{k,k'\\p,p'}} C^{\alpha_0}_{\lambda r}(k,k')\mu^{\alpha}_{pk}(\tau)\,\mu^{\alpha'}_{p'k'}(\tau)\xi^{\alpha}_{pq}\otimes\xi^{\alpha'}_{p'q'}$$

$$=\sum C^{\alpha_0}_{\lambda r}(k,k')C^{\beta}_{\nu s}(p,p')\mu^{\beta}_{ss'}(\tau)\,C^{\beta}_{\nu s'}(k,k')\xi^{\alpha}_{pq}\otimes\xi^{\alpha'}_{p'q'}\ ,$$

summing over repeated indices, using (VI. 3) and the fact that the C's are real valued. Using orthonormality,

$$\tau(g^{\alpha_0}_{\lambda r})=\sum_{s,p,p'} C^{\alpha_0}_{\lambda s}(p,p')\mu^{\alpha_0}_{sr}(\tau)\xi^{\alpha}_{pq}\otimes\xi^{\alpha'}_{p'q'}$$

$$=\sum_{s}\mu^{\alpha_0}_{sr}(\tau)g^{\alpha_0}_{\lambda s},$$

and so the representation $\mu^{\alpha}\otimes\mu^{\alpha'}$, restricted to $V_{\lambda}^{\alpha_0}$, is an irreducible representation of class α_0. Hence,

$$V^{\alpha,q}\otimes V^{\alpha' q'}=\sum_{\alpha_0,\lambda}\oplus V_{\lambda}^{\alpha_0} \tag{VI.20}$$

gives the direct sum reduction for the tensor product of two irreducible representations.

We also note, using (V.5), that

$$\sum_{\alpha_0,\lambda,r} C_{\lambda r}^{\alpha_0}(p,p')g_{\lambda r}^{\alpha_0}=\xi_{kq}^{\alpha}\otimes\xi_{k'q'}^{\alpha'}\,. \tag{VI.21}$$

Returning to $V[x]\otimes V[y]$, we see that $g_{\lambda r}^{\alpha_0}(q,\alpha;q',\alpha')$ is identified with

$$\begin{aligned}&g_{\lambda r}^{\alpha_0}(q,\alpha;q',\alpha')(x,y)\\ &\quad=g_{\lambda r}^{\alpha_0}(x,y)=\sum_{k,k'} C_{\lambda r}^{\alpha_0}(k,k')\xi_{kq}(x)\xi_{k'q'}(y).\end{aligned} \tag{VI.22}$$

Moreover,

$$\begin{aligned}&\theta^{\alpha}\theta^{\alpha'}g_{\lambda r}^{\alpha_0}(x,y)\\ &\quad=\sum_{k,k'}\theta^{\alpha}\theta^{\alpha'}C_{\lambda r}^{\alpha_0}(k,k')\xi_{kq}^{\alpha}(x)\xi_{k'q'}^{\alpha'}(y)\\ &\quad=\sum_{\substack{k,k'\\ l,\sigma\\ s,t}}\theta^{\alpha_0}C_{\lambda r}^{\alpha_0}(k,k')\mathcal{A}_{st}^{\alpha_0}(k,q,\alpha;k',l,\alpha')\\ &\qquad\times\mu_{q'l}^{\alpha'}(\sigma)\xi_{st}^{\alpha_0}(x\sigma y)\\ &\quad=\sum_{l,\sigma,t}C_{\lambda t}^{\alpha_0}(q,l)\mu_{q'l}^{\alpha'}(\sigma)\xi_{rt}^{\alpha_0}(x\sigma y),\end{aligned} \tag{VI.23}$$

using (IV.13), (VI.3), and orthonormality. Relation (VI.23) shows the vectors $\xi_{rt}^{\alpha_0}(x\sigma y)$ are combined to obtain vectors for reducing tensor products of irreducible representations. We see how subspaces involved in the direct sum for such a reduction are spanned by polynomials and not merely by products of polynomials as in (VI.18).

The polynomials we discussed in Sec. IV are labeled by equivalence class, representation, and vector. The space of products of pairs of polynomials does not reduce into a direct sum of subspaces of polynomials labeled in the same way. Neither does the expansion (IV.13) give a reduction into subspaces of the tensor product of two irreducible representations, even though we introduced further vectors, obtained from vectors labeled as above with certain changes of variables as specified by the elements, σ, of S_n. The $g_{\lambda r}^{\alpha_0}(x,y)$ are linear combinations summed over the representations and over the group of the latter states; this is the content of (VI.23). For each α_0 and λ we have a basis for a representation of class α_0 and when α_0 and λ are varied we obtain a basis for the tensor product of the two irreducible representations in question.

VII. THE ITERATIVE FORMULA

The tensor coupling coefficients can be calculated directly from the definitions (IV.8) and (IV.11). To do this we need each of $(n!)^2$ numbers $u_{pq}^{\alpha}(\tau)$, obtained by varying α, p, and q and $\tau\in S_n$. These would first have to be computed from the matrices for the neighboring transpositions.

We will use an alternate approach, and develop an iterative formula for the S_n coefficients which requires only the S_{n-1} coefficients and the matrix elements for that neighboring transposition which interchanges $n-1$ and n. We use the notation of the seminormal representation. All the results apply in the orthogonal case; they are obtained by replacing the A's with $\mathcal{A}$'s and the u's with μ's.

Our formula is based on the following decomposition of S_n in terms of S_{n-1}:

Lemma VII.1: $S_n=S_{n-1}\cup(S_{n-1}\tau_nS_{n-1})$ and

(a) $S_{n-1}\cap(S_{n-1}\tau_nS_{n-1})=\phi$;

(b) If σ_0, $\sigma_0'\in S_{n-1}$ then the number of pairs $\sigma,\sigma'\in S_{n-1}$ with $\sigma\tau_n\sigma'=\sigma_0\tau_n\sigma_0'$ is $(n-2)!$

Proof of Lemma VII.1: Elements of the set $S_{n-1}\tau_nS_{n-1}$ do not fix n so that (a) is immediate. To prove (b) we see that if σ_0,σ_0', $\sigma,\sigma'\in S_{n-1}$ with

$$\sigma\tau_n\sigma'=\sigma_0\tau_n\sigma_0', \tag{VII.1}$$

then $\tau_n=(\sigma_0^{-1}\sigma)\tau_n\sigma'\sigma_0'^{-1}$. Note that $\sigma_0^{-1}\sigma,\sigma'\sigma_0'^{-1}\in S_{n-1}$. Let $j\leqslant n-1$ be the image of $n-1$ under $\sigma'\sigma_0'^{-1}$. If j were less than $n-1$ then since $\tau_n(j)=j$ and $(\sigma_0^{-1}\sigma)(j)\leqslant n-1$, we get $n=\tau_n(n-1)=(\sigma_0^{-1}\sigma)(j)\leqslant n-1$, a contradiction. Thus, $\sigma'\sigma_0'^{-1}$ fixes $n-1$. The same is true of $\sigma_0^{-1}\sigma$. Both these products already fix n so that $\sigma_0^{-1}\sigma=\rho\in S_{n-2}$ and $\sigma'\sigma_0'^{-1}=\rho'\in S_{n-2}$. Then $\tau_n=\rho\tau_n\rho'$. Elements of S_{n-2} commute with τ_n which implies $\tau_n=\rho\rho'\tau_n$, and thus $\rho'=\rho^{-1}$. We conclude that if (VII.1) holds, there exists $\rho\in S_{n-2}$ with $\sigma=\sigma_0\rho$ and $\sigma'=\rho^{-1}\sigma_0'$. Clearly ρ is unique. If $\sigma=\sigma_0\rho$ and $\sigma'=\rho^{-1}\sigma_0'$ with $\rho\in S_{n-2}$, then (VIII.1) holds because elements of S_{n-2} commute with τ_n. Since the order of S_{n-2} is $(n-2)!$, result (b) is proved.

To establish the main part of the lemma we use a counting argument. From (b), $S_{n-1}\tau_nS_{n-1}$ has $[(n-1)!^2/(n-2)!]=(n-1)(n-1)!$ elements. Adding this number to the order of S_{n-1}, $(n-1)!$ gives $n!$, the order of S_n. The disjointness of S_{n-1} and $S_{n-1}\tau_nS_{n-1}$ gives the result

Next we recall the concept of starring. If r is a tableau of frame α for S_n, then r^* is the tableau obtained from r by deleting the box containing the number n. By $\alpha^*(r)$ we mean the frame, for S_{n-1}, to which r^* corresponds. The following result proved by Rutherford[17] relates the matrices u for α and α^*:

Theorem VII.1: If $\tau\in S_{n-1}$, then

$$u_{rs}^{\alpha}(\tau)=\begin{cases}u_{r^*s^*}^{\alpha^*(r)}(\tau), & \text{if } j_n(r)=j_n(s),\\ 0, & \text{otherwise.}\end{cases}$$

The iterative formula and its proof may now be given.

Theorem VII.2: If α_0,α, and α' are frames for S_n then

$$\begin{aligned}&A_{rs}^{\alpha_0}(p,q,\alpha;p',q',\alpha')\\ &\quad=\frac{\theta^{\alpha_0^*(r)}}{\theta^{\alpha_0}}\tilde{A}_{rs}^{\alpha_0}(p,q,\alpha;p',q',\alpha')\\ &\qquad+\left(\frac{\theta^{\alpha_0^*(r)}\theta^{\alpha_0^*(s)}}{(n-2)!\,\theta^{\alpha_0}}\right)\tilde{\tilde{A}}_{rs}^{\alpha_0}(p,q,\alpha;p',q',\alpha'),\end{aligned} \tag{VII.2}$$

where

$$\tilde{A}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')$$

$$=A^{\alpha_0^*(r)}_{r^*s^*}(p^*, q^*, \alpha^*(p); p'^*, q'^*, \alpha'^*(p')),$$

when $j_n(r)=j_n(s),\ j_n(p)=j_n(q),\ j_n(p')=j_n(q')$,

$$\tilde{A}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')=0 \qquad \text{(VII.3)}$$

otherwise, and

$$\tilde{\tilde{A}}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')$$

$$=\sum A^{\alpha_0^*(r)}_{r^*i_2^*}(p^*, k_1^*, \alpha^*(p); p'^*, m_1^*, \alpha'^*(p'))$$

$$\times A^{\alpha_0^*(s)}_{i_1^*s^*}(k_2^*, q^*, \alpha^*(q); m_2^*, q'^*, \alpha'^*(q'))$$

$$\times u^{\alpha_0}_{i_1i_2}(\tau_n)u^{\alpha}_{k_1k_2}(\tau_n)u^{\alpha'}_{m_1m_2}(\tau_n); \qquad \text{(VII.4)}$$

the sum is taken over only those values of the indices for which $j_n(i_1)=j_n(s),\ j_n(i_2)=j_n(r),\ j_n(k_1)=j_n(p),\ j_n(k_2)=j_n(q),\ j_n(m_1)=j_n(p'),\ j_n(m_2)=j_n(q')$.

Proof of Theorem VII.2: We use the decomposition given by the lemma. First, if $\sigma\in S_{n-1}$, then the product is

$$u^{\alpha_0}_{sr}(\sigma^{-1})u^{\alpha}_{pq}(\sigma)u^{\alpha'}_{p'q'}(\sigma)$$

$$=u^{\alpha_0^*(r)}_{s^*r^*}(\sigma^{-1})u^{\alpha^*(p)}_{p^*q^*}(\sigma)u^{\alpha'^*(p')}_{p'^*q'^*}(\sigma)$$

if $j_n(r)=j_n(s),\ j_n(p)=j_n(q),\ j_n(p')=j_n(q')$, and is otherwise zero. Thus, if the appropriate j's are equal

$$\sum_{\sigma\in S_{n-1}} u^{\alpha_0}_{sr}(\sigma^{-1})u^{\alpha}_{pq}(\sigma)u^{\alpha'}_{p'q'}(\sigma)$$

$$=\theta^{\alpha_0^*(r)}A^{\alpha_0^*(r)}_{r^*s^*}(p^*, q^*, \alpha^*(p); p'^*, q'^*, \alpha'^*(p')),$$

by the definition (IV.8), and the sum is zero if not all the pairs of j's are equal.

Next, we let $\sigma,\sigma'\in S_{n-1}$ and we consider $\sigma\tau_n\sigma'$. We have $(\sigma\tau_n\sigma')^{-1}=\sigma'^{-1}\tau_n\sigma^{-1}$ and

$$u_{sr}(\sigma^{-1}\tau_n\sigma'^{-1})u_{pq}(\sigma\tau_n\sigma')u_{p'q'}(\sigma\tau_n\sigma')$$

$$=\sum_{\substack{i_1,i_2\\k_1,k_2\\m_1,m_2}} u^{\alpha_0}_{si_1}(\sigma'^{-1})u^{\alpha_0}_{i_1i_2}(\tau_n)u^{\alpha_0}_{i_2r}(\sigma^{-1})$$

$$\times u^{\alpha}_{pk_1}(\sigma)u^{\alpha}_{k_1k_2}(\tau_n)u^{\alpha}_{k_2q}(\sigma')$$

$$\times u^{\alpha'}_{p'm_1}(\sigma)u^{\alpha'}_{m_1m_2}(\tau_n)u^{\alpha'}_{m_2q'}(\sigma').$$

The terms of the sum can be rearranged and the sum can be taken over fewer values of the indices so that it becomes

$$\sum u^{\alpha_0^*(r)}_{i_2^*r^*}(\sigma^{-1})u^{\alpha^*(p)}_{p^*k_1^*}(\sigma)u^{\alpha'^*(p')}_{p'^*m_1^*}(\sigma)$$

$$\times u^{\alpha_0^*(s)}_{s^*i_1^*}(\sigma'^{-1})u^{\alpha^*(q)}_{k_2^*q^*}(\sigma')u^{\alpha'^*(q')}_{m_2^*q'^*}(\sigma')$$

$$\times u^{\alpha_0}_{i_1i_2}(\tau_n)u^{\alpha}_{k_1k_2}(\tau_n)u^{\alpha'}_{m_1m_2}(\tau_n),$$

where the indices range over only those tableaux $i_1, i_2, k_1, k_2, m_1, m_2$ with $j_n(i_1)=j_n(s),\ j_n(i_2)=j_n(r),\ j_n(k_1)=j_n(p),\ j_n(k_2)=j_n(q),\ j_n(m_1)=j_n(p'),\ j_n(m_2)=j_n(q')$. If we now sum this over $\sigma,\sigma'\in S_{n-1}$, applying (IV.8), we obtain $\tilde{\tilde{A}}$.

By Lemma VII.1 the sum in (IV.8) splits into two parts, giving

$$A^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')$$

$$=\frac{\theta^{\alpha_0^*(r)}}{\theta^{\alpha_0}}\tilde{A}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')$$

$$+\frac{\theta^{\alpha_0^*(r)}\theta^{\alpha_0^*(s)}}{\theta^{\alpha_0}(n-2)!}\tilde{\tilde{A}}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')$$

as required.

We note that a condition requiring the j_n's to agree for two tableaux simply means that the number n must be in the same row for both tableaux.

VIII. THE SYMMETRIES

For the orthogonal representation several symmetry properties of the $\mathcal{A}$'s follow immediately from the definition (IV.11) and the unitarity of this representation. For the seminormal representation, corresponding properties may be obtained from those below by using (IV.14), and need not be stated explicitly here.

First,

$$\mathcal{A}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')=\mathcal{A}^{\alpha_0}_{sr}(q, p, \alpha; q', p', \alpha'). \qquad \text{(VIII.1)}$$

Also,

$$\theta^{\alpha_0}\mathcal{A}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha')$$

$$=\theta^{\alpha_0}\mathcal{A}^{\alpha_0}_{rs}(p', q', \alpha'; p, q, \alpha)$$

$$=\theta^{\alpha'}\mathcal{A}^{\alpha'}_{p'q'}(r, s, \alpha_0; p, q, \alpha), \qquad \text{(VIII.2)}$$

and so on. Further symmetry properties can be obtained by combining formulas (VIII.1) and (VIII.2) in an obvious way.

We recall that for any r the matrix $\mathcal{A}^{\alpha_0}_{rr}$, discussed earlier, is a projection. Hence, its trace equals its rank which we found was $m(\alpha_0, \alpha, \alpha')=m(\alpha_0)$. This now gives a proof that the multiplicity is independent of the order of the three classes $\alpha_0, \alpha, \alpha'$: We have

$$m(\alpha_0, \alpha, \alpha')=\sum_{p,p'}\mathcal{A}^{\alpha_0}_{rr}(p, p, \alpha; p', p', \alpha'), \qquad \text{(VIII.3)}$$

so that summing on r gives

$$\theta^{\alpha_0}f^{\alpha_0}m(\alpha_0, \alpha, \alpha')=\sum_{p,p',r}\theta^{\alpha_0}\mathcal{A}^{\alpha_0}_{rr}(p, p, \alpha; p', p', \alpha')$$

$$=\sum_{p,p',r}\theta^{\alpha}\mathcal{A}^{\alpha}_{pp}(r, r, \alpha_0; p', p', \alpha'),$$

by (VIII.2) and so, by (III.8),

$$n!\,m(\alpha_0, \alpha, \alpha')=\sum_{p,p',r}\theta^{\alpha}\mathcal{A}^{\alpha}_{pp}(r, r, \alpha_0; p', p', \alpha')$$

$$=n!\,m(\alpha, \alpha_0, \alpha'), \qquad \text{(VIII.4)}$$

as required. Clearly, also $m(\alpha_0, \alpha, \alpha')=m(\alpha_0, \alpha', \alpha)$.

The following constructions expand the discussions of Hamermesh.[18]

To find the symmetries of the Clebsch–Gordan coefficients we consider (V.6) and (V.7), and note that the numbers $C^{\alpha_0,\alpha',\alpha}_{\lambda r}(k', k)$ satisfy the same equations as the numbers $C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(k, k')$. (Here we must write the symbols α and α' on the C's; elsewhere these are suppressed.) We recall that the C's are not uniquely defined; any set of numbers satisfying (V.6) and (VI.3) is acceptable.

We order all the representations in some manner, and we solve for the coefficients $C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(k,k')$ for $\alpha \leqslant \alpha'$. The $\mathcal{A}$'s obey (VIII.2), so coefficients satisfying (V.6) and (VI.3) may be defined by

$$C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(k,k') = C^{\alpha_0,\alpha',\alpha}_{\lambda r}(k',k), \qquad \text{(VIII.5)}$$

for $\alpha' < \alpha$. The resulting coefficients are automatically symmetric.

For the symmetry arising from an interchange of the representations α_0 and α', we start from formula (V.7), which we rewrite as

$$\mu^{\alpha}_{pq}(\tau)\mu^{\alpha'}_{p'q'}(\tau)$$
$$= \sum_{\alpha_0,\lambda,r,r'} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')\mu^{\alpha_0}_{rr'}(\tau)C^{\alpha_0,\alpha,\alpha'}_{\lambda r'}(q,q'), \qquad \text{(VIII.6)}$$

taking the Clebsch–Gordan coefficients to be real. We multiply this equation by $C^{\tilde{\alpha}_0,\alpha,\alpha'}_{\tilde{\lambda}\tilde{r}}(q,q')$ and sum on q, q', using (V.6) and obtain

$$\sum_{q,q'} \mu^{\alpha}_{pq}(\tau)\mu^{\alpha'}_{p'q'}(\tau)C^{\alpha_0,\alpha,\alpha'}_{\lambda r'}(q,q')$$
$$= \sum_{r} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')\mu^{\alpha_0}_{rr'}(\tau). \qquad \text{(VIII.7)}$$

(For simplicity, we will continually drop the "~"'s.) Then, by unitarity, for any $\tilde{q}$,

$$\sum_{p} \mu^{\alpha}_{p\tilde{q}}(\tau)\mu^{\alpha}_{pq}(\tau) = \sum_{p} \mu_{\tilde{q}p}(\tau^{-1})\mu_{pq}(\tau)$$
$$= \mu^{\alpha}_{\tilde{q}q}(\epsilon) = \delta_{\tilde{q}q},$$

so multiplying (VIII.7) by $\mu^{\alpha}_{p\tilde{q}}(\tau)$ and summing on p, we have

$$\sum_{q'} \mu^{\alpha'}_{p'q'}(\tau)C^{\alpha_0,\alpha,\alpha'}_{\lambda r'}(q,q')$$
$$= \sum_{r,p} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')\mu^{\alpha_0}_{rr'}(\tau)\mu^{\alpha}_{pq}(\tau). \qquad \text{(VIII.8)}$$

Next, expanding $\mu^{\alpha_0}_{rr'}(\tau)\mu^{\alpha}_{pq}(\tau) = \mu^{\alpha}_{pq}(\tau)\mu^{\alpha_0}_{rr'}(\tau)$, using (V.7), gives

$$\sum_{q'} \mu^{\alpha'}_{p'q'}(\tau)C^{\alpha_0,\alpha,\alpha'}_{\lambda r'}(q,q')$$
$$= \sum_{\substack{r,p\\ t,t'\\ \beta,\nu}} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')C^{\beta,\alpha,\alpha_0}_{\nu t}(p,r)$$
$$\times \mu^{\beta}_{tt'}(\tau)C^{\beta,\alpha,\alpha_0}_{\nu t'}(q,r'). \qquad \text{(VIII.9)}$$

To move the last C over we multiply by $C^{\tilde{\beta},\alpha,\alpha_0}_{\tilde{\nu}\tilde{t}'}(q,r')$ and sum on q and r', so that

$$\sum_{r',q,q'} \mu^{\alpha'}_{p'q'}(\tau)C^{\alpha_0,\alpha,\alpha'}_{\lambda r'}(q,q')C^{\beta,\alpha,\alpha_0}_{\nu t'}(q,r')$$
$$= \sum_{r,p,t} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')C^{\beta,\alpha,\alpha_0}_{\nu t}(p,r)\mu^{\beta}_{tt'}(\tau). \qquad \text{(VIII.10)}$$

Now $\sum_{r,q} C^{\alpha_0,\alpha,\alpha'}_{\lambda r'}(q,q')C^{\beta,\alpha,\alpha_0}_{\nu t'}(q,r')$ may be viewed as the entry in row q, column t' of a matrix

$$\mathcal{K} = \mathcal{K}\begin{pmatrix}\alpha_0,\alpha,\alpha',\beta\\ \lambda,\nu\end{pmatrix}$$

and

$$\sum_{r,p} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')C^{\beta,\alpha,\alpha_0}_{\nu t}(p,r)$$

gives the row p' column t entry of the same matrix, and (VIII.10) states that

$$\mu^{\alpha'}\mathcal{K} = \mathcal{K}\mu^{\beta}.$$

By Schur's lemma,

$$\mathcal{K} = 0 \quad\text{or}\quad \mathcal{K}\begin{pmatrix}\alpha_0,\alpha,\alpha',\beta\\ \lambda,\nu\end{pmatrix} = K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}I$$

(I is the unit matrix of appropriate dimension) and representations α' and β must be equivalent, that is,

$$K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}$$

is a scalar such that

$$\sum_{r,p} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')C^{\beta,\alpha,\alpha_0}_{\nu t}(p,r)$$
$$= K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}\delta^{\alpha'\beta}\delta_{p't}. \qquad \text{(VIII.11)}$$

Next we see that

$$\sum_{\nu} C^{\alpha',\alpha,\alpha_0}_{\nu p'}(p,r)C^{\alpha',\alpha,\alpha_0}_{\nu\tilde{t}}(p,r)$$
$$= \mathcal{A}^{\alpha'}_{p'\tilde{t}}(r,\tilde{r},\alpha_0;p,\tilde{p},\alpha)$$
$$= \frac{\theta^{\alpha_0}}{\theta^{\alpha'}}\mathcal{A}^{\alpha_0}_{r\tilde{r}}(p,p,\alpha;p',t,\alpha') \qquad \text{(VIII.12)}$$

by (VI.3) and (VIII.2). Then an interchange of p' and t in (VIII.11), multiplication by $C^{\alpha',\alpha,\alpha_0}_{\nu\tilde{t}}(\tilde{p},\tilde{r})$, and summation on ν gives

$$\sum_{\nu} K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}\delta_{p't}C^{\alpha',\alpha,\alpha_0}_{\nu\tilde{t}}(\tilde{p},\tilde{r})$$
$$= \sum_{r,p,\nu} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,t)C^{\alpha',\alpha,\alpha_0}_{\nu p'}(p,r)C^{\alpha',\alpha,\alpha_0}_{\nu\tilde{t}}(\tilde{p},\tilde{r})$$
$$= \sum_{r,p} \frac{\theta^{\alpha_0}}{\theta^{\alpha'}} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,t)\mathcal{A}^{\alpha_0}_{r\tilde{r}}(p,\tilde{p},\alpha;p',\tilde{t},\alpha'). \qquad \text{(VIII.13)}$$

Taking $p' = t$ and summing on p' we see that

$$\sum_{r,p,p'} \left(\frac{\theta^{\alpha_0}}{\theta^{\alpha'}}\right) C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')\mathcal{A}^{\alpha_0}_{r\tilde{r}}(p,\tilde{p},\alpha;p',\tilde{t},\alpha')$$
$$= \sum_{r,p,p'} \left(\frac{\theta^{\alpha_0}}{\theta^{\alpha'}}\right)\mathcal{A}^{\alpha_0}_{\tilde{r}r}(\tilde{p},p,\alpha;\tilde{t},p',\alpha')C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')$$
$$= \sum_{r} \left(\frac{\theta^{\alpha_0}}{\theta^{\alpha'}}\right) C^{\alpha_0,\alpha,\alpha'}_{\lambda\tilde{r}}(\tilde{p},\tilde{t}), \qquad \text{(VIII.14)}$$

by (VI.4). The terms in the sum on r are independent of r so that it becomes

$$f^{\alpha_0}\left(\frac{\theta^{\alpha_0}}{\theta^{\alpha'}}\right) C^{\alpha_0,\alpha,\alpha'}_{\lambda\tilde{r}}(\tilde{p},\tilde{t}) = \frac{n!}{\theta^{\alpha'}} C^{\alpha_0,\alpha,\alpha'}_{\lambda\tilde{r}}(\tilde{p},\tilde{t})$$
$$= f^{\alpha'}C^{\alpha_0,\alpha,\alpha'}_{\lambda\tilde{r}}(\tilde{p},\tilde{t}),$$

by (III.8). Hence

$$f^{\alpha'}C^{\alpha_0,\alpha,\alpha'}_{\lambda\tilde{r}}(\tilde{p},\tilde{t}) = \sum_{\nu,p'} K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}C^{\alpha',\alpha,\alpha_0}_{\nu\tilde{t}}(\tilde{p},\tilde{r})$$
$$= f^{\alpha'}\sum_{\nu} K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}C^{\alpha',\alpha,\alpha_0}_{\nu\tilde{t}}(\tilde{p},\tilde{r}).$$
$$\text{(VIII.15)}$$

Also, from (V.6) and (VIII.15)

$$
\begin{aligned}
f^{\alpha_0}\delta_{\lambda\mu} &= \sum_{\tilde r}\delta_{\lambda\mu}\\
&= \sum_{\tilde r}\sum_{\tilde p,\tilde t} C^{\alpha_0,\alpha,\alpha'}_{\lambda\tilde r}(\tilde p,\tilde t)C^{\alpha_0,\alpha,\alpha'}_{\mu\tilde r}(\tilde p,\tilde t)\\
&= \sum_{\substack{\tilde p,\tilde t,\tilde r\\ \nu,\nu'}} K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu'\end{pmatrix}\\
&\quad\times C^{\alpha',\alpha,\alpha_0}_{\nu\tilde t}(\tilde p,\tilde r)C^{\alpha',\alpha,\alpha_0}_{\nu'\tilde t}(\tilde p,\tilde r)\\
&= \sum_{\nu,\tilde t} K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda,\nu\end{pmatrix}K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \mu\nu\end{pmatrix}\\
&= f^{\alpha'}\sum_{\nu} K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \lambda\nu\end{pmatrix}K\begin{pmatrix}\alpha_0,\alpha,\alpha'\\ \mu\nu\end{pmatrix}. \qquad \text{(VIII. 16)}
\end{aligned}
$$

From this relation and (VIII.15) it follows that

$$
\begin{aligned}
&\sum_{p,p'} C^{\alpha',\alpha,\alpha_0}_{\lambda p'}(p,r)C^{\alpha',\alpha,\beta_0}_{\mu p'}(p,s)\\
&\quad= \sum_{\substack{p,p'\\ \nu,\nu'}} K\begin{pmatrix}\alpha',\alpha,\alpha_0\\ \lambda\nu\end{pmatrix}K\begin{pmatrix}\alpha',\alpha,\beta_0\\ \mu\nu'\end{pmatrix}\\
&\qquad\times C^{\alpha_0,\alpha,\alpha'}_{\nu r}(p,p')C^{\beta_0,\alpha,\alpha'}_{\nu' s}(p,p')\\
&\quad= \delta^{\alpha_0\beta_0}\sum_{\nu} K\begin{pmatrix}\alpha',\alpha,\alpha_0\\ \lambda\nu\end{pmatrix}K\begin{pmatrix}\alpha',\alpha,\alpha_0\\ \mu\nu\end{pmatrix}\delta_{rs}\delta_{\lambda\mu}\\
&\quad= \frac{f^{\alpha'}}{f^{\alpha_0}}\delta^{\beta_0,\alpha_0}\delta_{\lambda\mu}\delta_{rs}. \qquad \text{(VIII. 17)}
\end{aligned}
$$

Moreover, by (V.5) and (VIII.15),

$$
\begin{aligned}
&\sum_{p,p'} C^{\alpha',\alpha,\alpha_0}_{\lambda p'}(p,r)C^{\beta_0,\alpha,\alpha'}_{\mu s}(p,p')\\
&\quad= \sum_{p,p',\nu} K\begin{pmatrix}\alpha',\alpha,\alpha_0\\ \lambda\nu\end{pmatrix}C^{\alpha_0,\alpha,\alpha'}_{\nu r}(p,p')C^{\beta_0,\alpha,\alpha'}_{\mu s}(p,p')\\
&\quad= K\begin{pmatrix}\alpha',\alpha,\alpha_0\\ \lambda\mu\end{pmatrix}\delta_{rs}\delta^{\alpha_0\beta_0}. \qquad \text{(VIII. 18)}
\end{aligned}
$$

Now suppose we have found coefficients $C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')$ satisfying (V.6) and (VI.3), for $\alpha' \leqslant \alpha_0$. Then for $\alpha_0 < \alpha'$, we define

$$
C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p') = (f^{\alpha_0}/f^{\alpha'})^{1/2}C^{\alpha',\alpha,\alpha_0}_{\lambda p'}(p,r). \qquad \text{(VIII. 19)}
$$

Note that the multiplicity indices for both triplets of representations are independent of the order in which the representations occur. We must show these coefficients satisfy (VI.3) and (V.6) in all cases.

First,

$$
\begin{aligned}
&\sum_{\lambda} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')C^{\alpha_0,\alpha,\alpha'}_{\lambda s}(q,q')\\
&\quad= \frac{f^{\alpha_0}}{f^{\alpha'}}\sum_{\lambda} C^{\alpha',\alpha,\alpha_0}_{\lambda p'}(p,r)C^{\alpha',\alpha,\alpha_0}_{\lambda q'}(q,s)\\
&\quad= \frac{f^{\alpha_0}}{f^{\alpha'}}\mathcal{A}^{\alpha'}_{p'q'}(p,q,\alpha;r,s,\alpha_0)\\
&\quad= \frac{f^{\alpha_0}}{f^{\alpha'}}\frac{\theta^{\alpha_0}}{\theta^{\alpha'}}\mathcal{A}^{\alpha_0}_{rs}(p,q,\alpha;p',q',\alpha')\\
&\quad= \mathcal{A}^{\alpha_0}_{rs}(p,q,\alpha;p',q',\alpha'),
\end{aligned}
$$

by (VIII.2). Next

$$
\begin{aligned}
&\sum_{p,p'} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')C^{\alpha_0,\alpha,\alpha'}_{\mu s}(p,p')\\
&\quad= \frac{f^{\alpha_0}}{f^{\alpha'}}\sum_{p,p'} C^{\alpha',\alpha,\alpha_0}_{\lambda p'}(p,r)C^{\alpha',\alpha,\alpha_0}_{\mu p'}(p,s) = \delta_{\lambda\mu}\delta_{rs}
\end{aligned}
$$

by (V.6). We must also show that $\alpha_0 \neq \beta_0$ implies that

$$
\sum_{p,p'} C^{\alpha_0,\alpha,\alpha'}_{\lambda r}(p,p')C^{\beta_0,\alpha,\alpha'}_{\mu s}(p,p') = 0.
$$

If $\alpha' \leqslant \alpha_0, \beta_0$ we already know this formula holds. In the cases $\alpha_0 < \alpha' \leqslant \beta_0$, or $\beta_0 < \alpha' \leqslant \alpha_0$, we use the definition (VIII.19) and (VIII.18). For $\alpha_0, \beta_0 < \alpha'$, (VIII.19) and (VIII.17) are used.

Interchanging representations and their conjugates gives further symmetries. We recall (III.4) and (III.6). The completely symmetric (one-dimensional) representation is labeled by

$$
\alpha^+ = (n, 0, \ldots, 0) \qquad \text{(VIII. 20)}
$$

and the completely antisymmetric (one-dimensional) representation by

$$
\alpha^- = (1, 1, \ldots, 1). \qquad \text{(VIII. 21)}
$$

For any frame α, we fix a tableau p_0. If p is another tableau of this frame, then p may be obtained from p_0 by applying some permutation $\sigma_{pp_0} \in S_n$ on the numbers of p_0. We let $\Lambda(p) = \operatorname{sign}(\sigma_{pp_0})$. The symbol "$\longleftrightarrow$" will indicate a mapping which gives an isomorphism preserving the action of S_n. We will also use the associativity and commutativity of the tensor product,

$$
(V_1 \otimes V_2) \otimes V_3 \longleftrightarrow V_1 \otimes (V_2 \otimes V_3),
$$

and

$$
V_1 \otimes V_2 \longleftrightarrow V_2 \otimes V_1.
$$

Our first isomorphism is

$$
\xi^{\alpha}_{pq} \otimes \xi^{\alpha^+} \longleftrightarrow \xi^{\alpha}_{pq}. \qquad \text{(VIII. 22)}
$$

The argument is that if τ_j is any neighboring transposition, then $\tau_j\xi^{\alpha^+} = \xi^{\alpha^+}$, and thus, for any $\tau \in S_n$,

$$
\tau\xi^{\alpha^+} = \xi^{\alpha^+}. \qquad \text{(VIII. 23)}
$$

Hence, for $\tau \in S_n$,

$$
\begin{aligned}
\tau(\xi^{\alpha}_{pq} \otimes \xi^{\alpha^+}) &= \tau\xi^{\alpha}_{pq} \otimes \tau\xi^{\alpha^+}\\
&= \tau\xi^{\alpha}_{pq} \otimes \xi^{\alpha^+}
\end{aligned}
$$

and "$\longleftrightarrow$" follows.

Since, for τ_j any neighboring transposition, we have $\tau_j\xi^{\alpha^-} = -\xi^{\alpha^-}$, we obtain, for $\tau \in S_n$

$$
\tau\xi^{\alpha^-} = \operatorname{sign}(\tau)\xi^{\alpha^-}. \qquad \text{(VIII. 24)}
$$

It then follows, from (VIII.23) and (VIII.24), that

$$
\xi^{\alpha^-} \otimes \xi^{\alpha^-} \longleftrightarrow \xi^{\alpha^+}. \qquad \text{(VIII. 25)}
$$

We now proceed to relate a representation and its conjugate. We will show

$$
\xi^{\alpha}_{pq} \longleftrightarrow \Lambda(p)[\xi^{\bar\alpha}_{\bar p\bar q} \otimes \xi^{\alpha^-}]. \qquad \text{(VIII. 26)}
$$

To show this isomorphism, we apply any $\tau_j = (j-1, j)$ to both sides and verify that the correspondence of (VIII.26) is preserved.

$$\tau_j \xi^{\alpha}_{pq} = \sum_k \mu^{\alpha}_{kp}(\tau) \xi^{\alpha}_{kq}$$

and

$$\tau_j(\xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\alpha^-}) = -1(\tau_j \xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\alpha^-})$$

$$= -1 \sum_m \mu^{\bar{\alpha}}_{m\bar{p}}(\tau_j)(\xi^{\bar{\alpha}}_{m\bar{q}} \otimes \xi^{\alpha^-})$$

$$= -1 \sum_m \rho(m,p) \mu^{\alpha}_{\bar{m}p}(\tau_j) \xi^{\bar{\alpha}}_{m\bar{q}} \otimes \xi^{\alpha^-}, \qquad \text{(VIII. 27)}$$

where

$$\rho(m,p) = \begin{cases} -1, & m = p, \\ +1, & m \neq p, \end{cases}$$

using (III. 22) and (III. 23). Now τ_j, acting on the left-hand side of (VIII. 26) gives

$$\sum_k \mu^{\alpha}_{kp}(\tau_j) \Lambda(k)(\xi^{\bar{\alpha}}_{\bar{k}\bar{q}} \otimes \xi^{\alpha^-}), \qquad \text{(VIII. 28)}$$

and τ_j acting on the right-hand side yields

$$-1\Lambda(p) \sum_m \rho(m,p) \mu^{\alpha}_{\bar{m}p}(\tau_j) \xi^{\bar{\alpha}}_{m\bar{q}} \otimes \xi^{\alpha^-}. \qquad \text{(VIII. 29)}$$

The expressions (VIII. 28) and (VIII. 29) are equal since

$$-\Lambda(p)\rho(m,p) = \Lambda(m), \qquad \text{(VIII. 30)}$$

whenever $\mu^{\alpha}_{\bar{m}p}(\tau_j) \neq 0$, that is, whenever m arises from p by interchanging the numbers $j-1$ and j. The asserted isomorphism follows.

We combine our results and obtain the following sequence of isomorphisms:

$$\Lambda(p)\Lambda(p')(\xi^{\alpha}_{pq} \otimes \xi^{\alpha'}_{p'q'})$$

$$\longleftrightarrow (\xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\alpha^-}) \otimes (\xi^{\bar{\alpha}'}_{\bar{p}'\bar{q}'} \otimes \xi^{\alpha^-})$$

$$\longleftrightarrow (\xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\bar{\alpha}'}_{\bar{p}'\bar{q}'}) \otimes (\xi^{\alpha^-} \otimes \xi^{\alpha^-})$$

$$\longleftrightarrow (\xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\bar{\alpha}'}_{\bar{p}'\bar{q}'}) \otimes \xi^{\alpha^+}$$

$$\longleftrightarrow \xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\bar{\alpha}'}_{\bar{p}'\bar{q}'},$$

that is

$$\Lambda(p)\Lambda(p')(\xi^{\alpha}_{pq} \otimes \xi^{\alpha'}_{p'q'}) \longleftrightarrow \xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\bar{\alpha}'}_{\bar{p}'\bar{q}'}. \qquad \text{(VIII. 31)}$$

Each equivalence in the preceding sequence of isomorphisms is unitary, thus (VIII. 31) also gives a unitary equivalence.

Now suppose the Clebsch–Gordan coefficients $C^{\alpha_0, \alpha, \alpha'}$ have been computed for each α_0, for $\alpha \leqslant \alpha'$. Then

$$\xi^{\alpha}_{pq} \otimes \xi^{\alpha'}_{p'q'} \longleftrightarrow \sum C^{\alpha_0, \alpha, \alpha'}_{\lambda r}(p,p') \xi^{\alpha_0}_{rs}$$

and thus, by (VIII. 31),

$$\xi^{\bar{\alpha}}_{\bar{p}\bar{q}} \otimes \xi^{\bar{\alpha}'}_{\bar{p}'\bar{q}'} \longleftrightarrow \sum \Lambda(p)\Lambda(p') C^{\alpha_0, \alpha, \alpha'}_{\lambda r}(p,p') \xi^{\alpha_0}_{rs}; \qquad \text{(VIII. 32)}$$

both isomorphisms are unitary.

We then define, for $\alpha' < \alpha$,

$$C^{\alpha_0, \alpha, \alpha'}_{\lambda r}(p,p') = \Lambda(\bar{p})\Lambda(\bar{p}') C^{\alpha_0, \bar{\alpha}, \bar{\alpha}'}_{\lambda r}(\bar{p}, \bar{p}'). \qquad \text{(VIII. 33)}$$

Since (VIII. 31) does give a unitary equivalence, these numbers are Clebsch–Gordan coefficients in the case $\alpha' < \alpha$.

We have seen that starting with certain subsets of Clebsch–Gordan coefficients, it is possible to construct the other coefficients such that the total set has the symmetry properties of a particular class. The different procedures do not commute so that not all the symmetries will hold simultaneously.

An example showing that all classes of symmetries cannot be made to hold simultaneously is given by a triplet $\alpha_0, \alpha, \bar{\alpha}$. For interchange symmetry, we must have $C^{\alpha_0, \alpha, \bar{\alpha}}_{\lambda r}(p,p') = C^{\alpha_0, \bar{\alpha}, \alpha}_{\lambda r}(\bar{p}', p)$, for conjugation symmetry, $C^{\alpha_0, \alpha, \bar{\alpha}}_{\lambda r}(p, \bar{p}') = \Lambda(p)\Lambda(p') C^{\alpha_0, \bar{\alpha}, \alpha}_{\lambda r}(\bar{p}, p')$. These coefficients are not equal unless α is one-dimensional.

We will later describe a consistent procedure for using all the symmetry related constructions to reduce the number of coefficients that need to be calculated.

To find the symmetries of the $\mathcal{A}$'s under conjugation, we use (VI. 3) and get

$$\mathcal{A}^{\alpha_0}_{rs}(\bar{p}, \bar{q}, \bar{\alpha}; \bar{p}', \bar{q}', \bar{\alpha}')$$

$$= \Lambda(p)\Lambda(p')\Lambda(q)\Lambda(q') \mathcal{A}^{\alpha_0}_{rs}(p, q, \alpha; p', q', \alpha'). \qquad \text{(VIII. 34)}$$

Using the multiplicity formula (VIII. 3), it immediately follows that the multiplicity is invariant under conjugation of α and α'. Using the interchange symmetry as well, we see that the multiplicity is invariant under conjugation of any two of the representations.

We cannot construct coefficients satisfying, simultaneously, all the symmetries discussed. However, we can use all our symmetry constructions, in some sequence, to define coefficients, allowing us to compute, simply, large sets of coefficients from just one coefficient. This avoids the need to use detailed calculations for most of them.

We use different types of operations on ordered triplets:

$T_{1,a}$: an interchange of the last two members of the triplet;

$T_{1,b}$: an interchange of the first and last members of the triplet;

T_2: conjugation of the last two members of the triplet.

The triplets of representations split into equivalence classes. Two triplets are said to be equivalent if one can be obtained from the other by any sequence of operations of the above types.

The representations may be ordered. We do this by defining "<" as follows: If $m_k(\alpha) = m_k(\beta)$ for $k > k_0$, but $m_{k_0}(\alpha) > m_{k_0}(\beta)$ then $\alpha < \beta$. We divide the representation into three sets. The first contains the self-conjugate representations. The second is given by $\{\alpha : \alpha < \bar{\alpha}\}$ and the third by $\{\alpha : \bar{\alpha} < \alpha\}$.

In each class of triplets we use the methods of Secs. VI and VII to compute all the coefficients for one triplet, the "working triplet." The coefficients for the other triplets will then be obtainable using constructions going with operations of types $T_{1,a}$, $T_{1,b}$, and T_2. To select the working triplet $(\alpha_0, \alpha, \alpha')$, whose coefficients $C^{\alpha_0, \alpha, \alpha'}$ are to be computed, we consider all triplets $(\alpha_0, \alpha, \alpha')$ for which $\alpha_0 \leqslant \alpha \leqslant \alpha'$. Of these we pick out those for which α_0 is minimum with respect to the ordering. From that collection we pick out ones with α the smallest, and finally, we pick out the one with α' the smallest. This triplet is unique. We compute all the coefficients for it.

To find the coefficients for any triplet in an equivalence class, we first list, in order, the operations which transform the required triplet into the working triplet. Then we apply the constructions corresponding to the operations, but in reverse order, to the coefficients of the working triplet. This chain of constructions then gives the coefficients for the required triplet. The reason we use this simple reversal is that each of the relevant operations is its own inverse.

There are a number of suitable chains of operations; we discuss one. Suppose the required triplet is (β_0, β, β'). We tabulate the steps necessary to give the working triplet $(\alpha_0, \alpha, \alpha')$.

Triplet	Conditions	Operations (where necessary)	Result
(β_0, β, β')	$\alpha_0 = \min\left\{\begin{matrix}\beta_0, \beta, \beta' \\ \bar\beta_0, \bar\beta, \bar\beta'\end{matrix}\right\}$		
	I. (a) α_0 in triplet	$T_{1,b}$	$- - \alpha_0,$
		$T_{1,a}$	$\alpha_0 - - .$
	(b) α_0 not in triplet	$T_{1,a}$ or $T_{1,b}$	$- - \bar\alpha_0,$
		T_2	$- - \alpha_0,$
		$T_{1,b}$	$\alpha_0 - - .$

In either case, we obtain a triplet denoted by:

Triplet	Conditions	Operations (where necessary)	Result
$(\alpha_0, \gamma, \gamma')$	$\alpha = \min\left\{\begin{matrix}\gamma, \gamma' \\ \bar\gamma, \bar\gamma'\end{matrix}\right\}$		
	II. (a) α in pair γ, γ'	$T_{1,a}$	$\alpha_0, \alpha, -.$
	(b) α not in pair γ, γ'	$T_{1,a}$	$\alpha_0, \bar\alpha, -,$
		T_2	$\alpha_0, \alpha, -.$

Either case leads to a triplet:

Triplet	Conditions	Operations (where necessary)	Result
(α_0, α, ρ)			
	III. (a) $\alpha \neq \bar\alpha$	Let $\alpha' = \rho$.	$\alpha_0, \alpha, \alpha'.$
	(b) $\alpha = \bar\alpha$	Let $\alpha' = \min\{\rho, \bar\rho\}$ and T_2.	$\alpha_0, \alpha, \alpha'.$

Since the coefficients are not unique, we would expect different chains of operations to give different coefficients.

Finally we note that there is no symmetry corresponding to the conjugation of all three representations of a triplet. In fact, of the two coefficients $C_{\lambda r}^{\alpha_0, \alpha, \alpha'}(p, p')$ and $C_{\lambda r}^{\bar\alpha_0, \bar\alpha, \bar\alpha'}(\bar p, \bar p')$, at most one is nonzero. We see this from (VIII. 7) with $\tau = (12)$, giving

$$\mu_{pp}^{\alpha}(12)\mu_{p'p'}^{\alpha'}(12)C_{\lambda r}^{\alpha_0, \alpha, \alpha'}(p, p') \tag{VIII. 35}$$

$$= \mu_{rr}^{\alpha_0}(12)C_{\lambda r}^{\alpha_0, \alpha, \alpha'}(p, p').$$

If an odd number of p, p', r contain the numbers 1 and 2 in the same column, $C_{\lambda r}^{\alpha_0, \alpha, \alpha'}(p, p') = -C_{\lambda r}^{\alpha_0, \alpha, \alpha'}(p, p')$, so this coefficient is zero. Under conjugation, rows and columns are interchanged, so the result follows.

IX. THE HOMOGENEOUS EQUATIONS

The tensor coupling coefficients satisfy two systems of homogeneous linear equations, which are of some use. We now derive these, using the symbols for the orthogonal representation. The formulas in the seminormal case are the same.

We start with (IV. 12),

$$\xi_{rs}^{\alpha_0}(xy)$$

$$= \sum_{\substack{p, q, \alpha \\ p', q', \alpha'}} \mathcal{A}_{rs}^{\alpha_0}(p, q, \alpha; p', q', \alpha')\xi_{pq}^{\alpha}(x)\xi_{p'q'}^{\alpha'}(y),$$

where $x = x_1 x_2 \cdots x_n$, and $y = y_1 y_2 \cdots y_n$. We can replace x by $\tau x = x_{\tau(1)} \cdots x_{\tau(n)}$, and y by τy, giving

$$\xi_{rs}^{\alpha_0}(\tau x \tau y) = \sum_{\substack{p, q, \alpha \\ p', q', \alpha'}} \mathcal{A}_{rs}^{\alpha_0}(p, q, \alpha; p', q', \alpha')$$

$$\times \xi_{pq}^{\alpha}(\tau x)\xi_{p'q'}^{\alpha'}(\tau y).$$

Since

$$\xi_{pq}^{\alpha}\tau = \sum_{\beta, k, j} \xi_{pq}^{\alpha}\mu_{kj}^{\beta}(\tau)\xi_{kj}^{\beta} = \sum_{j} \mu_{qj}^{\alpha}(\tau)\xi_{pj}^{\alpha}$$

we get

$$\xi_{rs}^{\alpha_0}(\tau x \tau y) = \sum_{\substack{p, q, \alpha \\ p', q', \alpha' \\ j, j'}} \mathcal{A}_{rs}^{\alpha_0}(p, q, \alpha; p', q', \alpha')$$

$$\times \mu_{qj}^{\alpha}(\tau)\mu_{q'j'}^{\alpha'}(\tau)\xi_{pj}^{\alpha}(x)\xi_{p'j'}^{\alpha'}(y). \tag{IX. 1}$$

Also

$$\xi^{\alpha_0}_{rs}(\tau x \tau y) = \xi^{\alpha_0}_{rs}(\tau(xy)) = \sum_k \mu^{\alpha_0}_{sk}(\tau)\xi^{\alpha_0}_{rk}(xy)$$

$$= \sum_{\substack{k,p,q,\alpha\\ p',q',\alpha'}} \mu^{\alpha_0}_{sk}(\tau)\mathcal{A}^{\alpha_0}_{rk}(p,q,\alpha;p',q',\alpha')$$

$$\times \xi^{\alpha}_{pq}(x)\,\xi^{\alpha'}_{p'q'}(y).$$

But the products of the ξ's are independent, so for each p, q, p', q', r, s we get, by interchanging j and q, j' and q' in (IX.1),

$$\sum_{j,j'} \mathcal{A}^{\alpha_0}_{rs}(p,j,\alpha;p',j',\alpha')\mu^{\alpha}_{jq}(\tau)\mu^{\alpha'}_{j'q'}(\tau)$$

$$= \sum_k \mathcal{A}^{\alpha_0}_{rk}(p,q,\alpha;p',q',\alpha')\mu^{\alpha_0}_{sk}(\tau). \qquad \text{(IX.2)}$$

We get another set of homogeneous equations by applying the symmetry property (VIII.1) of the $\mathcal{A}$'s and the unitarity of the representation to obtain

$$\sum_{j,j'} \mathcal{A}^{\alpha_0}_{rs}(j,q,\alpha;j',q',\alpha')\mu^{\alpha}_{pj}(\tau)\mu^{\alpha'}_{p'j'}(\tau)$$

$$= \sum_k \mathcal{A}^{\alpha_0}_{ks}(p,q,\alpha;p',q',\alpha')\mu^{\alpha}_{kr}(\tau). \qquad \text{(IX.3)}$$

Since the μ's are unitary, (IX.2) and (IX.3) both describe the same system of homogeneous equations. Moreover, (VIII.7) is the statement that the Clebsch—Gordan coefficients satisfy this system.

Formulas (IX.2) and (IX.3) may be derived for the A's and the u's of the seminormal representation by applying (III.16) and (IV.14).

We also note that any solution of the equations (IX.2) or (IX.3) for the neighboring transpositions satisfies the system obtained when τ ranges over S_n.

X. EXAMPLES

To illustrate our procedures we consider a few examples, doing the calculation by means of the iterative formula (VII.2) for the orthogonal situation. For S_1 there is only one $\mathcal{A}$; it has the value 1.

For S_2 there are two representations, both one-dimensional, numbered as listed:

$$\alpha = \boxed{1}\,;\ r=1: \begin{array}{|c|}\hline 1\\\hline 2\\\hline\end{array}\,, \ \text{and}\ \ \alpha=\boxed{2}\,;\ r=2: \begin{array}{|c|c|}\hline 1&2\\\hline\end{array}\,,$$

$$\mu^{\boxed{1}}(12) = -1, \quad \mu^{\boxed{2}}(12) = 1. \qquad \text{(X.1)}$$

Both representations star down to $\boxed{1}$ of S_1. For both representations, we also have

$$\theta = 2!/f = 2/1 = 2. \qquad \text{(X.2)}$$

Since we have, for S_1, $\theta = 1$, we get

$$\mathcal{A}^{\boxed{1}}_{11}(1,1,\boxed{1}\,;1,1,\boxed{1}\,)$$

$$= \tfrac{1}{2}(1) + \tfrac{1}{2}(1)(1)(-1)(-1)(-1) = 0, \qquad \text{(X.3)}$$

$$\mathcal{A}^{\boxed{2}}_{11}(1;1,\boxed{1}\,;1,1,\boxed{1}\,)$$

$$= \tfrac{1}{2}(1) + \tfrac{1}{2}(1)(1)(1)(-1)(-1) = 1. \qquad \text{(X.4)}$$

It is now obvious that if, in the triplet $(\alpha_0, \alpha, \alpha')$, there is an odd number of $\boxed{1}$'s, the corresponding $\mathcal{A}$ is 0; if there is an even number, it is one. [See also (VIII.35).]

These results can also be obtained from the definition (IV.11),

$$\mathcal{A}^{\boxed{1}}_{11}(1,1,\boxed{1}\,;1,1,\boxed{1}\,)$$

$$= \frac{1}{\theta^{\boxed{1}}}[\mu^{\boxed{1}}(12)\mu^{\boxed{1}}(12)\mu^{\boxed{1}}(12) + \mu^{\boxed{1}}(\epsilon)\mu^{\boxed{1}}(\epsilon)\mu^{\boxed{1}}(\epsilon)]$$

$$= \tfrac{1}{2}[(-1)^3 + (1)^3] = 0,$$

$$\mathcal{A}^{\boxed{2}}_{11}(1,1,\boxed{1}\,;1,1,\boxed{1}\,) = \tfrac{1}{2}[(1)(-1)^2 + (1)^3] = 1.$$

To avoid ambiguity, we will use "a"'s to denote the $\mathcal{A}$'s for starred representations and tableaux. For S_3 there are three representations; two are one-dimensional, one is two-dimensional:

$$\alpha = \boxed{1}\,;\ r=1: \begin{array}{|c|}\hline 1\\\hline 2\\\hline 3\\\hline\end{array}\,,\ \ \begin{array}{l} u^{\boxed{1}}(12) = u^{\boxed{1}}(23) = -1,\\ \mu^{\boxed{1}}(12) = \mu^{\boxed{1}}(23) = -1, \quad \theta = 6,\end{array}$$

$$\alpha = \boxed{2}\,;\ r=1: \begin{array}{|c|c|}\hline 1&2\\\hline 3\\\cline{1-1}\end{array}\,, \quad r=2: \begin{array}{|c|c|}\hline 1&3\\\hline 2\\\cline{1-1}\end{array}\,, \quad \theta = 3,$$

$$u^{\boxed{2}}(12) = \mu^{\boxed{2}}(12) = \begin{pmatrix} 1 & 0\\ 0 & -1\end{pmatrix},$$

$$u^{\boxed{2}}(23) = \begin{pmatrix} -1/2 & 3/4\\ 1 & 1/2\end{pmatrix},$$

$$\mu^{\boxed{2}}(33) = \begin{pmatrix} -1/2 & \sqrt{3/4}\\ \sqrt{3/4} & 1/2\end{pmatrix},$$

$$\alpha = \boxed{3}\,;\ r=1: \begin{array}{|c|c|c|}\hline 1&2&3\\\hline\end{array}\,, \quad \theta = 6,$$

$$u^{\boxed{3}}(12) = u^{\boxed{3}}(23) = 1,$$

$$\mu^{\boxed{3}}(12) = \mu^{\boxed{3}}(23) = 1.$$

Then $\alpha = \boxed{1}$ gives $\alpha^* = \boxed{1}$, $\alpha = \boxed{2}$ gives $\alpha^*(1) = \boxed{2}$, $\alpha^*(2) = \boxed{1}$, and $\alpha = \boxed{3}$ gives $\alpha^* = \boxed{2}$.

To compute $\mathcal{A}^{\boxed{1}}_{11}(1,1,\boxed{2}\,;1,1,\boxed{2}\,)$, we note that all of the a's are of the form

$$a^{\boxed{1}}_{11}(1,1,\boxed{2}\,;1,1,\boxed{2}\,) = 0,$$

so that

$$\mathcal{A}^{\boxed{1}}_{11}(1,1,\boxed{2}\,;1,1,\boxed{2}\,) = 0.$$

Another example is

$$\mathcal{A}^{\boxed{1}}_{11}(1,2,\boxed{2}\,;2,1,\boxed{2}\,)$$

$$= 0 + \tfrac{4}{6}a^{\boxed{1}}_{11}(1,1,\boxed{2}\,;1,1,\boxed{1}\,)$$

$$\times a^{\boxed{1}}_{11}(1,1,\boxed{1}\,;1,1,\boxed{2}\,)\mu^{\boxed{1}}\mu^{\boxed{2}}_{12}\mu^{\boxed{2}}_{21}$$

$$= 0 + \tfrac{4}{6}(1)(1)(-1)\sqrt{3/4}\,\sqrt{3/4} = -\tfrac{1}{2}$$

[where $\tilde{\mathcal{A}} = 0$ since $j_n(p) \neq j_n(q)$ (or $j_n(p') \neq j_n(q')$)].

Further,

$$\mathcal{A}^{\boxed{2}}_{11}(1,1,\underline{\overline{2}};1,1,\boxed{2})$$

$$=\tfrac{2}{3}a^{\boxed{2}}_{11}(1,1,\boxed{2};1,1,\boxed{2})$$

$$+\tfrac{4}{3}(a^{\boxed{2}}_{11}(1,1,\boxed{2};1,1,\boxed{2}))^2(\mu^{\boxed{2}}_{11})^3$$

$$=\tfrac{2}{3}-\tfrac{4}{3}(1)^2(-\tfrac{1}{2})^3=\tfrac{1}{2}.$$

In the same way we obtain

$$\mathcal{A}^{\boxed{2}}_{11}(1,2,\boxed{2},;2,2,\boxed{2})=0,$$

$$\mathcal{A}^{\boxed{1}}_{11}(1,1,\boxed{2};2,2,\boxed{2})=\tfrac{1}{2}.$$

Shifting our attention, now, to the Clebsch–Gordan coefficients, we note that for S_3, all the multiplicities are one. Using (VI. 8) and (VI. 9) which includes a sign convention, we have

$$C_1^{\boxed{1},\boxed{2},\boxed{2}}(1,2)=\mathcal{A}^{\boxed{1}}_{11}(1,1,\boxed{2};2,2,\boxed{2})$$

$$=\sqrt{1/2}=\sqrt{2}/2,$$

and

$$C_1^{\boxed{1},\boxed{2},\boxed{2}}(2,1)=\frac{\mathcal{A}^{\boxed{1}}_{11}(2,1,\boxed{2};1,2,\boxed{2})}{C_1^{\boxed{1},\boxed{2},\boxed{2}}(1,2)}$$

$$=\frac{-1/2}{\sqrt{2/2}}=-\sqrt{2/}2\,.$$

Our next examples are for the tableau function ψ. We give a table of values for some tableaux:

$$S_1: r=\boxed{1},\quad \psi_1^{\boxed{1}}=1,$$

$$S_2: r=\begin{array}{|c|}\hline 1\\\hline 2\\\hline\end{array},\quad \psi_1^{\boxed{1}}=(1+\tfrac{1}{1})=2,$$

$$r=\begin{array}{|c|c|}\hline 1&2\\\hline\end{array},\quad \psi_1^{\boxed{2}}=1,$$

$$S_3: r=\begin{array}{|c|}\hline 1\\\hline 2\\\hline 3\\\hline\end{array},\quad \psi_1^{\boxed{1}}=(1+\tfrac{1}{2})(1+\tfrac{1}{1})(2)=6,$$

$$r=\begin{array}{|c|c|}\hline 1&2\\\hline 3&\\\cline{1-1}\end{array},\quad \psi_1^{\boxed{2}}=(1+1/2)(1)(1)=3/2,$$

$$r=\begin{array}{|c|c|}\hline 1&3\\\hline 2&\\\cline{1-1}\end{array},\quad \psi_2^{\boxed{2}}=(1)(2)=2,$$

$$r=\begin{array}{|c|c|c|}\hline 1&2&3\\\hline\end{array},\quad \psi_1^{\boxed{3}}=1.$$

Thus, by (IV. 14), for the seminormal coefficients for S_2, we get $A=\mathcal{A}$ in every case. For S_3 we obtain, for example,

$$A^{\boxed{1}}_{11}(1,2,\boxed{2};2,1,\boxed{2})=\mathcal{A}^{\boxed{1}}_{11}(1,2,\boxed{2};2,1,\boxed{2})$$

and

$$A^{\boxed{1}}_{11}(1,2,\boxed{2};2,2,\boxed{2})=\left(\frac{2}{3/2}\right)^{1/2}\mathcal{A}^{\boxed{1}}_{11}(1,2,\boxed{2};2,\ 2,\boxed{2}).$$

Finally, we list the basis vectors, in the space $V[x]$ [see (IV. 1)], for the two representations corresponding to the frame $\alpha=2$: $\begin{array}{|c|c|}\hline \ &\ \\\hline \ &\\\cline{1-1}\end{array}$, of S_3. The examples are for the orthogonal case. The tableaux are $r=1$: $\begin{array}{|c|c|}\hline 1&2\\\hline 3&\\\cline{1-1}\end{array}$ and $r=2$: $\begin{array}{|c|c|}\hline 1&3\\\hline 2&\\\cline{1-1}\end{array}$. We have, for the first representation,

$$\xi^{\boxed{2}}_{11}(x)=3[x_1x_2x_3+x_2x_1x_3+\tfrac{1}{2}(-x_1x_3x_2$$

$$-x_3x_2x_1-x_2x_3x_1-x_3x_1x_2)],$$

$$\xi^{\boxed{2}}_{21}(x)=3\sqrt{3/2}[x_1x_3x_2-x_3x_2x_1-x_2x_3x_1+x_3x_1x_2],$$

and, for the second,

$$\xi^{\boxed{2}}_{12}(x)=3\sqrt{3/2}[x_1x_3x_2-x_3x_2x_1+x_2x_3x_1-x_3x_1x_2],$$

$$\xi^{\boxed{2}}_{22}(x)=3[x_1x_2x_3-x_2x_1x_3+\tfrac{1}{2}(x_1x_3x_2$$

$$+x_3x_2x_1-x_2x_3x_1-x_3x_1x_2)].$$

XI. CONCLUSION

A formalism has been developed for describing the decomposition of the tensor product of two irreducible representations of S_n into a direct sum of irreducible representations. This has been discussed both for the tensor product of the vectors and for the tensor product of the matrices of the representations. We have obtained the tensor coupling coefficients and the Clebsch–Gordan coefficients. We have considered relations between the two sets and developed methods of calculating the latter from the former.

Although done for the symmetric group, the discussion is essentially unchanged for any finite group.

An iterative procedure for calculating the tensor coupling coefficients has been derived and discussed. We have considered, in detail, a systematic procedure for obtaining the Clebsch–Gordan coefficients from the tensor coupling coefficients, emphasizing the general case when the multiplicity may be greater than one.

Thus the paper forms a foundation for an actual calculation of these coefficients, which we consider in a future work.[1]

ACKNOWLEDGMENTS

We would like to thank the Baruch College Small Grants Program and the CUNY Research Foundation for their assistance.

[1] S. Schindler and R. Mirman, J. Math. Phys. 18, xxxx (1977).
[2] Daniel Edwin Rutherford, *Substitutional Analysis* (Cambridge U.P., Cambridge, 1948).
[3] Hermann Boerner, *Representations of Groups*, translated by P.G. Murphy with J. Mayer-Kalkschmidt and P. Carr (North-Holland, Amsterdam, 1970).
[4] Morton Hamermesh, *Group Theory and its Application to Physical Problems* (Addison-Wesley, Reading, Massachusetts, 1962).
[5] H. Weyl, *Theory of Groups and Quantum Mechanics* (Princeton U.P., Princeton, N.J., 1931).
[6] H. Weyl, *The Classical Groups* (Princeton U.P., Princeton, N.J., 1946).
[7] G. De B. Robinson, *Representation Theory of the Symmetric Group* (University of Toronto Press, Toronto, 1961).
[8] D.E. Littlewood, *The Theory of Group Characters* (Oxford U.P., New York, 1950).
[9] P. Rudra and M.K. Sikdar, J. Math. Phys. 17, 463 (1976); R. Berenson and J.L. Birman, J. Math. Phys. 16, 227 (1975); R. Berenson, I. Itzkan, and J.L. Birman, J. Math. Phys. 16, 236 (1975); and references cited in these papers.
[10] Boerner, Ref. 3, Chap. III.
[11] Ref. 10, Sec. 3.
[12] Ref. 10, Sec. 4.
[13] Boerner, Ref. 3, Chap. IV, Sec. 5; Rutherford, Ref. 2, Sec. 17.
[14] Rutherford, Ref. 2, Sec. 27.
[15] Boerner, Ref. 3, Chap. IV, Sec. 6.
[16] Hamermesh, Ref. 4, p. 150.
[17] Rutherford, Ref. 2, Sec. 20.
[18] Hamermesh, Ref. 4, Secs. 7–14.

The Clebsch–Gordan coefficients of S_n*

Susan Schindler
Department of Mathematics, Baruch College, City University of New York, New York, New York 10010

R. Mirman
155 East 34th Street, New York, New York 10016

(Received 12 July 1976)

Ordering schemes for the frames and tableaux of S_n are presented, and some results, expressible in terms of these, are developed. A formula is derived for the sign function on tableaux. A table of the nonzero Clebsch–Gordon coefficients for the "working triplets" is given. Methods are described, and a needed table supplied, for finding the other coefficients from the tabulated ones. The values are for $n = 2, \ldots, 6$, with the coefficients for $n = 6$ relegated to PAPS.

I. INTRODUCTION

In a previous paper[1] we have investigated the decomposition of the tensor product of irreducible representations of the symmetric group into a direct sum of irreducible representations. The decomposition may be specified by the matrix of the similarity transformation between the representation matrices for the product and the direct sum. The Clebsch—Gordan coefficients are the entries of the matrix of the similarity transformation.

Our earlier paper contains much of the theoretical groundwork for the construction of the coefficients. In order to calculate them, we must first consider a number of technical problems. This is done in the present paper.

The final results, of course, are the Clebsch—Gordan coefficients. We present them in Table I. However, for reasons of space, and to reduce the amount of computation, coefficients for all the triplets of representations are not listed, but only those for the "working triplet." We give simple instructions for finding the others.

Each representation, and each tableau within a representation, is assigned an ordinal. In Sec. II, the ordering schemes are discussed. Algorithms for generating the numbers describing the tableaux for the symmetric group S_n are given. In Sec. III, we treat the operation of "starring" tableaux for S_n to produce the S_{n-1} tableaux which are needed in the iterative formula, and we relate this operation to the ordering schemes. We also indicate a way of specifying pairs of tableaux that are needed in the iterative formula. Conjugation and its relationship to the ordering schemes are discussed in Sec. IV. In addition, a formula is derived for the sign function (on tableaux); this is needed for conjugation. In Sec. V, we describe how the working triplet may be found, and we give specific instructions for finding the coefficients of any triplet from those of the working triplet.

We conclude with the tables. Table I lists the Clebsch—Gordan coefficients. They are given as decimals. In theory, at least, these coefficients are algebraic functions of integers, that is, they may be obtained from integers by taking roots, sums, differences, products and quotients. However, such expressions become increasingly complicated as we proceed from S_{n-1} to S_n by iteration. Thus, it is more practical to work with, and express the coefficients, as decimals.

Finally, Table II gives the dimension of each representation, the ordinal of its conjugate, and the sign and psi functions for the tableaux.

The tables are for $n = 2, \ldots, 6$, with the coefficients for $n = 6$ relegated to PAPS.[2]

II. ORDERING AND SPECIFICATIONS

Frames are described by their row lengths and tableaux by the sets of numbers giving the rows in which each of the numerals $1, \ldots, n$ appear. We give, in this section, ordering schemes for the frames and tableaux (last letter ordering) and describe algorithms for generating the row lengths and the sets of numbers specifying the tableaux, in order.

We recall that a frame for S_n is an array of n boxes, in rows, each beginning in the same initial column, and whose lengths form a nonincreasing sequence. If α denotes a frame, we let $m_i(\alpha)$ equal the length of its ith row. Then,

$$\sum_{i=1}^{n} m_i(\alpha) = n, \qquad \text{(II. 1)}$$

and

$$m_1(\alpha) \geq m_2(\alpha) \geq \cdots \geq m_n(\alpha) \geq 0. \qquad \text{(II. 2)}$$

We will identify α by the sequence of its row lengths, that is,

$$\alpha: (m_1(\alpha), m_2(\alpha), \ldots, m_n(\alpha)). \qquad \text{(II. 3)}$$

The set of frames for S_n may be ordered in the following way. If α and β label frames and

$$m_k(\alpha) = m_k(\beta), \quad k > k_0, \quad m_{k_0}(\alpha) > m_{k_0}(\beta) \qquad \text{(II. 4)}$$

for some k_0 with $1 \leq k_0 \leq n$, we say $\alpha < \beta$. Then $\alpha < \beta$ and $\beta < \gamma$ imply $\alpha < \gamma$, and for any two frames, α and β, either $\alpha < \beta$ or $\beta < \alpha$ or $\alpha = \beta$. That is, "<" totally orders the frames for S_n.

To each frame corresponds a set of standard tableaux. A standard tableau, we recall, is obtained from a frame for S_n by inserting each of the numbers $1, \ldots, n$ into a box of the frame so that the numbers in each row and in each column increase. If r is a standard tableau, we identify it with a sequence, that is,

$$r: (j_1(r), j_2(r), \ldots, j_n(r)),$$

TABLE I. The Clebsch—Gordan coefficients for the symmetric group. Listed here are the nonzero Clebsch—Gordan coefficients for the working triplets with nonzero multiplicity, for the orthogonal class of representations of S_n, for $n=2$ through $n=5$. If a triplet does not appear, either it is equivalent to a working triplet, or it has zero multiplicity (or both). An unlisted coefficient for a listed triplet is zero. The rules for obtaining the coefficients for nonworking triplets are given in Sec. V. The table is read as follows: First the value of n is listed. Then, for each triplet there are four numbers, the ordinals specifying the representation, α_0, α, α', where α_0 labels a summand in the decomposition of the tensor product of α with α', and, slightly separated from these three numbers, the multiplicity. Under these four numbers are indices r, p and p', going with, respectively, α_0, α and α', followed by the multiplicity index, which labels the occurrence of α_0. The final number is the Clebsch—Gordan coefficient. For a particular triplet we read across and then down. Note that this table corrects a misprint in Ref. 3, for the multiplicity of the triplet $(3,1^3)\rightarrow(2^3)\times(3,2,1)$, which, in our notation, is the triplet (4,5,6). See Ref. 1.

```
S(2)

1  1  2    1
 1 1 1  1    1.000000

S(3)

1  1  3    1
 1 1 1  1    1.000000

1  2  2    1
 1 1 2  1    0.707107    1 2 1  1   -0.707107

2  2  2    1
 1 1 1  1    0.707107    1 2 2  1   -0.707107    2 1 2  1   -0.707107    2 2 1  1   -0.707107

S(4)

1  1  5    1
 1 1 1  1    1.000000

1  2  4    1
 1 1 3  1    0.577350    1 2 2  1   -0.577350    1 3 1  1    0.577350

1  3  3    1
 1 1 2  1    0.707107    1 2 1  1   -0.707107

2  2  2    1
 1 2 3  1    0.707107    1 3 2  1   -0.707107    2 1 3  1   -0.707107    2 3 1  1    0.707107    3 1 2  1    0.707107
 3 2 1  1   -0.707107

2  2  3    1
 1 1 1  1    0.500000    1 2 2  1   -0.500000    1 3 2  1   -0.707107    2 1 2  1   -0.500000    2 2 1  1   -0.500000
 2 3 1  1    0.707107    3 1 2  1   -0.707107    3 2 1  1    0.707107

2  2  4    1
 1 1 1  1    0.408248    1 1 2  1    0.577350    1 2 3  1   -0.577350    1 3 3  1    0.408248    2 1 3  1   -0.577350
 2 2 1  1    0.408248    2 2 2  1   -0.577350    2 3 2  1   -0.408248    3 1 3  1    0.408248    3 2 2  1   -0.408248
 3 3 1  1   -0.816497

3  3  3    1
 1 1 1  1    0.707107    1 2 2  1   -0.707107    2 1 2  1   -0.707107    2 2 1  1   -0.707107

S(5)

1  1  7    1
 1 1 1  1    1.000000

1  2  6    1
 1 1 4  1    0.500000    1 2 3  1   -0.500000    1 3 2  1    0.500000    1 4 1  1   -0.500000

1  3  5    1
 1 1 5  1    0.447214    1 2 4  1   -0.447214    1 3 3  1   -0.447214    1 4 2  1    0.447214    1 5 1  1   -0.447214

1  4  4    1
 1 1 6  1    0.408248    1 2 5  1   -0.408248    1 3 4  1    0.408248    1 4 3  1    0.408248    1 5 2  1   -0.408248
 1 6 1  1    0.408248

2  2  4    1
 1 2 6  1    0.577350    1 3 5  1   -0.577350    1 4 3  1    0.577350    2 1 6  1   -0.577350    2 3 4  1    0.577350
 2 4 2  1   -0.577350    3 1 5  1    0.577350    3 2 4  1   -0.577350    3 4 1  1    0.577350    4 1 3  1   -0.577350
 4 2 2  1    0.577350    4 3 1  1   -0.577350

2  2  5    1
 1 1 1  1    0.149071    1 1 2  1    0.210819    1 1 4  1    0.365148    1 2 3  1   -0.210819    1 2 5  1   -0.365148
 1 3 3  1    0.149071    1 3 5  1   -0.516398    1 4 3  1    0.577350    2 1 3  1   -0.210819    2 1 5  1   -0.365148
 2 2 1  1    0.149071    2 2 2  1   -0.210819    2 2 4  1   -0.365148    2 3 2  1   -0.149071    2 3 4  1    0.516398
 2 4 2  1   -0.577350    3 1 3  1    0.149071    3 1 5  1   -0.516398    3 2 2  1   -0.149071    3 2 4  1    0.516398
 3 3 1  1   -0.298142    3 4 1  1    0.577350    4 1 3  1    0.577350    4 2 2  1   -0.577350    4 3 1  1    0.577350
```

TABLE I. (*Continued*).

2 2 6 1

1	1	1	1	0.288675	1	1	2	1	0.372678	1	1	3	1	0.527046	1	2	4	1	-0.527046	1	3	4	1	0.372678
1	4	4	1	-0.288675	2	1	4	1	-0.527046	2	2	1	1	0.288675	2	2	2	1	0.372678	2	2	3	1	-0.527046
2	3	3	1	-0.372678	2	4	3	1	0.288675	3	1	4	1	0.372678	3	2	3	1	-0.372678	3	3	1	1	0.288675
3	3	2	1	-0.745356	3	4	2	1	-0.288675	4	1	4	1	-0.288675	4	2	3	1	0.288675	4	3	2	1	-0.288675
4	4	1	1	-0.866025																				

2 3 3 1

1	1	3	1	0.223607	1	2	4	1	-0.223607	1	2	5	1	-0.316228	1	3	1	1	-0.223607	1	4	2	1	0.223607
1	4	5	1	-0.547723	1	5	2	1	0.316228	1	5	4	1	0.547723	2	1	4	1	-0.223607	2	1	5	1	0.316228
2	2	3	1	-0.223607	2	3	2	1	0.223607	2	3	5	1	0.547723	2	4	1	1	0.223607	2	5	1	1	-0.316228
2	5	3	1	-0.547723	3	1	4	1	0.316228	3	2	3	1	-0.316228	3	3	2	1	0.316228	3	3	4	1	-0.547723
3	4	1	1	-0.316228	3	4	3	1	0.547723	4	1	2	1	-0.707107	4	2	1	1	0.707107					

2 3 4 1

1	1	1	1	0.129099	1	1	2	1	-0.091287	1	1	4	1	-0.353553	1	2	3	1	0.091287	1	2	5	1	0.353553
1	2	6	1	0.500000	1	3	1	1	0.223607	1	3	2	1	0.316228	1	4	3	1	-0.316228	1	4	6	1	-0.288675
1	5	3	1	0.223607	1	5	5	1	0.288675	2	1	3	1	0.091287	2	1	5	1	0.353553	2	1	6	1	-0.500000
2	2	1	1	0.129099	2	2	2	1	0.091287	2	2	4	1	0.353553	2	3	3	1	-0.316228	2	3	6	1	0.288675
2	4	1	1	0.223607	2	4	2	1	-0.316228	2	5	2	1	-0.223607	2	5	4	1	-0.288675	3	1	3	1	0.129099
3	1	5	1	-0.500000	3	2	2	1	-0.129099	3	2	4	1	0.500000	3	3	3	1	0.223607	3	3	5	1	-0.288675
3	4	2	1	-0.223607	3	4	4	1	0.288675	3	5	1	1	-0.447214	4	3	3	1	-0.577350	4	4	2	1	0.577350
4	5	1	1	-0.577350																				

2 3 5 1

1	1	1	1	0.408248	1	1	2	1	-0.288675	1	2	3	1	0.288675	1	3	1	1	-0.235702	1	3	2	1	-0.333333
1	3	4	1	0.288675	1	4	3	1	0.333333	1	4	5	1	-0.288675	1	5	3	1	-0.235702	1	5	5	1	-0.408248
2	1	3	1	0.288675	2	2	1	1	0.408248	2	2	2	1	0.288675	2	3	3	1	0.333333	2	3	5	1	-0.288675
2	4	1	1	-0.235702	2	4	2	1	0.333333	2	4	4	1	-0.288675	2	5	2	1	0.235702	2	5	4	1	0.408248
3	1	3	1	0.408248	3	2	2	1	-0.408248	3	3	3	1	-0.235702	3	3	5	1	-0.408248	3	4	2	1	0.235702
3	4	4	1	0.408248	3	5	1	1	0.471405	4	1	5	1	-0.547723	4	2	4	1	0.547723	4	3	3	1	-0.365148
4	4	2	1	0.365148	4	5	1	1	-0.365148															

2 4 4 1

1	1	2	1	0.204124	1	1	4	1	0.263523	1	2	1	1	-0.204124	1	2	4	1	0.372678	1	3	5	1	-0.372678
1	3	6	1	0.263523	1	4	1	1	-0.263523	1	4	2	1	-0.372678	1	5	3	1	0.372678	1	5	6	1	-0.204124
1	6	3	1	-0.263523	1	6	5	1	0.204124	2	1	3	1	0.204124	2	1	5	1	0.263523	2	2	5	1	-0.372678
2	2	6	1	-0.263523	2	3	1	1	-0.204124	2	3	4	1	-0.372678	2	4	3	1	0.372678	2	4	6	1	0.204124
2	5	1	1	-0.263523	2	5	2	1	0.372678	2	6	2	1	0.263523	2	6	4	1	-0.204124	3	1	6	1	-0.527046
3	2	3	1	0.204124	3	2	5	1	-0.263523	3	3	2	1	-0.204124	3	3	4	1	0.263523	3	4	3	1	-0.263523
3	4	5	1	-0.204124	3	5	2	1	0.263523	3	5	4	1	0.204124	3	6	1	1	0.527046	4	1	6	1	0.408248
4	2	5	1	-0.408248	4	3	4	1	0.408248	4	4	3	1	-0.408248	4	5	2	1	0.408248	4	6	1	1	-0.408248

3 3 3 1

1	1	1	1	0.353553	1	2	2	1	-0.353553	1	3	3	1	-0.353553	1	4	4	1	0.353553	1	4	5	1	-0.500000
1	5	4	1	-0.500000	2	1	2	1	-0.353553	2	2	1	1	-0.353553	2	3	4	1	0.353553	2	3	5	1	0.500000
2	4	3	1	0.353553	2	5	3	1	0.500000	3	1	3	1	-0.353553	3	2	4	1	0.353553	3	2	5	1	0.500000
3	3	1	1	-0.353553	3	4	2	1	0.353553	3	5	2	1	0.500000	4	1	4	1	0.353553	4	1	5	1	-0.500000
4	2	3	1	0.353553	4	3	2	1	0.353553	4	4	1	1	0.353553	4	5	1	1	-0.500000	5	1	4	1	-0.500000
5	2	3	1	0.500000	5	3	2	1	0.500000	5	4	1	1	-0.500000										

3 3 4 1

1	3	1	1	0.456435	1	3	2	1	-0.322749	1	3	4	1	0.250000	1	4	3	1	0.322749	1	4	5	1	-0.250000
1	4	6	1	0.353553	1	5	3	1	0.456435	1	5	5	1	0.353553	2	3	3	1	0.322749	2	3	5	1	-0.250000
2	3	6	1	-0.353553	2	4	1	1	0.456435	2	4	2	1	0.322749	2	4	4	1	-0.250000	2	5	2	1	-0.456435
2	5	4	1	-0.353553	3	1	1	1	-0.456435	3	1	2	1	0.322749	3	1	4	1	-0.250000	3	2	3	1	-0.322749
3	2	5	1	0.250000	3	2	6	1	0.353553	3	4	6	1	0.408248	3	5	5	1	-0.408248	4	1	3	1	-0.322749
4	1	5	1	0.250000	4	1	6	1	-0.353553	4	2	1	1	-0.456435	4	2	2	1	-0.322749	4	2	4	1	0.250000
4	3	6	1	-0.408248	4	5	4	1	0.408248	5	1	3	1	-0.456435	5	1	5	1	-0.353553	5	2	2	1	0.456435
5	2	4	1	0.353553	5	3	5	1	0.408248	5	4	4	1	-0.408248										

3 3 5 1

1	1	4	1	0.612372	1	2	5	1	-0.612372	1	3	1	1	0.288675	1	3	2	1	-0.204124	1	4	3	1	0.204124
1	5	3	1	0.288675	2	1	5	1	-0.612372	2	2	4	1	-0.612372	2	3	3	1	0.204124	2	4	1	1	0.288675
2	4	2	1	0.204124	2	5	2	1	-0.288675	3	1	1	1	0.288675	3	1	2	1	-0.204124	3	2	3	1	0.204124
3	3	1	1	0.333333	3	3	2	1	0.471405	3	3	4	1	0.204124	3	4	3	1	-0.471405	3	4	5	1	-0.204124
3	5	3	1	0.333333	3	5	5	1	-0.288675	4	1	3	1	0.204124	4	2	1	1	0.288675	4	2	2	1	0.204124
4	3	3	1	-0.471405	4	3	5	1	-0.204124	4	4	1	1	0.333333	4	4	2	1	-0.471405	4	4	4	1	-0.204124
4	5	2	1	-0.333333	4	5	4	1	0.288675	5	1	3	1	0.288675	5	2	2	1	-0.288675	5	3	3	1	0.333333
5	3	5	1	-0.288675	5	4	2	1	-0.333333	5	4	4	1	0.288675	5	5	1	1	-0.666667					

3 4 4 2

1	1	2	1	0.322749	1	1	4	1	0.250000	1	1	4	2	0.408248	1	2	1	1	0.322749	1	2	2	1	0.228218
1	2	4	1	-0.176777	1	2	4	2	-0.288675	1	3	3	1	-0.228218	1	3	5	1	0.176777	1	3	5	2	0.288675
1	3	6	1	0.250000	1	3	6	2	0.408248	1	4	1	1	0.250000	1	4	1	2	-0.408248	1	4	2	1	-0.176777
1	4	2	2	0.288675	1	4	4	1	-0.228218	1	5	3	1	0.176777	1	5	3	2	-0.288675	1	5	5	1	0.228218
1	5	6	1	-0.322749	1	6	3	1	0.250000	1	6	3	2	-0.408248	1	6	5	1	-0.322749	2	1	3	1	0.322749
2	1	5	1	0.250000	2	1	5	2	0.408248	2	2	3	1	-0.228218	2	2	5	1	0.176777	2	2	5	2	0.288675
2	2	6	1	-0.250000	2	2	6	2	-0.408248	2	3	1	1	0.322749	2	3	2	1	-0.228218	2	3	4	1	0.176777
2	3	4	2	0.288675	2	4	3	1	0.176777	2	4	3	2	-0.288675	2	4	5	1	0.228218	2	4	6	1	0.322749
2	5	1	1	0.250000	2	5	1	2	-0.408248	2	5	2	1	0.176777	2	5	2	2	-0.288675	2	5	4	1	0.228218
2	6	2	1	-0.250000	2	6	2	2	0.408248	2	6	4	1	0.322749	3	1	2	2	-0.456435	3	1	4	1	-0.288675
3	1	4	2	0.117851	3	2	1	2	0.456435	3	2	4	1	-0.408248	3	2	4	2	0.166667	3	3	5	1	0.408248
3	3	5	2	-0.166667	3	3	6	1	-0.288675	3	3	6	2	0.117851	3	4	1	1	-0.288675	3	4	1	2	-0.117851
3	4	2	1	-0.408248	3	4	2	2	-0.166667	3	5	3	1	0.408248	3	5	3	2	0.166667	3	5	6	2	0.456435
3	6	3	1	-0.288675	3	6	3	2	-0.117851	3	6	5	2	-0.456435	4	1	3	2	-0.456435	4	1	5	1	-0.288675
4	1	5	2	0.117851	4	2	5	1	0.408248	4	2	5	2	-0.166667	4	2	6	1	0.288675	4	2	6	2	-0.117851
4	3	1	2	0.456435	4	3	4	1	0.408248	4	3	4	2	-0.166667	4	4	3	1	0.408248	4	4	3	2	0.166667
4	4	6	2	-0.456435	4	5	1	1	-0.288675	4	5	1	2	-0.117851	4	5	2	1	0.408248	4	5	2	2	0.166667
4	6	2	1	0.288675	4	6	2	2	0.117851	4	6	4	2	0.456435	5	1	6	1	0.577350	5	1	6	2	-0.235702
5	2	3	2	-0.456435	5	2	5	1	0.288675	5	2	5	2	-0.117851	5	3	2	2	0.456435	5	3	4	1	-0.288675
5	3	4	2	0.117851	5	4	3	1	-0.288675	5	4	3	2	-0.117851	5	4	5	2	0.456435	5	5	2	1	0.288675
5	5	2	2	0.117851	5	5	4	2	-0.456435	5	6	1	1	0.577350	5	6	1	2	0.235702					

4 4 4 1

1	2	4	1	0.500000	1	3	5	1	0.500000	1	4	2	1	-0.500000	1	5	3	1	-0.500000	2	1	4	1	-0.500000
2	3	6	1	0.500000	2	4	1	1	0.500000	2	6	3	1	-0.500000	3	1	5	1	-0.500000	3	2	6	1	-0.500000
3	5	1	1	0.500000	3	6	2	1	0.500000	4	1	2	1	0.500000	4	2	1	1	-0.500000	4	5	6	1	0.500000
4	6	5	1	-0.500000	5	1	3	1	0.500000	5	3	1	1	-0.500000	5	4	6	1	-0.500000	5	6	4	1	0.500000
6	2	3	1	0.500000	6	3	2	1	-0.500000	6	4	5	1	0.500000	6	5	4	1	-0.500000					

TABLE II. Functions on the representations. For $n=2$ through $n=6$ this table lists the dimensions, ordinals and row lengths of the representations. In addition, it gives the psi and sign functions and the ordinal and j-values of the tableaux. For each value of n, two numbers are listed on the first line: the number n and the number of frames (representations) for that n. Below we list, for each representation, its ordinal, its dimension, the ordinal of its conjugate, and, slightly separated from these, its row lengths. Then, for each tableau, there is a line containing its ordinal, the j's, and the psi function multiplied by the sign function. (Since the ψ's are intrinsically positive, this product gives the sign function.)

```
2    2

  1    1    2         1    1
1      1    2                        2.000000000
  2    1    1         2    0
1      1    1                        1.000000000


3    3

  1    1    3         1    1    1
1      1    2    3                   6.000000000
  2    2    2         2    1    0
1      1    1    2                   1.500000000
2      1    2    1                  -2.000000000
  3    1    1         3    0    0
1      1    1    1                   1.000000000


4    5

  1    1    5         1    1    1    1
1      1    2    3    4             24.000000000
  2    3    4         2    1    1    0
1      1    1    2    3              4.000000000
2      1    2    1    3             -5.333333333
3      1    2    3    1              6.000000000
  3    2    3         2    2    0    0
1      1    1    2    2              3.000000000
2      1    2    1    2             -4.000000000
  4    3    2         3    1    0    0
1      1    1    1    2              1.333333333
2      1    1    2    1             -1.500000000
3      1    2    1    1              2.000000000
  5    1    1         4    0    0    0
1      1    1    1    1              1.000000000


5    7

  1    1    7         1    1    1    1    1
1      1    2    3    4    5       120.000000000
  2    4    6         2    1    1    1    0
1      1    1    2    3    4        15.000000000
2      1    2    1    3    4       -20.000000000
3      1    2    3    1    4        22.500000000
4      1    2    3    4    1       -24.000000000
  3    5    5         2    2    1    0    0
1      1    1    2    2    3         6.000000000
2      1    2    1    2    3        -8.000000000
3      1    1    2    3    2        -8.000000000
4      1    2    1    3    2        10.666666667
5      1    2    3    1    2       -12.000000000
  4    6    4         3    1    1    0    0
1      1    1    1    2    3         3.333333333
2      1    1    2    1    3        -3.750000000
3      1    2    1    1    3         5.000000000
4      1    1    2    3    1         4.000000000
5      1    2    1    3    1        -5.333333333
6      1    2    3    1    1         6.000000000
  5    5    3         3    2    0    0    0
1      1    1    1    2    2         2.000000000
2      1    1    2    1    2        -2.250000000
3      1    2    1    1    2         3.000000000
4      1    1    2    2    1         3.000000000
5      1    2    1    2    1        -4.000000000
  6    4    2         4    1    0    0    0
1      1    1    1    1    2         1.250000000
2      1    1    1    2    1        -1.333333333
3      1    1    2    1    1         1.500000000
4      1    2    1    1    1        -2.000000000
  7    1    1         5    0    0    0    0
1      1    1    1    1    1         1.000000000


6   11

  1    1   11         1    1    1    1    1    1
1      1    2    3    4    5    6  720.000000000
  2    5   10         2    1    1    1    1    0
1      1    1    2    3    4    5   72.000000000
2      1    2    1    3    4    5  -96.000000000
3      1    2    3    1    4    5  108.000000000
4      1    2    3    4    1    5 -115.200000000
5      1    2    3    4    5    1  120.000000000
  3    9    9         2    2    1    1    0    0
1      1    1    2    2    3    4   20.000000000
2      1    2    1    2    3    4  -26.666666667
3      1    1    2    3    2    4  -26.666666667
4      1    2    1    3    2    4   35.555555556
5      1    2    3    1    2    4  -40.000000000
6      1    1    2    3    4    2   30.000000000
7      1    2    1    3    4    2  -40.000000000
8      1    2    3    1    4    2   45.000000000
9      1    2    3    4    1    2  -48.000000000
  4   10    7         3    1    1    1    0    0
1      1    1    1    2    3    4   12.000000000
2      1    1    2    1    3    4  -13.500000000
3      1    2    1    1    3    4   18.000000000
4      1    1    2    3    1    4   14.400000000
5      1    2    1    3    1    4  -19.200000000
6      1    2    3    1    1    4   21.600000000
7      1    1    2    3    4    1  -15.000000000
8      1    2    1    3    4    1   20.000000000
9      1    2    3    1    4    1  -22.500000000
10     1    2    3    4    1    1   24.000000000
  5    5    8         2    2    2    0    0    0
1      1    1    2    2    3    3   18.000000000
2      1    2    1    2    3    3  -24.000000000
3      1    1    2    3    2    3  -24.000000000
4      1    2    1    3    2    3   32.000000000
5      1    2    3    1    2    3  -36.000000000
  6   16    6         3    2    1    0    0    0
1      1    1    1    2    2    3    3.750000000
2      1    1    2    1    2    3   -4.218750000
3      1    2    1    1    2    3    5.625000000
4      1    1    2    2    1    3    5.625000000
5      1    2    1    2    1    3   -7.500000000
6      1    1    1    2    3    2   -5.000000000
7      1    1    2    1    3    2    5.625000000
8      1    2    1    1    3    2   -7.500000000
9      1    1    2    3    1    2   -6.000000000
10     1    2    1    3    1    2    8.000000000
11     1    2    3    1    1    2   -9.000000000
12     1    1    2    2    3    1   -6.000000000
13     1    2    1    2    3    1    8.000000000
14     1    1    2    3    2    1    8.000000000
15     1    2    1    3    2    1  -10.666666667
16     1    2    3    1    2    1   12.000000000
  7   10    4         4    1    1    0    0    0
1      1    1    1    1    2    3    3.000000000
2      1    1    1    2    1    3   -3.200000000
3      1    1    2    1    1    3    3.600000000
4      1    2    1    1    1    3   -4.800000000
5      1    1    1    2    3    1    3.333333333
6      1    1    2    1    3    1   -3.750000000
7      1    2    1    1    3    1    5.000000000
8      1    1    2    3    1    1    4.000000000
9      1    2    1    3    1    1   -5.333333333
10     1    2    3    1    1    1    6.000000000
  8    5    5         3    3    0    0    0    0
1      1    1    1    2    2    2    4.000000000
2      1    1    2    1    2    2   -4.500000000
3      1    2    1    1    2    2    6.000000000
4      1    1    2    2    1    2    6.000000000
5      1    2    1    2    1    2   -8.000000000
  9    9    3         4    2    0    0    0    0
1      1    1    1    1    2    2    1.666666667
2      1    1    1    2    1    2   -1.777777778
3      1    1    2    1    1    2    2.000000000
4      1    2    1    1    1    2   -2.666666667
5      1    1    1    2    2    1    2.000000000
6      1    1    2    1    2    1   -2.250000000
7      1    2    1    1    2    1    3.000000000
8      1    1    2    2    1    1    3.000000000
9      1    2    1    2    1    1   -4.000000000
 10    5    2         5    1    0    0    0    0
1      1    1    1    1    1    2    1.200000000
2      1    1    1    1    2    1   -1.250000000
3      1    1    1    2    1    1    1.333333333
4      1    1    2    1    1    1   -1.500000000
5      1    2    1    1    1    1    2.000000000
 11    1    1         6    0    0    0    0    0
1      1    1    1    1    1    1    1.000000000
```

where $j_k(r)$ gives the row of r containing k ($k=1, 2, \ldots, n$).

We may totally order the tableaux corresponding to a particular frame. We define "<" for tableaux by saying $r<s$ if

$$j_k(r)=j_k(s), \quad \text{for } k>k_0,$$
$$j_{k_0}(r)>j_{k_0}(s) \tag{II.5}$$

for some k_0 with $1\leq k_0\leq n$.

Visually, this scheme means that if n lies in a lower row of r than of s, then $r < s$. If n lies in the same row of r and s then the relative order of r and s is determined by the relative position of $n-1$, and so on. Our scheme is that of last letter ordering.

For example, the frame (2, 1, 1) of S_4 gives three tableaux, listed in order, identified with ordinals:

Ordinal	Tableau	Tableau labels (j_1, j_2, j_3)
1	$\begin{array}{\|c\|c\|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1} 4 \\ \cline{1-1}\end{array}$	(1, 1, 2, 3)
2	$\begin{array}{\|c\|c\|}\hline 1 & 3 \\ \hline 2 \\ \cline{1-1} 4 \\ \cline{1-1}\end{array}$	(1, 2, 1, 3)
3	$\begin{array}{\|c\|c\|}\hline 1 & 4 \\ \hline 2 \\ \cline{1-1} 3 \\ \cline{1-1}\end{array}$	(1, 2, 3, 1)

Note that the position of the numeral 1 in a tableau is listed first and of n, last. This scheme is in contrast to others which do the listing in the reverse order.

The first tableau of any frame is ordered lexicographically. That is, if we read across the rows, in turn, we get the numbers $1, 2, \ldots, n$ in order.

The frames are specified by their row lengths; next we give a method of producing them in the proper order. We construct the frames in stages, according to the number of rows they are to have. Since, if α has more rows than β, $\alpha < \beta$, we start with the single frame with n rows, and at each stage we decrease the number of rows by one. Then, to generate the totality of frames in order, it suffices to generate them in order at each stage. The following iterative procedure gives the required construction at each stage.

To obtain the frames with k rows, we use the following fact: For any frame of k rows the last row has at most $[n/k]$ boxes where $[n/k]$ is the largest integer less than or equal to n/k, because all rows of a frame are at least as long as the last one. For each $i = [n/k]$, $[n/k] - 1, \ldots, 1$, we adjoin a row of length i to the bottom of each frame of S_{n-i}. These frames are taken in order, skipping only those whose last row has length smaller than i. We obtain frames for S_n, whose last row, that is, the kth row, has length i.

Clearly, if two frames with the same number of rows correspond to different values of i, the last row length, then the one with the larger value of i comes first in the ordering scheme, and is also generated first in our construction (since the i's are used in decreasing order). Two frames with the same number, k, of rows, and the same last row length, i, are ordered according to the order of the frames given by their first $k-1$ rows, which is the order for the S_{n-i} frames.

This construction produces every frame for S_n exactly once, and in the correct order.

Finally, we note that the lengths, $m_j'(\alpha)$, of the columns of a frame α are given by

$$m_j'(\alpha) = \max\{i : j \leq m_i(\alpha)\}; \tag{II.6}$$

they also form a nonincreasing sequence.

Technically, frames and tableaux are labels. For simplicity we use the terms "frame" and "equivalence class," "tableau," and "basis vector" interchangeably.

III. STARRING

Our method of determining the Clebsch–Gordan coefficients uses an iterative procedure, based on the device of "starring." If r is a standard tableau corresponding to a frame α, for S_n, then deleting from α the box in which the number n appears in r, gives a frame $\alpha^*(r)$, for S_{n-1}. We let r^* be the tableau, corresponding to $\alpha^*(r)$, obtained by deleting this box from r. Then

$$m_i(\alpha^*(r)) = \begin{cases} m_i(\alpha), & i \neq j_n(r), \\ m_{j_n}(\alpha) - 1, & i = j_n(r). \end{cases} \tag{III.1}$$

We must show that this does give a frame for S_{n-1}. If $m_n(\alpha) \neq 0$, then α: $(1, 1, \ldots, 1)$ and so $m_n(\alpha^*) = 0$. Thus $\sum_{i=1}^{n-1} m_i(\alpha^*(r)) = n-1$. If $m_n(\alpha) = 0$, it is clear that $\sum_{i=1}^{n-1} m_i(\alpha^*(r)) = n-1$. The rows of any frame must be nonincreasing in length. We thus show that $i = j_n(r) \neq n$ implies $m_i(\alpha^*(r)) \geq m_{i+1}(\alpha^*(r))$. (This inequality is obvious for other values of i.) Since the numbers in the rows and columns of r increase, the number n must appear at the end of its row and of its column, so that we must have $m_{i+1}(\alpha) < m_i(\alpha)$ and so $m_{i+1}(\alpha) \leq m_i(\alpha) - 1$ as required.

We note that

$$r^*: j_k(r^*) = j_k(r), \tag{III.2}$$

for $k = i, \ldots, n-1$.

We have another means of labeling tableaux, and so the coefficients, which is based on starring. Each S_n tableau, r, is labeled by a pair of indices: the frame (for S_{n-1}) to which r^* corresponds (given by the row of r in which n appears), and the ordinal of r^*. Running through the tableaux, in order, for a particular frame of S_n, we obtain a set of S_{n-1} representations, not in order (which does not matter) and the tableaux of these representations, in correct order, for each starred representation.

Selection of tableaux satisfying the conditions in the iterative formula is facilitated by this labeling scheme. The conditions are stated in terms of certain pairs of indices (both of which give a tableau, for the same frame.) The condition $j_n(r) = j_n(s)$ amounts to the requirement that the first indices of r and s, in this labeling scheme, be the same.

IV. CONJUGATION

To use the operation of conjugation requires some facts about the effect of conjugation on the ordering schemes. We must also describe the sign function in terms of the tableau labels. We consider these questions here.

The conjugate, $\bar{\alpha}$, of α, is defined by

$$\bar{\alpha}: m_i(\bar{\alpha}) = m_i'(\alpha). \tag{IV.1}$$

For a tableau, r, corresponding to α, we let $j_k'(r)$ be the column containing $k = 1, \ldots, n$. The conjugate, $\bar{r}$, of r is then given by

$$\bar{r}: j_k(\bar{r}) = j_k'(r). \tag{IV. 2}$$

Then $\bar{r}$ is a standard tableau corresponding to $\bar{\alpha}$.

We observe that the operations of conjugation and starring commute, in the sense that

$$j_k(\overline{r^*}) = j_k((\bar{r})^*). \tag{IV. 3}$$

The effect of conjugation on the ordering scheme for tableaux is to reverse it.

Theorem: If r and s correspond to α and $r < s$, then $\bar{s} < \bar{r}$.

Proof: We proceed by induction. For S_1 the result is trivial. We take, as our hypotheses:

$$r < s \text{ implies } \bar{s} < \bar{r}$$

whenever r and s correspond to the same frame for S_{n-1}. We must prove the hypothesis holds for S_n.

Let r and s correspond to α, with $r < s$. First suppose $j_n(r) > j_n(s)$. Now n can only occupy the last position in its row, so that $j_n'(r) = m_{j_n(r)}(\alpha)$ and $j_n'(s) = m_{j_n(s)}(\alpha)$. The row lengths of α are nonincreasing so, by (IV. 2), $j_n(\bar{r}) < j_n(\bar{s})$ and $\bar{s} < \bar{r}$.

For the general situation we have

$$j_k(r) = j_k(s), \quad j_{k_0}(r) > j_{k_0}(s), \tag{IV. 4}$$

for $k > k_0$, some $k_0 \leq n-1$. Starring r and s gives tableaux, corresponding to $\alpha^*(r) = \alpha^*(s)$ for S_{n-1}, and from (III. 2) and (IV. 4), we see that $r^* < s^*$. We apply the inductive hypothesis and obtain $\overline{s^*} < \overline{r^*}$. That is,

$$j_k(\overline{r^*}) = j_k(\overline{s^*}), \quad j_{k_1}(\overline{r^*}) < j_{k_1}(\overline{s^*})$$

for $k > k_1$, some $k_1 \leq n-1$. Now, from (IV. 3) and (IV. 2), $j_k(\overline{r^*}) = j_k((\bar{r})^*) = j_k(\bar{r})$, and $j_k(\overline{s^*}) = j_k(\bar{s})$ for $k \leq n-1$, and $j_n'(r) = m_{j_n(r)}(\alpha) = j_n'(s)$ by (IV. 4). Then (IV. 2) gives

$$j_k(\bar{r}) = j_k(\bar{s}), \quad j_{k_1}(\bar{r}) < j_{k_1}(\bar{s}),$$

for $k > k_1$, so that $\bar{s} < \bar{r}$, as required.

For frames, one can show, by inspecting all cases, that, for $n \leq 5$, $\alpha < \beta$ implies $\bar{\beta} < \bar{\alpha}$. However, no such relationship holds beyond 5. For S_6, we have, for example,

$$(2, 2, 1, 1, 0, 0) < (3, 1, 1, 1, 0, 0) < (2, 2, 2, 0, 0, 0), \tag{IV. 5}$$

but

$$\overline{(3, 1, 1, 1, 0, 0)} < \overline{(2, 2, 2, 0, 0, 0)} < \overline{(2, 2, 1, 1, 0, 0)}. \tag{IV. 6}$$

In order to use conjugation we need the sign function on tableaux. For any frame α there is one tableau, r_0, whose entries are in lexicographic order, that is,

$$r_0: j_k(r_0) \leq j_{k+1}(r_0), \tag{IV. 7}$$

$k = 1, \ldots, n-1$. Then r_0 occurs first in the ordering scheme for the tableaux corresponding to α. If r is any other tableau of this frame, there is a unique permutation $\sigma_r \in S_n$ transforming r into r_0. That is $j_{\sigma_r(k)}(r) = j_k(r_0)$ and $j_{\sigma_r(k)}'(r) = j_k'(r_0)$, $k = 1, \ldots, n$, which we abbreviate as

$$\sigma_r(r) = r_0. \tag{IV. 8}$$

We define

$$\Lambda(r) = \operatorname{sign}(\sigma_r). \tag{IV. 9}$$

Clearly, $\Lambda(r_0) = 1$. We find $\Lambda(r)$, for any r, in stages.

Lemma: If r corresponds to α and r^* is lexicographic, that is, r^* is the first tableau corresponding to $\alpha^*(r)$, then

$$\Lambda(r) = (-1)^{n-i(r)}, \tag{IV. 10}$$

where

$$i(r) = \sum_{j \leq j_n(r)} m_j(\alpha) \tag{IV. 11}$$

Proof: Reading across the rows of r, in turn, gives the sequence

$$1, \ldots, i(r)-1, n, i(r), \ldots, n-1,$$

since n lies at the end of its row, in r. Then, multiplying from right to left, the product of transpositions

$$(n, n-1) \cdot \cdots \cdot (n, i(r)+1)(n, i(r)) \tag{IV. 12}$$

transforms the sequence into $1, 2, \ldots, n$, and thus σ_r is the product (IV. 12). Then σ_r is a product of $n - i(r)$ transpositions so that $\Lambda(r) = (-1)^{n-i(r)}$.

We move on the general case.

Theorem: For any tableau, r, for S_n, and $k = 2, \ldots, n$, let

$$\nu_k(r) = \text{number of } k' < k \text{ with } j_{k'}(r) > j_k(r) \tag{IV. 13}$$

and

$$\nu(r) = \sum_{k=2}^{n} \nu_k(r). \tag{IV. 14}$$

$$\text{Then } \Lambda(r) = (-1)^{\nu(r)}. \tag{IV. 15}$$

Proof: First suppose r^* is lexicographic. Then $k \leq n-1$ implies $\nu_k(r) = 0$. Thus $\nu(r) = \nu_n(r)$. Below the row $j_n(r)$ are $n - i(r)$ boxes and so $\nu_n(r) = n - i(r)$. Using the lemma gives the desired result in this case.

The theorem is obviously true for S_1; we assume, as an inductive hypothesis,

$$\Lambda(r) = (-1)^{\nu(r)},$$

for S_{n-1}. To prove this for S_n, we first suppose that $j_n(r) \geq j_k(r)$ for each $k \leq n$. Then $\nu_n(r) = 0$ so $\nu(r) = \sum_{k=2}^{n-1} \nu_k(r) = \nu(r^*)$. Moreover, $\sigma_{r^*} \in S_{n-1} \subset S_n$ and since $j_n(r) = j_n(r_0)$, $\sigma_{r^*}(r) = r_0$. Thus, $\Lambda(r) = \operatorname{sgn}(\sigma_{r^*}) = (-1)^{\nu(r^*)} = (-1)^{\nu(r)}$; the second equality follows from the inductive hypothesis.

In the general situation, we have r corresponding to α and $j_n(r) < j_k(r)$, for some $k < n$. Letting $(r^*)_0$ be the lexicographic tableau for $\alpha^*(r)$, we have

$$\sigma_{r^*}(r^*) = (r^*)_0, \tag{IV. 16}$$

where $\sigma_{r^*} \in S_{n-1} \subset S_n$. We define r', corresponding to α, by

$$j_k(r') = j_k((r^*)_0), \quad j_n(r') = j_n(r), \tag{IV. 17}$$

for $k \leq n-1$. Then $(r')^*$ is lexicographic, and, from the first paragraph of the proof,

$$\Lambda(r') = (-1)^{\nu(r')}. \qquad \text{(IV. 18)}$$

Also

$$\sigma_{r*}(r) = r' = \sigma_{r'}^{-1}(r_0),$$

and so

$$\sigma_{r'}\sigma_{r*}(r) = r_0.$$

Therefore,

$$\begin{aligned}\Lambda(r) &= \operatorname{sgn}(\sigma_{r'}\sigma_{r*})\\ &= \Lambda(r')\Lambda(r*)\\ &= (-1)^{\nu(r')+\nu(r^*)},\end{aligned} \qquad \text{(IV. 19)}$$

by (IV. 18) and the inductive hypothesis. But

$$\nu(r*) = \sum_{k=2}^{n-1} \nu_k(r*) = \sum_{k=2}^{n-1} \nu_k(r)$$

and

$$\nu(r') = n - i(r') = n - i(r) = \nu_n(r).$$

Thus (IV. 19) becomes $\Lambda(r) = (-1)^{\nu(r)}$, and the proof is complete.

V. THE WORKING TRIPLET AND ITS USE

Each triplet belongs to an equivalence class and the coefficients of only one member of this class (the working triplet) are calculated and listed. The definition of the classes and the operations for obtaining the other members are given in Ref. 1 (Sec. VIII). In practice, these operations need not be carried out as explicitly as stated there. Instead, we give a procedure for obtaining the working triplet, given any other member of its class, as well as a procedure for construction of the coefficients of a required triplet from the (listed) coefficients of the working triplet.

Given a triplet, to find the working triplet of its class, do the following:

Make a list of six representations, the first three those of the required triplet listed in the order in which they appear in the triplet, the last three, the conjugates of the first three, in the same order. Representations are referred to by their positions in the table. It is convenient to write the table in the form

$$\begin{array}{lll}\alpha_1 & \alpha_2 & \alpha_3\\ \bar{\alpha}_1 & \bar{\alpha}_2 & \bar{\alpha}_3.\end{array}$$

Count the number of self-conjugate representations in the given triplet.

Now pick out the smallest of the six representations. This is the first representation of the working triplet (first "working representation"). Strike out this one and the one which appears in the same column of the table; if there are repetitions, it does not matter which, if any, of the relevant columns is deleted. Of the remaining four representations, pick out the smallest. This is the second working representation. Strike out its column. This leaves one of the original three representations and its conjugate.

Suppose the original triplet contains no self-conjugate representations. If the first two working representations were picked from the same row of the table (that is, if they were both conjugates of the original representations, or neither were), then the third working representation is the remaining representation of the original triplet. Otherwise, it is the conjugate of the remaining one. If there are three self-conjugate representations in the original triplet, the third working representation is the only one remaining. For the case of one or two self-conjugate representations, the third working representation is the minimum of the pair remaining in the table.

These selections determine the working triplet; however, we must also specify a chain of operations to construct the required coefficient. We need an additional specification in the cases of one or two self-conjugate representations when exactly one of the non-self-conjugate representations has been conjugated in arriving at the working triplet.

For the case of one self-conjugate representation, we must conjugate its tableaux to construct the required coefficients. For two self-conjugate representations, the rule we use is to conjugate the tableaux of the first one appearing in the working triplet, reading from the left.

These are the only cases in which we conjugate the tableaux of a self-conjugate representation.

How do we find the coefficients for an arbitrary triplet? The first step is to find the working triplet for its equivalence class. The construction of the coefficients proceeds, according to the following rules, from the listed coefficients. We emphasize that conjugation refers to tableaux as well as to representations. Thus, we must keep track of conjugation, even for self-conjugate representations.

1. Write down, in a row, the ordinals of the frames of the given triplet. Below each put the ordinal of its conjugate. The result is a two-row, three-column table.

2. Determine the ordinals of the frames of the working triplet in accordance with the rules of the previous discussion. For each frame chosen note whether the choice came from the first or second row of this table.

3. From each column of the table exactly one entry has been chosen for the working triplet. Determine the position of this entry in the working triplet.

4. Write down, in order, the ordinals of the tableaux labeling the desired coefficient. If a frame ordinal was chosen from the second row of the table (that is, the representation was conjugated), replace the corresponding tableau ordinal by the ordinal of the conjugate. This latter number is obtained by subtracting the original ordinal from the dimension of the frame, and adding one.

5. Arrange these three resulting numbers. The final order must be the same as that in which their corresponding frames appear in the working triplet.

6. Look up that coefficient of the working triplet labeled by the resulting triplet of tableau ordinals.

7. For any frame of the working triplet which was selected from the second row of the table, write down the sign function of the corresponding working tableau. Multiply the coefficient by the product of all of these.

8. Multiply the number obtained by the square root of the quotient of the dimension of the first of the three given frames divided by the dimension of the first frame of the working triplet. The result is the required coefficient.

The coefficients obtained here may not be the same as the coefficients obtained by using the table in Ref. 1, Sec. VIII.

We list, for S_3, the working triplets together with the members of corresponding equivalence classes. First, we give the frames for S_3, specifying, for each, the frame ordinal, the corresponding tableaux, and the ordinal of the conjugate frame.

Frame ordinal	Corresponding tableaux	Conjugate frame ordinal
1	$r=1$: $\begin{array}{\|c\|}\hline 1\\ \hline 2\\ \hline 3\\ \hline\end{array}$	3
2	$r=1$ $\begin{array}{\|c\|c\|}\hline 1 & 2\\ \hline 3 \\ \cline{1-1}\end{array}$; $r=2$: $\begin{array}{\|c\|c\|}\hline 1 & 3\\ \hline 2 \\ \cline{1-1}\end{array}$	2
3	$r=1$: $\begin{array}{\|c\|c\|c\|}\hline 1 & 2 & 3\\ \hline\end{array}$	1

The working triplets, equivalence classes and instructions for obtaining a member of a class from its working triplet are as follows (frames are labeled by ordinals):

Working triplet	Equivalent triplets	Representations in equivalent triplet whose tableaux are conjugated
1, 1, 1	1, 1, 1	---
	1, 3, 3	second and third
	3, 1, 3	first and third
	3, 3, 1	first and second
1, 1, 2	1, 1, 2	---
	1, 2, 1	---
	1, 2, 3	second and third
	1, 3, 2	second and third
	2, 1, 1	---
	2, 1, 3	first and third
	2, 3, 1	first and second
	2, 3, 3	second and third
	3, 1, 2	first and third
	3, 2, 1	first and second
1, 1, 2	3, 2, 3	first and third
	3, 3, 2	first and second
1, 1, 3	1, 1, 3	---
	1, 3, 1	---
	3, 1, 1	---
	3, 3, 3	first and second
1, 2, 2	1, 2, 2	---
	2, 1, 2	---
	2, 2, 1	---
	2, 2, 3	first and third
	2, 3, 2	first and second
	3, 2, 2	first and second
2, 2, 2	2, 2, 2	---

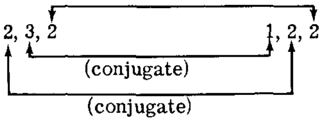
We give a specific example of the way in which the coefficients for a given triplet are found from those of the working triplet, for the 2, 3, 2 triplet. The working triplet is the 1, 2, 2 triplet; the multiplicity, $m(1, 2, 2)$ is one.

To find the required coefficients, we first note that

$$(f^{[2]}/f^{[1]})^{1/2}=\sqrt{2}.$$

The sign function, Λ, on the (only) tableau of the representation of the representation of the representation $\alpha=3$ is 1 and the sign function on the tableaux of $\alpha=2$ becomes $\Lambda(1)=+1$ and $\Lambda(2)=-1$.

The chain of operations used in arriving at the corresponding tableaux for the working triplet may be described schematically as



We thus obtain the following:

Ordinals, r,p,p', of tableaux of 2, 3, 2	Ordinals, $\bar{p},\tilde{r},p'$, of corresponding tableaux of 1, 2, 2	Multiply working coefficient by
1, 1, 1	1, 2, 1	$\sqrt{2}\,\Lambda(2)=-\sqrt{2}$
1, 1, 2	1, 2, 2	$\sqrt{2}\,\Lambda(2)=-\sqrt{2}$
2, 1, 1	1, 1, 1	$\sqrt{2}\,\Lambda(1)=\sqrt{2}$
2, 1, 2	1, 1, 2	$\sqrt{2}\,\Lambda(1)=\sqrt{2}$

The required coefficients are

$$C^{2,\underset{1}{3},2}(1, 1)=-\sqrt{2}\,C^{1,\underset{1}{2},2}(2, 1)=1,$$

$$C^{2,\underset{1}{3},2}(1, 2)=-\sqrt{2}\,C^{1,\underset{1}{2},2}(2, 2)=0,$$

$$C^{2,\underset{2}{3},2}(1, 1)=\sqrt{2}\,C^{1,\underset{1}{2},2}(1, 1)=0,$$

$$C^{2,\underset{2}{3},2}(1, 2)=\sqrt{2}\,C^{1,\underset{1}{2},2}(1, 2)=1,$$

where we have used the values of the working coefficients obtained from Table I, expressed in radical form, which are

$$C^{1,\underset{1}{2},2}(1, 2)=\sqrt{2}/2,\quad C^{1,\underset{1}{2},2}(2, 1)=-\sqrt{2}/2,\quad \text{and}$$

$$C^{1,\underset{1}{2},2}(1, 1)=C^{1,\underset{1}{2},2}(2, 2)=0.$$

ACKNOWLEDGMENTS

We would like to thank the Educational Computer Center of the City University of New York for a grant of computer time and Dr. Dina Bedi, Ray Rauth, and the staff of the Baruch College (CUNY) Computer Center, and Dr. Leon Landovitz and the staff of the Graduate Center (CUNY) Computer Center for their aid and advice. We would also like to thank the Baruch College Small Grants Program and the CUNY Research Foundation for their assistance.

*Supported by City University of New York, Educational Computer Center.

[1] Susan Schindler and R. Mirman, J. Math. Phys. **18**, 1678 (1977).

[2] See AIP document no. PAPS JMAPA-18-1697-84 for 84 pages of the Clebsch—Gordan coefficients to 16 decimal places for $n=2,\ldots,6$. Order by PAPS number and journal reference from American Institute of Physics, Physics Auxiliary Publication Service, 335 East 45th Street, New York, N.Y. 10017. The price is $1.50 for each microfiche (98 pages), or $5 for photocopies of up to 30 pages with $0.15 for each additional page over 30 pages. Airmail additional. Make checks payable to the American Institute of Physics.

[3] C. Itzykson and M. Nauenberg, Rev. Mod. Phys. **38**, 95 (1966).

Invariant solutions for a class of diffusion equations

Mayer Humi

Department of Mathematics, Worcester Polytechnic Institute, Worcester, Massachusetts 01609

(Received 9 August 1976; revised manuscript received 20 September 1976)

We show that a class of diffusion equations, related to superradiant emission, forms a subset of a wider class of equations which are invariant with respect to some one parameter group of transformations. This property gives rise to solutions which are invariant with respect to the aforementioned group. In particular we show that a solution found by L. H. Narducci *et al.* [Phys. Rev. A **11**, 1354 (1975)] for one of these equations is an invariant solution.

I. INTRODUCTION

Recently Narducci *et al.* showed[1,2] that some superradiant emission processes can be described by a single Fokker—Planck equation of the form

$$[x(1+x)u]_{xx}+[(1+Nx)u]_x=u_t,\quad x,t>0, \tag{1.1}$$

where N is a constant and subscripts indicate differentiation with respect to the indicated variables. In the lowest order approximation Eq. (1.1) reduces to

$$xu_{xx}+(1-Nx)u_x-Nu=u_t \tag{1.2}$$

and it was demonstrated in Ref. 2 that this last equation admits a solution of the form

$$u(x,t)=b^{-1}(t)\exp(-z),\quad z=xb^{-1}(t), \tag{1.3}$$

where

$$b(t)=N^{-1}[\exp(Nt)-1]. \tag{1.4}$$

This solution of Eq. (1.2) is interesting from a physical point of view as it represents a conservative diffusion process with a delta function initial conditions.

Lately[3] this result was generalized to a wider class of equations in the form

$$A(x)u_{xx}+B(x)u_x+C(x)u=u_t, \tag{1.5}$$

where the coefficients A, B, C are given by

$$A(x)=\alpha x^{\lambda+2},\quad B(x)=\beta_1x^{\lambda+1}+\beta_2x,\quad \alpha>0$$
$$C(x)=\lambda[\beta_1-\alpha(2+\lambda)]x^{\lambda}+\beta_2, \tag{1.6}$$

and where $\lambda, \beta_1, \beta_2$ are arbitrary constants.

We note, however, that these results were derived in Refs. 2 and 3 through analytic techniques. In this paper we approach this same problem from a group theoretical point of view and demonstrate that this approach can contribute a new insight into some of the equations mentioned above and their solutions. At the same time, nevertheless, this new approach can be used to solve other classes of equations of the form (1.5) which might be relevant to the description of other superradiant processes under current research.

In Sec. II we consider the general set of equations of the form (1.5) and find the constraints that the coefficients A, B, C have to satisfy for such an equation to be invariant with respect to a one-parameter group of transformations. We then exhibit explicitly in Sec. III some of these classes of invariant equations and show that Eq. (1.2) as well as Eqs. (1.5)—(1.6) for $\beta_2=0$ belong to this class of equations.

Invariant equations of the form (1.5) are reducible to an oridnary differential equation *in the invariants of the groups* (other reduction techniques are of no relevance in our context). The solutions of the reduced equations provide, then, invariant solutions for the respective original equations. This program is carried out explicitly in Sec. IV for those classes of invariant equations discussed in Sec. III. In particular we show that (1.3) is an invariant solution of (1.2) and construct a second one in the same similarity variable.

II. CONDITIONS FOR INVARIANCE

To find out those equations of the form (1.5) which admit a one-parameter group of transformations we subject them to a general infinitesimal transformation of the form

$$\begin{aligned}\bar{x}&=x+\epsilon X(x,t,u)+O(\epsilon^2),\\ \bar{t}&=t+\epsilon T(x,t,u)+O(\epsilon^2),\\ \bar{u}&=u+\epsilon U(x,t,u)+O(\epsilon^2),\end{aligned} \tag{2.1}$$

and require that the original equation remain invariant to first order for a proper choice of X, T and U. After a lengthy calculation[4] we thus obtain;

$$\begin{aligned}A(\bar{x})\bar{u}_{\bar{x}\bar{x}}+B(\bar{x})\bar{u}_{\bar{x}}+C(\bar{x})\bar{u}-\bar{u}_{\bar{t}}=&[A(x)u_{xx}+B(x)u_x+C(x)u-u_t]+\epsilon\{[A(x)U_{xx}+B(x)U_x+C(x)U-U_t]+u_x[A(x)(2U_{xu}-X_{xx})\\ &+(U_u-X_x)B(x)+XB_x+X_t]+u_t[-A(x)T_{xx}-B(x)T_x-(U_u-T_t)]+u_x^2[A(x)(U_{uu}-2X_{xu})-B(x)X_u]\\ &+u_xu_t[-A(x)T_{xu}+X_u-B(x)T_u]-u_x^3A(x)X_{uu}-u_x^2u_tA(x)T_{uu}+u_{xx}[A(x)(U_u-2X_{xx})+XA_x]-2u_{xt}A(x)T_x-3u_{xx}u_xA(x)X_u\\ &-u_{xx}u_tA(x)T_u-2u_{xt}u_xA(x)T_u+u_t^2T_u+uXC_x\}+O(\epsilon^2).\end{aligned} \tag{2.2}$$

To achieve invariance we must nullify the first order terms. The classical method to achieve this is to require that after substituting for u_t in terms of Eq. (1.5) each separate coefficient of u and its derivatives be zero.

This leads, after obvious simplifications, to the following system of equations:

$$AU_{xx}+BU_x+CU-U_t=0, \tag{2.3}$$

$$XA_x+A(T_t-2X_x)=0, \tag{2.4}$$

$$A(2U_{xu}-X_{xx})+B(x)(T_t-X_x)+XB_x+X_t=0, \tag{2.5}$$

$$XC_x-C(U_u-T_t)=0, \tag{2.6}$$

$$T_x=T_u=X_u=U_{uu}=0. \tag{2.7}$$

We stress at this point that the classical method provides only sufficient conditions for invariance of Eq. (1.5). In fact we could choose to nullify the first order terms in (2.2) in any appropriate combination, e.g., replacing Eqs. (2.3), (2.6) by a single equation

$$AU_{xx}+BU_x+CU-U_t+[XC_x-C(U_u-T_t)]u=0. \tag{2.8}$$

We shall see later on that while there exist no solution of the classical system which yield Eq. (1.2) we can recover it (and a whole family of related equations) by using the nonclassical system.

In the following we do not attempt to solve the classical or nonclassical system in all generality. In fact we shall concentrate our attention to some particular families of invariant equation and thus show that Eqs. (1.2), (1.5)—(1.6) form a subset of this wider class.

III. INVARIANT EQUATIONS

We start with some general remarks regarding the analysis of the classical system (2.3)—(2.7).

If we are given an equation of the form (1.5) then from (2.7) we infer that X, T, and U are of the form.

$$T=T(t),\quad X=X(x,t),\quad U=F(x,t)u+G(x,t). \tag{3.1}$$

To proceed, we assume, $X=a(x)g(t)$ and use (2.4) to obtain after separation of variables

$$a(x)(A_x/A)-2a_x=-\nu, \tag{3.2}$$

$$T_t/g(t)=\nu, \tag{3.3}$$

where ν is a constant. This yields

$$a(x)=A^{1/2}[(\nu/2)\int A^{-1/2}\,dx+k], \tag{3.4}$$

where k is a constant. Equation (2.6) then leads to

$$U=g(t)[\nu+(C_x/C)\,a(x)]u+G(x,t). \tag{3.5}$$

Equations (2.3), (2.5) can now be used to determine $G(x,t)$ and to check whether the system (2.3)—(2.7) is consistent with the given functional values of B and $F(x,t)$.

In this section, however, our approach is different. We usually choose an appropriate value for U and use A (or C) as a parameter to determine the possible values of B and C (or A).

We point out nevertheless that the preceding analysis will serve for guidance in the following:

Case 1: We assume $U=0$ and $C_x\neq 0$ and use the classical system.

As a result of (3.1) we can put $T_t=f(t)$. Equation (2.6) then implies

$$X=-f(t)(C/C_x). \tag{3.6}$$

Substitution of (3.6) in (2.4) and integration yields

$$A=\eta(C^3/C_x^2), \tag{3.7}$$

where η is a constant. From (2.5) we then obtain after separation of variables

$$f'(t)/f(t)=\beta, \tag{3.8}$$

$$\eta\frac{C^2}{C_x}\left(\frac{C}{C_x}\right)_{xx}+\frac{C_x}{C}\left[1+\left(\frac{C}{C_x}\right)_x\right]B-B_x=\beta. \tag{3.9}$$

Thus $f(t)=k\exp(\beta t)$ and Eq. (3.9) is a linear ordinary differential equation for B in terms of C and its derivatives.

Example 1: If $C=\mu x^\lambda$ $\lambda\neq 0$ then $A=\eta x^{\lambda+2}$ and $B=\delta x^{\lambda+1}+(\beta/\lambda)x$, where μ, η, δ and β are arbitrary constants. Thus we find that the family of differential equations of the form

$$\eta x^{\lambda+2}u_{xx}+[\delta x^{\lambda+1}+(\beta/\lambda)x]u_x+\mu x^\lambda u=u_t \tag{3.10}$$

admits a one-parameter group of transformations in the form

$$U=0,\quad X=-kx\exp(\beta t)/\lambda,\quad T=\beta^{-1}k(\exp(\beta t)-1)+\gamma,$$
$$k\neq 0,\quad \beta,\gamma\in R. \tag{3.11}$$

Remarks: 1. We note that the form of the infinitesimal generator T has been chosen so that the limit $\beta\to 0$ is meaningful.

2. The same set of invariant equations can be obtained by using the nonclassical system viz. Eqs. (2.4), (2.5), (2.7), and (2.8). In this case, however, U can take a more general form $U=\alpha u$, where α is an arbitrary constant while X, T remain the same.

3. When $\beta=0$ Eqs. (3.10) coincide with Eqs. (1.5)—(1.6) with $\beta_2=0$.

Case 2: We assume $C_x=0$, $U=F(x,t)u$, and use the classical system.

From (2.6) it follows that

$$U_u=T_t=f(t)$$

and Eq. (2.3), then yields $f=k\exp(Ct)$. Assuming $X=a(x)g(t)$ we can use Eqs. (3.3), (3.4) to determine $a(x)$ and $g(t)$ while (2.5) is a linear differential equation for B in terms of A, X, and f.

Example 2: If $A=x$ then

$$a(x)=\nu x+kx^{1/2}, \tag{3.12}$$

where k is a constant. If we put $k=0$ in (3.12) then we obtain for B

$$B(x)=-Cx+\rho, \tag{3.13}$$

where ρ is a constant. Thus the family of invariant equations obtained in this manner is

$$xu_{xx}+(\delta-Cx)u_x+Cu=u_t. \tag{3.14}$$

Although (3.14) resembles Eq. (1.2) there is still a sign discrepancy in the coefficient of $B(x)$.

We remark that we could have had adopted a different approach to this problem with $C_x=0$ by assuming $U=0$. From (2.6) we then obtain $T=$const and from Eq. (2.4) that $X=A^{1/2}f(t)$. Equation (2.5) then yields after separation of variables

$$f'(t)/f(t)=\beta, \tag{3.15}$$

$$-\frac{1}{4}\,\frac{(A_x)^2}{A}+\tfrac{1}{2}A_{xx}+\tfrac{1}{2}B\frac{A_x}{A}-B_x=\beta. \tag{3.16}$$

For $A=x$ we obtain

$$B=-2\beta x+\tfrac{1}{2}+kx^{1/2}. \tag{3.17}$$

Once again we do not recover Eq. (1.2). This and similar deliberations lead us to conclude that the classical system cannot generate a one-parameter group with respect to which Eq. (1.2) is invariant.

Case 3: We assume that $C=\text{const.}$, $U=F(t)u$ and use the nonclassical system.

By our assumptions Eq. (2.8) reduces to

$$F'(t)=CT_t. \tag{3.18}$$

Assuming $X=a(x)g(t)$, Eq. (2.4) yields Eqs. (3.3), (3.4), and Eq. (2.5) reduces after separation of variables to

$$g'(t)/g(t)=\beta \tag{3.19}$$

$$-A\frac{a_{xx}}{a}+B\ \frac{\nu-a_x}{a}\ +B_x=-\beta. \tag{3.20}$$

Thus $g(t)=k\exp(Bt)$ and hence $F(t)=(1/\beta)[KC\nu\exp(\beta t)+\gamma]$, where β,γ are arbitrary constants.

Example 3: If $A=x$ we still obtain Eq. (3.12) but if we now put $k=0$ and use Eq. (3.20) we obtain

$$B=\beta x+\delta. \tag{3.21}$$

Thus the family of equations obtained in this manner is

$$xu_{xx}+(\delta-\beta x)u_x+Cu=u_t. \tag{3.22}$$

Equation (1.2) obviously belongs to this family of equations.

IV. INVARIANTS AND SOLUTIONS

An equation of the form (1.5) which admits a one-parameter group of transformations characterized by X, T, and U can be reduced to an ordinary differential equation through the invariants of this group.

These invariants can be obtained as solutions of the linear equation[4]

$$X\frac{\partial I}{\partial x}+T\frac{\partial I}{\partial t}+U\frac{\partial I}{\partial u}=0. \tag{4.1}$$

This equation has two independent solutions. The first has to be used as the independence variable in the reduced equation while the second serves as the dependent variable. The solutions of the reduced equation then furnishes invariant solutions for the original equation.

In the following we compute explicitly the invariants and the reduced equations for the families found in examples 1—3 and thus show the origin of the similarity variables found in Refs. 2 and 3. Moreover we show that the solution of (1.2) given by (1.3) is invariant solution.

Example 1 (continuation): For equations of the form (3.10) and X, T, and U given by Eq. (3.11) the invariants are

$$I_1=C[\beta^{-1}k(\exp(\beta t)-1)+\gamma],\quad I_2=u. \tag{4.2}$$

Thus to reduce these equations we have to choose $z=I_1$ as the independent variable and $q(z)=I_2$ as the dependent one. The reduced equations are then in the form

$$\eta\lambda^2z^2q''+\{[\lambda\delta+\eta\lambda(\lambda-1)]z+\mu\beta\gamma-\mu k\}q'+\mu q=0, \tag{4.3}$$

where primes denote differentiation with respect to z.

These reduced equations can be easily solved by a power series or in an integral form through Euler transformation.[5] We note however that the parameter γ in (4.3) is arbitrary and this gives rise to a family of invariant solutions for the original equation. This is a result of the translational invariance in time of Eq. (1.5).

Remarks: 1. The subclass of Eqs. (1.5)—(1.6) which are in the form (3.10) is characterized by $\beta=\beta_2=0$. Equation (4.3) is then an Euler equation.[5]

2. If instead of using the infinitesimals given by Eq. (3.11) we used those obtained from the nonclassical system I_2 will change to

$$I_2=u^\alpha[CK\exp(\beta t)]^{-1/\gamma},$$

where α is an arbitrary constant. However the reduced equations are still similar to Eq. (4.3).

Example 2 (continuation): For equations in the form (3.14) X, T, and U are given by

$$X=Kx\exp(Ct),\quad T=C^{-1}[K\exp(Ct)+\gamma],$$
$$U=K\exp(Ct)u. \tag{4.4}$$

The invariants are

$$I_1=x[K\exp(Ct)+\gamma]^{-1},\quad I_2=u/x. \tag{4.5}$$

The reduced equations are then in the form

$$z^2q''+[(\delta+2)z-\gamma Cz^2]q'+\delta q=0. \tag{4.6}$$

Example 3 (continuation): For Eq. (3.19) the infinitesimal transformations are

$$\begin{aligned}X&=K\nu x\exp(\beta t),\\ T&=\beta^{-1}[K\nu\exp(\beta t)+\gamma],\\ U&=[(KC\nu\exp(\beta t)/\beta)+\rho]u.\end{aligned} \tag{4.7}$$

(In the following we put $\rho=0$.)

The invariants are

$$I_1=x[K\nu\exp(\beta t)+\gamma]^{-1}\quad I_2=x^{-C/\beta}u, \tag{4.8}$$

and the reduced equation is

$$z^2q''+[(\delta+2\alpha)z-\beta\gamma z^2]q'+[\delta\alpha+\alpha(\alpha-1)]q=0, \tag{4.9}$$

where $\alpha=C/\beta$. In Eq. (1.2) $\beta=N$, $C=-N$, and $\delta=1$. If we substitute these values in Eq. (4.9) and choose $\beta\gamma=-1$ we obtain

$$z^2q''+(z^2-z)q'+q=0. \tag{4.10}$$

A fundamental system of solutions for this equation is given by

$$q_1 = z\exp(-z), \quad q_2 = z\exp(-z)\int z^{-1}\exp(z)\,dz. \tag{4.11}$$

Thus two invariant solutions of (1.2) are $u_1 = x^{-1}q_1$ and $u_2 = x^{-1}q_2$. The solution given by u_1 coincide with the one given by Eq. (1.3), while u_2 is a new solution whose physical significance has to be explored.

[1] L.M. Narducci, C.A. Coulter and C.M. Bowden, Phys. Rev. A 9, 829 (1974).

[2] L.M. Narducci and V. Bluemel, Phys. Rev. A 11, 1354 (1975).

[3] S.H. Lehnigk, J. Math. Phys. 17, 973 (1976).

[4] W.F. Ames, *Nonlinear Partial Differential Equations in Engineering* (Academic Press, New York, 1972), Vol. II, see especially Chap. II.

[5] E.L. Ince, *Ordinary Differential Equations* (Dover Publications, New York, 1956).

ERRATA

Erratum: Zeros of the grand canonical partition function: Generalization of a result of Lee and Yang [J. Math. Phys. 17, 2166 (1976)]

M. E. Baur and R. M. Sinder

Department of Chemistry, University of California, Los Angeles, California 90024

(Received 18 April 1977)

In this article, a proof was claimed for the proposition that sufficient conditions for finite polynomials of grand canonical partition function form, $\mathcal{Q}(M) = \sum_n Q_n(M) z^n$ to have all zeros on the unit circle $|z| = 1$ are:

(1) symmetry, $Q_n(M) = Q_{M-n}(M)$,

(2) boundedness by the binomial series, $Q_n(M) \leqslant \binom{M}{n}$.

The proof given is incorrect. The error is in the assertion $|\mathcal{Q}^{(k)}| \geqslant |N_k|\ |R(k; k+1)|$ following Eq. (10); the stated conditions are not sufficiently strong to guarantee validity of this inequality.

It is instructive to consider the character of counterexamples. Those which we have been able to construct either have oscillating coefficients $Q_n(M)$ (as in $z^4 + \frac{17}{4}z^2 + 1$) or sharp jumps in magnitude of some $Q_n(M)$ (as in $z^4 + z^3 + 6z^2 + z + 1$). Such properties do not violate conditions (1) or (2) above, but are inconsistent with the criteria for thermodynamic stability.[1] A plausible conjecture is that the additional conditions required to guarantee that all zeros lie on the unit circle in the z plane amount to a certain smooth convexity in the $Q_n(M)$, perhaps no more restrictive than that necessary for thermodynamic stability.

[1] K. Huang, *Statistical Mechanics* (Wiley, New York, 1963), Chap. 8.

Erratum: A C^*-algebra formulation of the quantization of the electromagnetic field [J. Math. Phys. 18, 629 (1977)]

A. L. Carey, J. M. Gaffney, and C. A. Hurst

Department of Mathematical Physics, University of Adelaide, South Australia 5001

(Received 27 April 1977)

The proof of Theorem 3.1 is incorrect as $\overline{\Delta(N)} \neq \Delta_c(N)$ (cf. Ref. 1). A correct proof is obtained by observing that π is continuous provided its restriction f, say, to $\Delta(T)$ is continuous. Now the Schrödinger representation, ρ, of $\Delta_c(M)$ provides a faithful representation of $\Delta_c(T)$. The direct integral decomposition of ρ restricted to $\Delta_c(N)$ described in Sec. 3, is a direct integral over the spectrum Σ of $\Delta_c(T)$. As ρ is faithful, we have $\Delta_c(T)$ isomorphic to the C^*-algebra of continuous functions, $C(\Sigma)$ on Σ. The map f is an evaluation map on $C(\Sigma)$ and so is continuous in the sup norm on $C(\Sigma)$ and hence is continuous on $\Delta_c(T)$. Thus π is continuous. We would like to thank the referee for pointing out this method of proof.

[1] J. Manuceau, M. Sirigue, D. Testard, and A. Verbeure, Commun. Math. Phys. 32, 231—43 (1973).

A fundamental system of solutions for this equation is given by

$$q_1 = z\exp(-z), \quad q_2 = z\exp(-z)\int z^{-1}\exp(z)\,dz. \tag{4.11}$$

Thus two invariant solutions of (1.2) are $u_1 = x^{-1}q_1$ and $u_2 = x^{-1}q_2$. The solution given by u_1 coincide with the one given by Eq. (1.3), while u_2 is a new solution whose physical significance has to be explored.

[1] L.M. Narducci, C.A. Coulter and C.M. Bowden, Phys. Rev. A 9, 829 (1974).
[2] L.M. Narducci and V. Bluemel, Phys. Rev. A 11, 1354 (1975).
[3] S.H. Lehnigk, J. Math. Phys. 17, 973 (1976).
[4] W.F. Ames, *Nonlinear Partial Differential Equations in Engineering* (Academic Press, New York, 1972), Vol. II, see especially Chap. II.
[5] E.L. Ince, *Ordinary Differential Equations* (Dover Publications, New York, 1956).

ERRATA

Erratum: Zeros of the grand canonical partition function: Generalization of a result of Lee and Yang [J. Math. Phys. 17, 2166 (1976)]

M. E. Baur and R. M. Sinder

Department of Chemistry, University of California, Los Angeles, California 90024

(Received 18 April 1977)

In this article, a proof was claimed for the proposition that sufficient conditions for finite polynomials of grand canonical partition function form, $\mathcal{Q}(M) = \sum_n Q_n(M) z^n$ to have all zeros on the unit circle $|z| = 1$ are:

(1) symmetry, $Q_n(M) = Q_{M-n}(M)$,

(2) boundedness by the binomial series, $Q_n(M) \leqslant \binom{M}{n}$.

The proof given is incorrect. The error is in the assertion $|\mathcal{Q}^{(k)}| \geqslant |N_k|\ |R(k;k+1)|$ following Eq. (10); the stated conditions are not sufficiently strong to guarantee validity of this inequality.

It is instructive to consider the character of counterexamples. Those which we have been able to construct either have oscillating coefficients $Q_n(M)$ (as in $z^4 + \frac{17}{4}z^2 + 1$) or sharp jumps in magnitude of some $Q_n(M)$ (as in $z^4 + z^3 + 6z^2 + z + 1$). Such properties do not violate conditions (1) or (2) above, but are inconsistent with the criteria for thermodynamic stability.[1] A plausible conjecture is that the additional conditions required to guarantee that all zeros lie on the unit circle in the z plane amount to a certain smooth convexity in the $Q_n(M)$, perhaps no more restrictive than that necessary for thermodynamic stability.

[1] K. Huang, *Statistical Mechanics* (Wiley, New York, 1963), Chap. 8.

Erratum: A C^*-algebra formulation of the quantization of the electromagnetic field [J. Math. Phys. 18, 629 (1977)]

A. L. Carey, J. M. Gaffney, and C. A. Hurst

Department of Mathematical Physics, University of Adelaide, South Australia 5001

(Received 27 April 1977)

The proof of Theorem 3.1 is incorrect as $\overline{\Delta(N)} \neq \Delta_c(N)$ (cf. Ref. 1). A correct proof is obtained by observing that π is continuous provided its restriction f, say, to $\Delta(T)$ is continuous. Now the Schrödinger representation, ρ, of $\Delta_c(M)$ provides a faithful representation of $\Delta_c(T)$. The direct integral decomposition of ρ restricted to $\Delta_c(N)$ described in Sec. 3, is a direct integral over the spectrum Σ of $\Delta_c(T)$. As ρ is faithful, we have $\Delta_c(T)$ isomorphic to the C^*-algebra of continuous functions, $C(\Sigma)$ on Σ. The map f is an evaluation map on $C(\Sigma)$ and so is continuous in the sup norm on $C(\Sigma)$ and hence is continuous on $\Delta_c(T)$. Thus π is continuous. We would like to thank the referee for pointing out this method of proof.

[1] J. Manuceau, M. Sirigue, D. Testard, and A. Verbeure, Commun. Math. Phys. 32, 231—43 (1973).